UNIVERSIDAD POLITÉCNICA DE VALENCIA

Programa de doctorado en Industrias de la Comunicación y Culturales

Departamento de Comunicación Audiovisual, Documentación e Historia del Arte

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

TESIS DOCTORAL

Diseño de una metodología cibermétrica de cálculo del éxito para la optimización de contenidos web

Víctor Manuel Yeste Moreno

Directores

Dr. Jorge Ignacio Serrano Cobos

Dra. María de los Ángeles Calduch Losa

Valencia, septiembre de 2021

TOMO I

“Equipado con sus cinco sentidos,
el hombre explora el universo que
lo rodea y a sus aventuras las llama
ciencia.”

Edwin Powell Hubble
The Nature of Science, 1954

Agradecimientos

En primer lugar, me gustaría expresar mi agradecimiento a todas esas personas que me han ayudado respondiendo a mis preguntas, aportando así su pequeño granito de arena en esta gran aventura: Enrique Orduña-Malea y Sara Kophamel, entre otros.

Sara forma parte, además, de ese grupo de amigos tan especial que me ha apoyado durante años para que esto se hiciera realidad. Amigos y familia como Javier Sánchez, Alexandra Celi, Héctor Solís, Jano Solís, Carla Carretero, Marta Soler, Cristina Torre, Sandra Moreno, Charlotte Goodwin, Juanjo Grau, Anabel Botella, Vanesa Martínez, Sara Torres, Rubén Pozo y Mar Martínez.

Esta tesis también ha sido posible gracias al esfuerzo diario del equipo de Hello Friki, la web que ha servido de caso de uso, con gente maravillosa como Valentín Alcocer, Daniel Collado, Marta Catalán, Maite Araez, Israel Gordon, Tony Rey, Raúl Martín, Álvaro Gekko, Luís Martínez y Eva Gómez, entre otros.

Un agradecimiento especial a José-Antonio Ontalba-Ruipérez, por animarme a hacer esta tesis doctoral y a ir más allá de lo conocido mediante la investigación.

Muchísimas gracias a mis directores, Jorge Serrano-Cobos y Ángeles Calduch-Losa, por su dedicación y por haber estado ahí apoyándome hasta el final, siempre con una sonrisa y palabras de ánimo.

Por último, le dedico esta tesis doctoral a mis padres Agripino y Antonia, a mi hermano David y a mis abuelos Victoriano, Octavia y Remigia, por haber tenido tanta fe en mí, desde pequeñito. Siempre os tendré en mi corazón.

Índice General

TESIS DOCTORAL ***1***

Agradecimientos ***5***

Índice General ***7***

Índice de Figuras ***23***

Índice de Tablas ***45***

Resumen ***51***

Resum ***53***

Abstract ***55***

1. Introducción ***57***

1.1. Objeto de estudio **59**

a. Temático 62

b. Idiomático 63

c. Temporal 63

d. Unidad de observación 64

1.2. Objetivos **65**

1.2.1. Objetivo principal 65

1.2.2. Objetivos específicos 65

1.3. Justificación de la investigación **67**

a. Metodológica 67

b. Práctica 68

c. Académica 68

1.4. Tipo de investigación, metodología y fuentes **71**

1.4.1. Tipo de investigación 71

1.4.2. Estructura del trabajo 72

1.5. Fuentes de datos **73**

1.6. Fuentes de información **75**

2. Estado de la cuestión ***77***

2.1. Periodismo digital **81**

2.1.1. Consumo de contenidos 81

2.1.1.1. La transformación digital 81
2.1.1.2. Tipos de medios y contenidos periodísticos 88
2.1.1.3. Estructura de los contenidos en el periodismo digital 91
2.1.1.4. Modelos de comunicación 92
2.1.1.5. Periodismo automatizado y el procesamiento de la información 96
2.1.1.6. Medición de audiencias online 99
2.1.1.7. La publicidad en el periodismo digital 100
2.1.1.8. El periodismo y la Blogosfera 103
2.1.2. Twitter 105
2.1.2.1. La red social 105
2.1.2.2. Uso y tipos de contenidos 107
2.1.2.3. Escucha activa 110
2.1.3. Difusión de las noticias en Twitter 113
2.1.3.1. La adaptación del periodismo a Twitter 113
2.1.3.2. Twitter y la comunicación con la audiencia 115
2.1.3.3. Twitter como fuente de noticias en tiempo real 118
2.2. Medición web de éxito 123
2.2.1. Analítica web 123
2.2.1.1. Periodismo medible 123
2.2.1.2. Minería de uso de la web 125
2.2.1.3. KPI de la analítica web 130
2.2.1.4. Metodologías de la analítica web 145
2.2.1.5. Tests y experimentación 154
a) Test A/B 154
b) Test multivariante 155
c) Test de experiencia 156
2.2.2. Cibermetría 159
2.2.2.1. Definición 159
2.2.2.2. Áreas de trabajo e indicadores del análisis cibermétrico 163
2.2.2.3. Rastreo de contenidos y redes de enlaces 166
2.2.2.4. Áreas de estudio 177
2.2.2.5. Análisis de la competencia 179
2.2.2.6. Altmetría 184
2.2.3. Analítica en Twitter 188
2.2.3.1. Minería de datos en Twitter 188
2.2.3.2. Análisis de la actividad 189
2.2.3.3. Análisis de la autoridad 193
2.2.3.4. Procesamiento de hashtags y spam 196
2.2.3.5. Social Media Analytics 199

2.2.3.6. Análisis de sentimientos 206
2.2.4. Análisis de tendencias en Twitter 211
2.2.4.1. La difusión de la información 211
2.2.4.2. Detección y clasificación de tendencias 214
2.2.4.3. Noticias de última hora 218
2.2.4.4. Procesamiento semántico 220
2.2.4.5. Predicción de tendencias 223
2.2.4.6. Otros tipos de análisis 225
2.2.5. Publicidad digital en la web 227
2.2.5.1. La publicidad en la World Wide Web 227
2.2.5.2. Optimización de la publicidad web 235
2.2.5.3. Sistemas de anuncios web 249
2.2.5.4. El fraude y el bloqueo de anuncios en la publicidad web 256
2.3. Conclusiones del estado de la cuestión 259
3. Metodología 261
3.1. Antecedentes y marco metodológico 263
3.2. Selección de los datos de estudio 269
3.3. Selección de temáticas 273
3.4. Indicadores 275
3.4.1. Analítica del contenido en la web 276
3.4.2. Analítica del contenido en la cuenta de Twitter 278
3.4.3. Análisis de tendencias 278
3.5. Instrumentos de recolección de datos 281
3.5.1. API de informes de Google Analytics v4 281
3.5.2. Twitter Standard API 284
3.6. Técnicas de procesamiento y análisis de datos 289
3.6.1. Tablas de la base de datos 290
3.6.2. Procesos de recogida de datos 295
3.7. Acotación y selección de las unidades de tiempo 299
3.8. Procedimiento para la representación e interpretación de los datos 301
3.9. Limitaciones metodológicas 307
4. Resultados de la investigación 311
4.1. Datos 313

4.1.1. Resumen 313
4.1.2. Variables 317
4.1.2.1. Variables de éxito 319
4.1.2.2. Variables de predicción 320
4.1.2.3. Tabla de variables 321
4.1.3. Objetivos estadísticos 322
4.2. Fase 1. Análisis de los datos de entrenamiento 325
4.2.1. Análisis de todos los artículos 326
4.2.1.1. Variables de éxito 326
a) Páginas vistas únicas (total) 326
b) AdSense eCPM (promedio) 338
c) Duración de la visita (promedio) 339
d) Páginas vistas por sesión (promedio) 340
e) Nº de retuits en la cuenta del medio (promedio) 341
f) Nº de favoritos en la cuenta del medio (promedio) 341
g) Nº de tuits de la tendencia 14 días después (total) 342
h) Nº de retuits de la tendencia 14 días después (total) 343
i) Nº de retuits de la tendencia 14 días después (promedio) 344
j) Normalidad y equidistribución de los residuos 345
k) Filtro de alta correlación 351
4.2.1.2. Variables de predicción 353
a) Número de tuits de la tendencia inicial (total) 353
b) Número de retuits de la tendencia inicial (total) 354
c) Número de retuits de la tendencia inicial (promedio) 355
d) Número de favoritos de la tendencia inicial (total) 355
e) Número de favoritos de la tendencia inicial (promedio) 356
f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 357
g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 358
h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio) 359
i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio) 360
j) Ratio de inclusión de URLs en los tuits de la tendencia inicial 361
k) Normalidad y equidistribución de los residuos 362
l) Filtro de alta correlación (colinealidad) 368
m) Análisis de componentes principales (ACP) 370
4.2.1.3. Regresión lineal múltiple 372
a) Páginas únicas (total) 373

b) Duración de la visita (promedio) 375
c) Páginas vistas por sesión (promedio) 377
d) Nº de favoritos en la cuenta del medio (promedio) 379
e) Nº de retuits de la tendencia 14 días después (promedio) 381
f) Resumen de relaciones entre variables de predicción y de éxito 383
4.2.1.4. Regresión binomial negativa o de Poisson 384
a) Filtro de alta correlación (colinealidad) 385
b) Páginas únicas (total) 387
c) Nº de tuits de la tendencia 14 días después (total) 388
d) Nº de retuits de la tendencia 14 días después (total) 391
e) Resumen de relaciones entre variables de predicción y de éxito 393
4.2.2. Análisis de otros conjuntos de artículos 395
4.2.3. Resumen de todos los análisis 396
4.3. Fase 2. Análisis de los datos de test 399
4.3.1. Validación de la predicción de todos los artículos 401
4.3.2. Validación de la predicción de los artículos de la categoría Cine 403
4.3.3. Validación de la predicción de los artículos de la categoría Series 405
4.3.4. Validación de la predicción de los artículos de la categoría Videojuegos 407
4.3.5. Validación de la predicción de los artículos sobre Tráileres 409
4.4. Selección de ecuaciones de predicción 411
4.5. Discusión de los resultados finales 415
5. Conclusiones 417
5.1. Objetivos específicos 421
5.1.1. Investigar el concepto de éxito en periodismo digital, la red social Twitter, la analítica web y la publicidad en la web 421
5.1.2. Diseñar la metodología y determinar qué herramientas y reportes son necesarios 421
5.1.3. Extraer los datos y procesarlos para obtener los indicadores 422
5.1.4. Realizar regresiones que permitan obtener ecuaciones de predicción de las variables de éxito seleccionadas 422
5.1.5. Validar las ecuaciones de predicción con datos de test y obtener su precisión 423
5.2. Conclusiones generales 425
5.3. Investigaciones relacionadas 427
5.3.1. COMRED 2021, II Congreso Internacional Comunicación y Redes Sociales de la Sociedad de la Información 427
5.3.2. CIMED 2021, I Congreso Internacional de Museos y Estrategias Digitales 428
5.4. Limitaciones y futuras líneas de investigación 431

Bibliografía *437*

Glosario *483*

6. Anexos *491*

6.1. Análisis de los subconjuntos de artículos **493**

6.1.1. Análisis de los artículos de la categoría Cine 493

6.1.1.1. Variables de éxito 493

a) Páginas vistas únicas (total) 493

b) AdSense eCPM (promedio) 494

c) Duración de la visita (promedio) 495

d) Páginas vistas por sesión (promedio) 496

e) Nº de retuits en la cuenta del medio (promedio) 497

f) Nº de favoritos en la cuenta del medio (promedio) 498

g) Nº de tuits de la tendencia 14 días después (total) 499

h) Nº de retuits de la tendencia 14 días después (total) 500

i) Nº de retuits de la tendencia 14 días después (promedio) 501

j) Normalidad y equidistribución de los residuos 502

k) Filtro de alta correlación 508

6.1.1.2. Variables de predicción 510

a) Número de tuits de la tendencia inicial (total) 510

b) Número de retuits de la tendencia inicial (total) 511

c) Número de retuits de la tendencia inicial (promedio) 512

d) Número de favoritos de la tendencia inicial (total) 513

e) Número de favoritos de la tendencia inicial (promedio) 514

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 515

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 516

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio) 517

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio) 518

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial 519

k) Normalidad y equidistribución de los residuos 520

l) Filtro de alta correlación (colinealidad) 527

m) Análisis de componentes principales (ACP) 529

6.1.1.3. Regresión lineal múltiple 531

a) Páginas únicas (total) 532

b) Duración de la visita (promedio) 534

c) Páginas vistas por sesión (promedio) 536
d) Nº de favoritos en la cuenta del medio (promedio) 538
e) Nº de retuits de la tendencia 14 días después (promedio) 540
f) Resumen de relaciones entre variables de predicción y de éxito 542
6.1.1.4. Regresión binomial negativa o de Poisson 544
a) Filtro de alta correlación (colinealidad) 544
b) Páginas únicas (total) 546
c) Nº de tuits de la tendencia 14 días después (total) 548
d) Nº de retuits de la tendencia 14 días después (total) 550
e) Resumen de relaciones entre variables de predicción y de éxito 553
6.1.2. Análisis de los artículos de la categoría Series 556
6.1.2.1. Variables de éxito 556
a) Páginas vistas únicas (total) 556
b) AdSense eCPM (promedio) 557
c) Duración de la visita (promedio) 557
d) Páginas vistas por sesión (promedio) 558
e) Nº de retuits en la cuenta del medio (promedio) 559
f) Nº de favoritos en la cuenta del medio (promedio) 560
g) Nº de tuits de la tendencia 14 días después (total) 561
h) Nº de retuits de la tendencia 14 días después (total) 562
i) Nº de retuits de la tendencia 14 días después (promedio) 563
j) Normalidad y equidistribución de los residuos 564
k) Filtro de alta correlación 571
6.1.2.2. Variables de predicción 574
a) Número de tuits de la tendencia inicial (total) 574
b) Número de retuits de la tendencia inicial (total) 574
c) Número de retuits de la tendencia inicial (promedio) 575
d) Número de favoritos de la tendencia inicial (total) 576
e) Número de favoritos de la tendencia inicial (promedio) 577
f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 578
g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 579
h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio) 580
i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio) 581
j) Ratio de inclusión de URLs en los tuits de la tendencia inicial 582
k) Normalidad y equidistribución de los residuos 583
l) Filtro de alta correlación (colinealidad) 591

m) Análisis de componentes principales (ACP) 593
6.1.2.3. Regresión lineal múltiple 595
a) Páginas únicas (total) 596
b) AdSense eCPM (promedio) 598
g) Duración de la visita (promedio) 600
h) Páginas vistas por sesión (promedio) 602
c) Nº de retuits en la cuenta del medio (promedio) 604
d) Nº de favoritos en la cuenta del medio (promedio) 606
e) Nº de tuits de la tendencia 14 días después (total) 608
f) Nº de retuits de la tendencia 14 días después (promedio) 610
g) Resumen de relaciones entre variables de predicción y de éxito 613
6.1.2.4. Regresión binomial negativa o de Poisson 614
a) Filtro de alta correlación (colinealidad) 614
b) Páginas únicas (total) 616
c) Nº de tuits de la tendencia 14 días después (total) 619
d) Nº de retuits de la tendencia 14 días después (total) 621
e) Resumen de relaciones entre variables de predicción y de éxito 624
6.1.3. Análisis de los artículos de la categoría Videojuegos 626
6.1.3.1. Variables de éxito 626
a) Páginas vistas únicas (total) 626
b) AdSense eCPM (promedio) 627
c) Duración de la visita (promedio) 627
d) Páginas vistas por sesión (promedio) 628
e) Nº de retuits en la cuenta del medio (promedio) 629
f) Nº de favoritos en la cuenta del medio (promedio) 630
g) Nº de tuits de la tendencia 14 días después (total) 631
h) Nº de retuits de la tendencia 14 días después (total) 632
i) Nº de retuits de la tendencia 14 días después (promedio) 633
j) Normalidad y equidistribución de los residuos 634
k) Filtro de alta correlación 641
6.1.3.2. Variables de predicción 644
a) Número de tuits de la tendencia inicial (total) 644
b) Número de retuits de la tendencia inicial (total) 645
c) Número de retuits de la tendencia inicial (promedio) 646
d) Número de favoritos de la tendencia inicial (total) 647
e) Número de favoritos de la tendencia inicial (promedio) 648
f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 649

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 650
h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio) 651
i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio) 652
j) Ratio de inclusión de URLs en los tuits de la tendencia inicial 653
k) Normalidad y equidistribución de los residuos 654
l) Filtro de alta correlación (colinealidad) 662
m) Análisis de componentes principales (ACP) 665
6.1.3.3. Regresión lineal múltiple 667
a) Páginas únicas (total) 668
b) AdSense eCPM (promedio) 670
c) Duración de la visita (promedio) 671
d) Páginas vistas por sesión (promedio) 672
e) Nº de favoritos en la cuenta del medio (promedio) 674
f) Nº de tuits de la tendencia 14 días después (total) 675
g) Nº de retuits de la tendencia 14 días después (promedio) 677
h) Resumen de relaciones entre variables de predicción y de éxito 679
6.1.3.4. Regresión binomial negativa o de Poisson 681
a) Filtro de alta correlación (colinealidad) 681
b) Páginas únicas (total) 683
c) Nº de tuits de la tendencia 14 días después (total) 685
d) Nº de retuits de la tendencia 14 días después (total) 687
e) Resumen de relaciones entre variables de predicción y de éxito 690
6.1.4. Análisis de los artículos de la categoría Tráileres 692
6.1.4.1. Variables de éxito 692
a) Páginas vistas únicas (total) 692
b) AdSense eCPM (promedio) 693
c) Duración de la visita (promedio) 693
d) Páginas vistas por sesión (promedio) 694
e) Nº de retuits en la cuenta del medio (promedio) 695
f) Nº de favoritos en la cuenta del medio (promedio) 696
g) Nº de tuits de la tendencia 14 días después (total) 697
h) Nº de retuits de la tendencia 14 días después (total) 698
i) Nº de retuits de la tendencia 14 días después (promedio) 699
j) Normalidad y equidistribución de los residuos 700
k) Filtro de alta correlación 708
6.1.4.2. Variables de predicción 710
a) Número de tuits de la tendencia inicial (total) 711

b) Número de retuits de la tendencia inicial (total) 711
c) Número de retuits de la tendencia inicial (promedio) 712
d) Número de favoritos de la tendencia inicial (total) 713
e) Número de favoritos de la tendencia inicial (promedio) 714
f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 715
g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio) 716
h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio) 717
i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio) 718
j) Ratio de inclusión de URLs en los tuits de la tendencia inicial 719
k) Normalidad y equidistribución de los residuos 720
l) Filtro de alta correlación (colinealidad) 728
m) Análisis de componentes principales (ACP) 730
6.1.4.3. Regresión lineal múltiple 732
a) Páginas únicas (total) 733
b) AdSense eCPM (promedio) 735
c) Duración de la visita (promedio) 737
d) Páginas vistas por sesión (promedio) 739
e) Nº de retuits en la cuenta del medio (promedio) 742
f) Nº de favoritos en la cuenta del medio (promedio) 744
g) Nº de retuits de la tendencia 14 días después (promedio) 746
h) Resumen de relaciones entre variables de predicción y de éxito 748
6.1.4.4. Regresión binomial negativa o de Poisson 750
a) Filtro de alta correlación (colinealidad) 750
b) Páginas únicas (total) 752
c) Nº de tuits de la tendencia 14 días después (total) 755
d) Nº de retuits de la tendencia 14 días después (total) 757
e) Resumen de relaciones entre variables de predicción y de éxito 760

6.2. API de informes de Google Analytics v4 763

6.2.1. Cuerpo de la solicitud 763
6.2.2. Cuerpo de la respuesta 764
6.2.3. Autorización 764
6.2.4. ReportRequest 765
6.2.5. DateRange 770
6.2.6. Muestreo 771
6.2.7. Dimensión 772
6.2.8. DimensionFilterClause 774

6.2.9. FilterLogicalOperator 775
6.2.10. DimensionFilter 776
6.2.11. Operador 777
6.2.12. Métrica 779
6.2.13. MetricType 780
6.2.14. MetricFilterClause 781
6.2.15. MetricFilter 782
6.2.16. Operador 783
6.2.17. OrderBy 784
6.2.18. OrderType 785
6.2.19. SortOrder 786
6.2.20. Segmento 786
6.2.21. DynamicSegment 787
6.2.22. SegmentDefinition 788
6.2.23. SegmentFilter 789
6.2.24. SimpleSegment 791
6.2.25. OrFiltersForSegment 792
6.2.26. SegmentFilterClause 793
6.2.27. SegmentDimensionFilter 794
6.2.28. Operador 795
6.2.29. SegmentMetricFilter 797
6.2.30. Alcance 798
6.2.31. Operador 799
6.2.32. SequenceSegment 800
6.2.33. SegmentSequenceStep 801
6.2.34. MatchType 801
6.2.35. Tabla dinámica 802
6.2.36. CohortGroup 805
6.2.37. Cohorte 808
6.2.38. Tipo 809
6.2.39. Informe 810
6.2.40. ColumnHeader 811
6.2.41. MetricHeader 812
6.2.42. MetricHeaderEntry 812
6.2.43. PivotHeader 813
6.2.44. PivotHeaderEntry 814
6.2.45. ReportData 815
6.2.46. ReportRow 818
6.2.47. DateRangeValues 819

6.2.48. PivotValueRegion 820
6.2.49. Parámetros de consulta estándar 820
6.2.50. Límites y cuotas en las solicitudes a la API 822
6.2.51. Respuestas de error 822
6.2.51.1. Tabla de errores 823
6.2.52. Informes de Google Ad Manager 826

6.3. Listado completo de dimensiones y métricas de la API de informes de Google Analytics 829

6.3.1. Usuario 829
6.3.1.1. Dimensiones 829
6.3.1.2. Métricas 829
6.3.2. Sesión 829
6.3.2.1. Dimensiones 829
6.3.2.2. Métricas 829
6.3.3. Fuentes de tráfico 830
6.3.3.1. Dimensiones 830
6.3.3.2. Métricas 830
6.3.4. Adwords 830
6.3.4.1. Dimensiones 830
6.3.4.2. Métricas 831
6.3.5. Conversiones de objetivos 831
6.3.5.1. Dimensiones 831
6.3.5.2. Métricas 832
6.3.6. Plataforma o dispositivo 832
6.3.6.1. Dimensiones 832
6.3.7. Red geográfica 833
6.3.7.1. Dimensiones 833
6.3.8. Sistema 833
6.3.8.1. Dimensiones 833
6.3.9. Seguimiento de página 834
6.3.9.1. Dimensiones 834
6.3.9.2. Métricas 834
6.3.10. Agrupación de contenido 834
6.3.10.1. Dimensiones 834
6.3.10.2. Métricas 835
6.3.11. Búsqueda interna 835
6.3.11.1. Dimensiones 835
6.3.11.2. Métricas 835
6.3.12. Velocidad del sitio 836

6.3.12.1. Métricas 836
6.3.13. Seguimiento de aplicaciones 836
6.3.13.1. Dimensiones 836
6.3.13.2. Métricas 837
6.3.14. Seguimiento de eventos 837
6.3.14.1. Dimensiones 837
6.3.14.2. Métricas 837
6.3.15. Comercio electrónico 837
6.3.15.1. Dimensiones 837
6.3.15.2. Métricas 838
6.3.16. Interacciones sociales 839
6.3.16.1. Dimensiones 839
6.3.16.2. Métricas 839
6.3.17. Tiempos de usuario 840
6.3.17.1. Dimensiones 840
6.3.17.2. Métricas 840
6.3.18. Excepciones 840
6.3.18.1. Dimensiones 840
6.3.18.2. Métricas 840
6.3.19. Experimentos de contenido 840
6.3.19.1. Dimensiones 840
6.3.20. Variables o columnas personalizadas 841
6.3.20.1. Dimensiones 841
6.3.20.2. Métricas 841
6.3.21. Tiempo 841
6.3.21.1. Dimensiones 841
6.3.22. DoubleClick Campaign Manager 842
6.3.22.1. Dimensiones 842
6.3.22.2. Métricas 843
6.3.23. Audiencia 843
6.3.23.1. Dimensiones 843
6.3.24. Adsense 844
6.3.24.1. Métricas 844
6.3.25. Editor 844
6.3.25.1. Métricas 844
6.3.26. Ad Exchange 844
6.3.26.1. Métricas 844
6.3.27. Reabastecimiento de DoubleClick para editores 845
6.3.27.1. Dimensiones 845

6.3.27.2. Métricas 845
6.3.28. DoubleClick para editores 845
6.3.28.1. Métricas 845
6.3.29. Valor del tiempo de vida y cohortes 846
6.3.29.1. Dimensiones 846
6.3.29.2. Métricas 846
6.3.30. Agrupación de canales 847
6.3.30.1. Dimensiones 847
6.3.31. DoubleClick Bid Manager 847
6.3.31.1. Dimensiones 847
6.3.31.2. Métricas 848
6.3.32. Búsqueda de DoubleClick 848
6.3.32.1. Dimensiones 848
6.3.32.2. Métricas 848
6.4. Twitter Standard API 850
6.4.1. Métodos de autentificación 853
6.4.2. Tweet JSON 854
6.4.2.1. Tuits 855
6.4.2.2. Tuits Extendidos 856
6.4.2.3. Retuits y citas 858
6.4.2.4. Retuits 858
6.4.2.5. Citas 860
6.4.2.6. Buenas prácticas 862
6.4.3. Objeto Tweet 862
6.4.3.1. Diccionario de datos de tuits 863
6.4.3.2. Atributos adicionales de Tuit 874
6.4.3.3. Atributos obsoletos 876
6.4.4. Objeto Usuario 876
6.4.4.1. Diccionario de datos de usuario 877
6.4.4.2. Atributos ya no admitidos (obsoletos) 881
6.4.4.3. Objeto de usuario de ejemplo 884
6.4.5. Objeto Entidades 886
6.4.5.1. Diccionario de datos de entidades 888
6.4.5.2. Objeto hashtag 895
6.4.5.3. Objeto multimedia 896
6.4.5.4. Objetos de tamaño de medio 901
6.4.5.5. Formato de URL de medios fotográficos 903
6.4.5.6. Objeto URL 907
6.4.5.7. Objeto de mención de usuario 909

6.4.5.8. Objeto símbolo 910
6.4.5.9. Objeto de encuesta 911
6.4.5.10. Retuits y Citas de Tuits 912
6.4.5.11. Retuits 913
6.4.5.12. Citas 913
6.4.5.13. Entidades para objeto de usuario 914
6.4.6. GET statuses/user_timeline 916
6.4.6.1. URL del recurso 916
6.4.6.2. Información de recursos 916
6.4.6.3. Parámetros 917
6.4.6.4. Paginación 919
6.4.7. API de búsqueda estándar 921
6.4.7.1. URL del recurso 921
6.4.7.2. Información de recursos 921
6.4.7.3. Parámetros 922
6.4.7.4. Operadores 925
6.4.7.5. Operadores independientes 926
6.4.7.6. Operadores de entidad 930
6.4.7.7. Paginación 933
6.4.8. GET followers/ids 935
6.4.8.1. URL del recurso 936
6.4.8.2. Información de recursos 936
6.4.8.3. Parámetros 936
6.4.9. Códigos de respuesta 939
6.4.9.1. Códigos de estado HTTP 939
6.4.9.2. Mensajes de error 941
6.4.9.3. Códigos de error 941
6.4.10. Límites y cuotas en las solicitudes a la API 951
6.4.10.1. Por usuario o por aplicación de desarrollador 951
6.4.10.2. Ventanas de 15 minutos 951
6.4.10.3. Encabezados HTTP y códigos de respuesta 951
6.4.10.4. Límites de las solicitudes GET y POST 952
6.4.10.5. Consejos para evitar tener una tarifa limitada 953
6.4.10.6. Streaming API 954
6.4.10.7. Patrón de retroceso exponencial para streaming 954
6.4.10.8. Límites por ventana por recurso 955

Índice de Figuras

Figura 1. Modelo de generación de noticias. Esta tesis se enmarca en la priorización de la información (Graefe, 2016).....60
Figura 2. Línea de tiempo con el marco temporal del objeto de estudio.....64
Figura 3. Distribución por medio de los medios de comunicación online en enero de 2000: 11.428 en total (Navarro Zamora, 2000). 83
Figura 4. Distribución por zonas geográficas de los medios de comunicación online en enero de 2000 (Navarro Zamora, 2000).....83
Figura 5. Distribución por continente y Estados Unidos de los periódicos online en enero de 2000: 4.322 en total (Navarro Zamora, 2000). 83
Figura 6. Estructura de funcionamiento del cibermedio (Farías de Estany & María Prieto, 2009). 84
Figura 7. Proceso de producción del contenido (Farías de Estany & María Prieto, 2009). 85
Figura 8. Tipos de periodismo online (Deuze, 2001).....88
Figura 9. Estructura de análisis longitudinal de contenidos (Deuze, 1998).....91
Figura 10. Número de enlaces a material relacionado por artículo y año (Tremayne, 2004). 92
Figura 11. Elaboración de noticias por parte de los algoritmos (Graefe, 2016). 98
Figura 12. Comparación del modelo tradicional y RTB (Gutiérrez Argüello, 2013).....101
Figura 13. Comparación entre una segmentación normal y una basada en RTB (Gutiérrez Argüello, 2013). 102
Figura 14. Muestra del formulario de envío de tuits en 2021, con la pregunta actual de "¿Qué está pasando?" 106
Figura 15. Uso de los usuarios adultos de Twitter a finales de 2018 (Wojcik & Hughes, 2019) 108
Figura 16. Correlación entre el número de mensajes y el número de seguidores (Huberman, et al., 2009).....108
Figura 17. Marco teórico de entendimiento de la relevancia de las noticias en las redes sociales (Heikkilä & Ahva, 2015). 113
Figura 18. Embudo de conversión a través del contenido social y de búsqueda (Hochuli, 2015). 116
Figura 19. Lectura de periódicos según el medio de lectura y si son usuarios de Twitter o no (Fox & Lenhart, 2009).....119
Figura 20. Proceso de minería de uso de la web (Srivastava, et al., 2000). 126

Figura 21. Arquitectura de una plataforma web de aprendizaje con almacenamientos analíticos internos y externos (Rohloff, et al., 2019) 130
Figura 22. Relación entre los Objetivos de Negocio y los Datos Digitales (Graham, 2014) 131
Figura 23. Estructura de la estrategia digital (Graham, 2014).....131
Figura 24. Métricas básicas y su descripción (Alvarez Intriago, et al., 2016).....133
Figura 25. Evolución de la Web (Spivack, 2007).....134
Figura 26. Canales de tráfico más activos en diferentes tipos de contenidos (Coleman, 2016) 135
Figura 27. Tabla de parámetros de analítica web y calidad de una web (López, et al., 2017). 135
Figura 28. Lista de KPI según el tamaño del negocio al que pertenece la web (Kaushik, 2011) 139
Figura 29. Estudio del contenido consumido en Trust según Hart (Kaushik, 2007).....142
Figura 30. Métricas del contenido social y de búsqueda (Hochuli, 2015). 143
Figura 31. Proceso recomendado para la mejora continua gracias a la analítica digital (Chaffey & Patron, 2012).....146
Figura 32. Tabla de objetivos de negocio (Kaushik, 2011). 147
Figura 33. Tabla de metas de la web (Kaushik, 2011). 147
Figura 34. Tabla de KPI (Kaushik, 2011). 148
Figura 35. Tabla de metas de KPI (Kaushik, 2011).....149
Figura 36. Tabla de segmentos a analizar (Kaushik, 2011). 150
Figura 37. Ejemplo de segmentación de visitantes por la duración de la visita y el buscador utilizado (Kaushik, 2006) 151
Figura 38. Ejemplo de comparación de páginas según su número de visitas (Kaushik, 2007) 153
Figura 39. Ejemplo de test A/B cambiando la imagen y la llamada de acción, según Quintero (2008) 155
Figura 40. Ejemplo de test multivariante cambiando la imagen y la llamada de acción, según Quintero (2008).....156
Figura 41. Métodos usados por las compañías para mejorar su ratio de conversión (Chaffey & Patron, 2012) 157
Figura 42. Clasificación de los métodos de mejora de rendimiento según su tipología y enfoque (Chaffey & Patron, 2012).....158
Figura 43. Las relaciones entre los campos de investigación de la informetría, bibliometría, cienciometría, cibermetría y webometría (Ingwersen & Björneborn, 2004, p. 339).....160

Figura 44. Modelo simplificado de relaciones entre sitios web (Almind & Ingwersen, 1997). 162
Figura 45. Interacciones de las áreas de trabajo de la cibermetría (Orduña-Malea & Aguillo, 2015). 163
Figura 46. Número de webs y de usuarios estimados según el año (Orduña-Malea & Alonso-Arroyo, 2018, p. 10).....165
Figura 47. Esquema de profundidad del recorrido de un rastreador (Sánchez-Pita & Alonso-Berrocal, 2013) 167
Figura 48. Fórmula del soporte de la relación entre una página A y una página B (Ortega Priego & Aguillo, 2009).....168
Figura 49. Fórmula de la confianza de la relación entre una página A y una página B (Ortega Priego & Aguillo, 2009).....168
Figura 50. Fórmula del PageRank (Alonso Berrocal, et al., 2008). 171
Figura 51. Reciprocidad verdadera entre dos páginas de dos webs (Danman Fugl, 2001). 171
Figura 52. Reciprocidad entre tres páginas y dos webs (Danman Fugl, 2001).....172
Figura 53. Reciprocidad entre cuatro páginas y dos webs (Danman Fugl, 2001).....172
Figura 54. Grafo para terminología básica de enlaces webométricos en el que cada letra representa un nodo web diferente (Ingwersen & Björneborn, 2004, p. 344).....173
Figura 55. Fórmula para calcular la densidad de enlaces de una red (Sánchez-Pita & Alonso-Berrocal, 2013).....174
Figura 56. Diagrama simplificado de nodos web de un sitio web contenedor de subsitios web y sub-subsitios web (Björneborn & Ingwersen, 2004).....175
Figura 57. Dispersión web de tipo interna y externa (Orduña-Malea & Alonso-Arroyo, 2018, p. 39). 176
Figura 58. Análisis SWOT según el origen y si ayuda o no a obtener el objetivo (Lavinsky, 2013). 180
Figura 59. Las relaciones entre los campos de investigación de la informetría, bibliometría, cienciometría, cibermetría, webmetría y, también, altmetría (Manzano Zambruno, et al., 2019, p. 208). 185
Figura 60. Usuarios activos y conservados (Java, et al., 2009). 190
Figura 61. Fórmulas de autoridad y concentrador (hub) (Java, et al., 2009). 191
Figura 62. Alice es el usuario con más retuits, por lo que es la que tiene una centralidad de grado mayor (Kumar, et al., 2014, p. 35).....192
Figura 63. Bob y Gary son los nodos que aportan la información más retuiteada, por lo que son los de mayor centralidad de vector propio (Kumar, et al., 2014). 193

Figura 64. Alice es la que aporta la información a más usuarios en un camino más corto, por lo que tiene la mayor centralidad del camino más corto (Kumar, et al., 2014, p. 35). 193
Figura 65. Grafo Usuario-Tuit (Yamaguchi, et al., 2010).....195
Figura 66. Fórmula de la desviación estándar de un hashtag (Huang, et al., 2010).....196
Figura 67. Fórmula del sesgo de un hashtag (Huang, et al., 2010). 196
Figura 68. Fórmula de la curtosis de un hashtag (Huang, et al., 2010). 197
Figura 69. Valor económico en los diferentes canales (Kaushik, 2011).....201
Figura 70. Gráfica de mediciones de redes sociales según Ohlen (Kaushik, 2011). 201
Figura 71. Fórmula de la métrica Klout (Kaushik, 2011). 202
Figura 72. Métricas de redes sociales según Klout (Kaushik, 2011).....202
Figura 73. Tabla de indicadores (Alvarez Intriago, et al., 2016). 204
Figura 74. Tabla de indicadores de uso según su estrategia en Twitter y Facebook (López, et al., 2017).....205
Figura 75. Universo de acciones respecto de una cuenta de Twitter, según Navas (2018) 205
Figura 76. Ejemplo de modelo de grafos de hashtags (Wang, et al., 2011). 208
Figura 77. Ejemplo de clasificación mejorada según el significado literal de los hashtags (Wang, et al., 2011). 209
Figura 78. Ejemplo de frecuencia de intercambio de información entre nodos (Kossinets, et al., 2008).....212
Figura 79. Flujo de trabajo de la técnica de minería de datos (Azam, et al., 2015). 216
Figura 80. Procedimiento de la detección de tendencias (Gaglio, et al., 2016). 218
Figura 81. Nº de tuits por hora en las primeras horas tras la muerte de Michael Jackson (Sankaranarayanan, et al., 2009). 220
Figura 82. Arquitectura del sistema de TEDAS (Li, et al., 2012).....221
Figura 83. Cambio repentino en la correlación entre dos etiquetas (Alvanaki, et al., 2011). 221
Figura 84. Comparativa de comunicación según su fuente y si se filtra o no su semántica (Okazaki & Matsuo, 2008) 222
Figura 85. Detección, seguimiento y visualización de eventos en Twitter (Cai, et al., 2015). 223
Figura 86. Propuesta de predicción de eventos (Kursuncu, et al., 2019).....224
Figura 87. Modelo de comunicación online (Kiani, 1998).....228
Figura 88. Formatos IAB de banners de publicidad (Oliva Rodríguez, 2014). 231

Figura 89. Modelo de asignación simple de anuncios a páginas web (Aggarwal, et al., 1998). 236
Figura 90. Distribución de audiencia expuesta a los anuncios e impresiones de éstos según el número de exposiciones (Bruner & Gluck, 2006). 237
Figura 91. CTR según el tamaño de los anuncios (Bruner & Gluck, 2006).....237
Figura 92. Intención de compra según el tamaño de los anuncios (Bruner & Gluck, 2006). 238
Figura 93. CTR según el tipo de anuncio (Bruner & Gluck, 2006). 238
Figura 94. Intención de compra según el formato de anuncio (Bruner & Gluck, 2006)...239
Figura 95. Colocación de anuncios en lugares vacíos de una página web (Li, et al., 2010). 241
Figura 96. Formatos de anuncio por tamaño de Google AdSense (Li, et al., 2010). 242
Figura 97. Porcentaje de webs con espacio en blanco, en franjas de 10.000 píxeles (Li, et al., 2010).....242
Figura 98. Gráfico de usuario del General Click Model propuesto por Zhu et al. (2010) 244
Figura 99. Modelo de factores que fomentan la intención de compra (Ka Po, 2006). 245
Figura 100. Esquema de técnicas involucradas en el sistema AD ROSA (Kazienko & Adamski, 2004).....250
Figura 101. Modelo de arquitectura de un sistema de entrega de anuncios (Vratonjic, et al., 2010). 251
Figura 102. Diagrama de conexiones entre Google AdWords y AdSense (Desnica, et al., 2014). 252
Figura 103. Plataforma de publicidad contextual (Wu, et al., 2013).....253
Figura 104. Ecosistema simplificado de la publicidad en Internet (Yuan, et al., 2012). ..254
Figura 105. Captura de la página principal de la web de Hello Friki, realizada el día 2 de junio de 2021 263
Figura 106. Esquema de los componentes de Google Analytics (Google Developers, s.f.) 282
Figura 107. Estructura de procesos de la recogida de datos en la fase 1. Fuente: Elaboración propia.....296
Figura 108. Línea de tiempo de la toma de datos de un artículo publicado el 23 de septiembre de 2020 299
Figura 109. Logo de STATGRAPHICS Centurion XVII 301
Figura 110. Ejemplo de gráfico de Caja y Bigotes (Anon., 2005).....302
Figura 111. Ejemplo de gráfico de Probabilidad Normal (Anon., 2007) 302
Figura 112. Ejemplo de Matriz de correlaciones Pearson.....303

Figura 113. Ejemplo de gráfico de Componentes Principales 2D (Anon., 2007) 304
Figura 114. Ejemplo de gráfico Observado contra Predicho (Anon., 2006) 304
Figura 115. Ejemplo de gráfico de frecuencia de búsquedas en Google Trends (Anon., 2020) 305
Figura 116. Todos: Gráfico de Caja y Bigotes para el valor uniquepageviews_total 327
Figura 117. Análisis del término HeroQuest en Google vía Google Trends en el periodo analizado 328
Figura 118. Análisis del término Amazon Prime en Google vía Google Trends en el periodo analizado 328
Figura 119. Análisis del término Craig Rosenberg en Google vía Google Trends en el periodo analizado 328
Figura 120. Análisis del término The Boys en Google vía Google Trends en el periodo analizado 328
Figura 121. Análisis del término Raised by Wolves en Google vía Google Trends en el periodo analizado 329
Figura 122. Análisis del término Animales fantásticos 3 en Google vía Google Trends en el periodo analizado 329
Figura 123. Análisis del término Eddie Redmayne en Google vía Google Trends en el periodo analizado 329
Figura 124. Análisis del término Conan en Google vía Google Trends en el periodo analizado 329
Figura 125. Análisis del término Netflix en Google vía Google Trends en el periodo analizado 330
Figura 126. Análisis del término the CW en Google vía Google Trends en el periodo analizado 330
Figura 127. Análisis del término the Flash en Google vía Google Trends en el periodo analizado 330
Figura 128. Análisis del término Black Adam en Google vía Google Trends en el periodo analizado 331
Figura 129. Análisis del término Dune en Google vía Google Trends en el periodo analizado 331
Figura 130. Análisis del término Matrix 4 en Google vía Google Trends en el periodo analizado 331
Figura 131. Análisis del término Minecraft en Google vía Google Trends en el periodo analizado 331

Figura 132. Análisis del término The Batman en Google vía Google Trends en el periodo analizado 331
Figura 133. Análisis del término Warner Bros. en Google vía Google Trends en el periodo analizado 332
Figura 134. Análisis del término Carly Mensch en Google vía Google Trends en el periodo analizado 332
Figura 135. Análisis del término Glow en Google vía Google Trends en el periodo analizado 332
Figura 136. Análisis del término Liz Flahive en Google vía Google Trends en el periodo analizado 332
Figura 137. Análisis del término Ecc Ediciones en Google vía Google Trends en el periodo analizado 333
Figura 138. Análisis del término noviembre 2020 en Google vía Google Trends en el periodo analizado 333
Figura 139. Análisis del término Benedict Cumberbatch en Google vía Google Trends en el periodo analizado 333
Figura 140. Análisis del término Doctor Strange en Google vía Google Trends en el periodo analizado 333
Figura 141. Análisis del término Marvel en Google vía Google Trends en el periodo analizado 333
Figura 142. Análisis del término The Amazing Spider-Man 3 en Google vía Google Trends en el periodo analizado 334
Figura 143. Análisis del término Nacon en Google vía Google Trends en el periodo analizado 334
Figura 144. Análisis del término PC en Google vía Google Trends en el periodo analizado 334
Figura 145. Análisis del término Revolution en Google vía Google Trends en el periodo analizado 334
Figura 146. Análisis del término Windows en Google vía Google Trends en el periodo analizado 334
Figura 147. Análisis del término Xbox en Google vía Google Trends en el periodo analizado 335
Figura 148. Análisis del término Buenos días tristeza en Google vía Google Trends en el periodo analizado 335
Figura 149. Análisis del término Planeta Cómic en Google vía Google Trends en el periodo analizado 335

Figura 150. Análisis del término Tortugas Ninja en Google vía Google Trends en el periodo analizado 335
Figura 151. Análisis del término Jurassic World Evolution: Complete Edition en Google vía Google Trends en el periodo analizado.....336
Figura 152. Análisis del término Nintendo Switch en Google vía Google Trends en el periodo analizado 336
Figura 153. Análisis del término Sobrenatural en Google vía Google Trends en el periodo analizado 336
Figura 154. Análisis del término Supernatural en Google vía Google Trends en el periodo analizado 336
Figura 155. Análisis del término diciembre 2020 en Google vía Google Trends en el periodo analizado 337
Figura 156. Análisis del término Saga en Google vía Google Trends en el periodo analizado 337
Figura 157. Análisis del término Wonder Girl en Google vía Google Trends en el periodo analizado 337
Figura 158. Todos: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean 338
Figura 159. Todos: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean 339
Figura 160. Todos: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean 340
Figura 161. Todos: Gráfico de Caja y Bigotes para el valor retweet_count_mean 341
Figura 162. Todos: Gráfico de Caja y Bigotes para el valor favorite_count_mean 342
Figura 163. Todos: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets 343
Figura 164. Todos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total 344
Figura 165. Todos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean 345
Figura 166. Todos: Gráfico de probabilidad normal de la variable log(uniquepageviews_total).....349
Figura 167. Todos: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean).....349
Figura 168. Todos: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean) 350
Figura 169. Todos: Gráfico de probabilidad normal de la variable log(favorite_count_mean) 350

Figura 170. Todos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean).....351
Figura 171. Todos: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente.....352
Figura 172. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets.....354
Figura 173. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total 354
Figura 174. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean 355
Figura 175. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total 356
Figura 176. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean 357
Figura 177. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate 358
Figura 178. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean.....359
Figura 179. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean.....360
Figura 180. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean 361
Figura 181. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate 362
Figura 182. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean) 366
Figura 183. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean) 367
Figura 184. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean).....367
Figura 185. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate).....368
Figura 186. Todos: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente.....369
Figura 187. Todos: Gráfica de pesos de cada componente principal 372
Figura 188. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1 375

Figura 189. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1 377
Figura 190. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1 ...379
Figura 191. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1 381
Figura 192. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1 383
Figura 193. Todos: Matriz de correlaciones Pearson entre las variables de predicción .385
Figura 194. Todos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1 390
Figura 195. Todos: Gráfico de la relación entre la función de predicción según la regresión de binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1 393
Figura 196. Logo del congreso COMRED 2021 427
Figura 197. Logo del congreso CIMED 2021 428
Figura 198. Cine: Gráfico de Caja y Bigotes para el valor uniquepageviews_total.....494
Figura 199. Cine: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean.....495
Figura 200. Cine: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean.....496
Figura 201. Cine: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean 497
Figura 202. Cine: Gráfico de Caja y Bigotes para el valor retweet_count_mean.....498
Figura 203. Cine: Gráfico de Caja y Bigotes para el valor favorite_count_mean.....499
Figura 204. Cine: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets.....500
Figura 205. Cine: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total 501
Figura 206. Cine: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean 502
Figura 207. Cine: Gráfico de probabilidad normal de la variable log(uniquepageviews_total) 506
Figura 208. Cine: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean) 506
Figura 209. Cine: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean) 507
Figura 210. Cine: Gráfico de probabilidad normal de la variable log(favorite_count_mean) 507

Figura 211. Cine: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total) 508
Figura 212. Cine: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean) 508
Figura 213. Cine: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente 509
Figura 214. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets 511
Figura 215. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total 512
Figura 216. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean 513
Figura 217. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total 514
Figura 218. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean 515
Figura 219. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate 516
Figura 220. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean 517
Figura 221. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean 518
Figura 222. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean 519
Figura 223. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate 520
Figura 224. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean) 524
Figura 225. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean) 525
Figura 226. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate) 525
Figura 227. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean) 526
Figura 228. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean) 526
Figura 229. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate) 527

Figura 230. Cine: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente 528
Figura 231. Cine: Gráfica de pesos de cada componente principal.....531
Figura 232. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1.....534
Figura 233. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1.....536
Figura 234. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1.....538
Figura 235. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1.....540
Figura 236. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1.....542
Figura 237. Cine: Matriz de correlaciones Pearson entre las variables de predicción....544
Figura 238. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1 548
Figura 239. Cine: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1 550
Figura 240. Cine: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1553
Figura 241. Series: Gráfico de Caja y Bigotes para el valor uniquepageviews_total.....556
Figura 242. Series: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean 557
Figura 243. Series: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean 558
Figura 244. Series: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean 559
Figura 245. Series: Gráfico de Caja y Bigotes para el valor retweet_count_mean 560
Figura 246. Series: Gráfico de Caja y Bigotes para el valor favorite_count_mean 561
Figura 247. Series: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets 562
Figura 248. Series: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total 563
Figura 249. Series: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean 564
Figura 250. Series: Gráfico de probabilidad normal de la variable log(uniquepageviews_total).....567
Figura 251. Series: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean).....568

Figura 252. Series: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean) 568
Figura 253. Series: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean) 569
Figura 254. Series: Gráfico de probabilidad normal de la variable retweet_count_mean 569
Figura 255. Series: Gráfico de probabilidad normal de la variable log(favorite_count_mean) 570
Figura 256. Series: Gráfico de probabilidad normal de la variable terms_end_num_tweets 570
Figura 257. Series: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total) 571
Figura 258. Series: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean) 571
Figura 259. Series: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente 572
Figura 260. Series: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets 574
Figura 261. Series: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total 575
Figura 262. Series: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean 576
Figura 263. Series: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total 577
Figura 264. Series: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean 578
Figura 265. Series: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate 579
Figura 266. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean 580
Figura 267. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean 581
Figura 268. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean 582
Figura 269. Series: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate 583
Figura 270. Series: Gráfico de probabilidad normal de la variable terms_ini_num_tweets 587

Figura 271. Series: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean) 588
Figura 272. Series: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean) 588
Figura 273. Series: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate) 589
Figura 274. Series: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean).....589
Figura 275. Series: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean) 590
Figura 276. Series: Gráfico de probabilidad normal de la variable terms_ini_user_age_mean 590
Figura 277. Series: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate).....591
Figura 278. Series: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente 592
Figura 279. Series: Gráfica de pesos de cada componente principal 595
Figura 280. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1 598
Figura 281. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1 602
Figura 282. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1 ...604
Figura 283. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1 607
Figura 284. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de terms_end_num_tweets en la fase 1.....610
Figura 285. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1 612
Figura 286. Series: Matriz de correlaciones Pearson entre las variables de predicción.615
Figura 287. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1 618
Figura 288. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1 621

Figura 289. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1623
Figura 290. Videojuegos: Gráfico de Caja y Bigotes para el valor uniquepageviews_total 626
Figura 291. Videojuegos: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean 627
Figura 292. Videojuegos: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean 628
Figura 293. Videojuegos: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean 629
Figura 294. Videojuegos: Gráfico de Caja y Bigotes para el valor retweet_count_mean630
Figura 295. Videojuegos: Gráfico de Caja y Bigotes para el valor favorite_count_mean631
Figura 296. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets 632
Figura 297. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total 633
Figura 298. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean 634
Figura 299. Videojuegos: Gráfico de probabilidad normal de la variable log(uniquepageviews_total).....637
Figura 300. Videojuegos: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean) 638
Figura 301. Videojuegos: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean).....638
Figura 302. Videojuegos: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean) 639
Figura 303. Videojuegos: Gráfico de probabilidad normal de la variable retweet_count_mean 639
Figura 304. Videojuegos: Gráfico de probabilidad normal de la variable favorite_count_mean 640
Figura 305. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_num_tweets) 640
Figura 306. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total) 641
Figura 307. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean).....641

Figura 308. Videojuegos: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente 642
Figura 309. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets 645
Figura 310. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total.....646
Figura 311. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean 647
Figura 312. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total.....648
Figura 313. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean 649
Figura 314. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate 650
Figura 315. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean.....651
Figura 316. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean.....652
Figura 317. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean 653
Figura 318. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate.....654
Figura 319. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_num_tweets) 658
Figura 320. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean) 658
Figura 321. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_total) 659
Figura 322. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean) 659
Figura 323. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate) 660
Figura 324. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean).....660
Figura 325. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean) 661

Figura 326. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_age_mean).....661
Figura 327. Videojuegos: Gráfico de probabilidad normal de la variable terms_ini_url_inclusion_rate.....662
Figura 328. Videojuegos: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente.....663
Figura 329. Videojuegos: Gráfica de pesos de cada componente principal 666
Figura 330. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1 669
Figura 331. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_num_tweets) en la fase 1 677
Figura 332. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1.....679
Figura 333. Videojuegos: Matriz de correlaciones Pearson entre las variables de predicción 681
Figura 334. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1 685
Figura 335. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1 687
Figura 336. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1.....690
Figura 337. Tráileres: Gráfico de Caja y Bigotes para el valor uniquepageviews_total ..692
Figura 338. Tráileres: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean ...693
Figura 339. Tráileres: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean ...694
Figura 340. Tráileres: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean 695
Figura 341. Tráileres: Gráfico de Caja y Bigotes para el valor retweet_count_mean 696
Figura 342. Tráileres: Gráfico de Caja y Bigotes para el valor favorite_count_mean 697
Figura 343. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets 698

Figura 344. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total 699
Figura 345. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean 700
Figura 346. Tráileres: Gráfico de probabilidad normal de la variable log(uniquepageviews_total) 704
Figura 347. Tráileres: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean) 704
Figura 348. Tráileres: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean) 705
Figura 349. Tráileres: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean) 705
Figura 350. Tráileres: Gráfico de probabilidad normal de la variable log(retweet_count_mean) 706
Figura 351. Tráileres: Gráfico de probabilidad normal de la variable log(favorite_count_mean) 706
Figura 352. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_num_tweets) 707
Figura 353. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total) 707
Figura 354. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean) 708
Figura 355. Tráileres: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente 709
Figura 356. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets...711
Figura 357. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total 712
Figura 358. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean 713
Figura 359. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total 714
Figura 360. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean 715
Figura 361. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate 716

Figura 362. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean 717
Figura 363. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean 718
Figura 364. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean 719
Figura 365. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate 720
Figura 366. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_num_tweets) 724
Figura 367. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean) 725
Figura 368. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_total) 725
Figura 369. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean) 726
Figura 370. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate) 726
Figura 371. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean) 727
Figura 372. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean) 727
Figura 373. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate) 728
Figura 374. Tráileres: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente 729
Figura 375. Tráileres: Gráfica de pesos de cada componente principal 732
Figura 376. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1 735
Figura 377. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(adsense_ecpm_mean) en la fase 1 737
Figura 378. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1 739
Figura 379. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1 ... 741

Figura 380. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(retweet_count_mean) en la fase 1 743
Figura 381. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1 745
Figura 382. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1 748
Figura 383. Tráileres: Matriz de correlaciones Pearson entre las variables de predicción 751
Figura 384. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1 755
Figura 385. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1 757
Figura 386. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1.....760
Figura 387. Tabla de características de los tipos de API de Twitter (Twitter Developers, s.f.) 851

Índice de Tablas

Tabla 1.....63
Tabla 2.....313
Tabla 3.....314
Tabla 4.....314
Tabla 5.....315
Tabla 6.....316
Tabla 7.....321
Tabla 8.....345
Tabla 9.....346
Tabla 10.....347
Tabla 11.....348
Tabla 12.....352
Tabla 13.....362
Tabla 14.....363
Tabla 15.....365
Tabla 16.....365
Tabla 17.....369
Tabla 18.....370
Tabla 19.....372
Tabla 20.....373
Tabla 21.....374
Tabla 22.....375
Tabla 23.....376
Tabla 24.....377
Tabla 25.....378
Tabla 26.....379
Tabla 27.....380
Tabla 28.....381
Tabla 29.....382
Tabla 30.....384
Tabla 31.....386
Tabla 32.....387
Tabla 33.....388
Tabla 34.....389

Tabla 35....390
Tabla 36....391
Tabla 37....392
Tabla 38....394
Tabla 39....396
Tabla 40....398
Tabla 41....401
Tabla 42....401
Tabla 43....403
Tabla 44....403
Tabla 45....405
Tabla 46....405
Tabla 47....407
Tabla 48....407
Tabla 49....409
Tabla 50....410
Tabla 51....412
Tabla 52....502
Tabla 53....503
Tabla 54....504
Tabla 55....505
Tabla 56....509
Tabla 57....520
Tabla 58....521
Tabla 59....523
Tabla 60....523
Tabla 61....528
Tabla 62....530
Tabla 63....531
Tabla 64....532
Tabla 65....533
Tabla 66....534
Tabla 67....535
Tabla 68....536
Tabla 69....537
Tabla 70....538

Tabla 71.....539
Tabla 72.....540
Tabla 73.....541
Tabla 74.....543
Tabla 75.....545
Tabla 76.....546
Tabla 77.....547
Tabla 78.....548
Tabla 79.....549
Tabla 80.....551
Tabla 81.....552
Tabla 82.....554
Tabla 83.....564
Tabla 84.....565
Tabla 85.....566
Tabla 86.....573
Tabla 87.....583
Tabla 88.....584
Tabla 89.....586
Tabla 90.....586
Tabla 91.....592
Tabla 92.....594
Tabla 93.....595
Tabla 94.....596
Tabla 95.....597
Tabla 96.....598
Tabla 97.....599
Tabla 98.....600
Tabla 99.....601
Tabla 100.....602
Tabla 101.....603
Tabla 102.....604
Tabla 103.....605
Tabla 104.....606
Tabla 105.....606
Tabla 106.....608

Tabla 107.....609
Tabla 108.....610
Tabla 109.....611
Tabla 110.....613
Tabla 111.....615
Tabla 112.....616
Tabla 113.....617
Tabla 114.....619
Tabla 115.....620
Tabla 116.....621
Tabla 117.....622
Tabla 118.....624
Tabla 119.....634
Tabla 120.....635
Tabla 121.....636
Tabla 122.....643
Tabla 123.....654
Tabla 124.....655
Tabla 125.....657
Tabla 126.....664
Tabla 127.....665
Tabla 128.....667
Tabla 129.....668
Tabla 130.....668
Tabla 131.....670
Tabla 132.....671
Tabla 133.....671
Tabla 134.....672
Tabla 135.....672
Tabla 136.....673
Tabla 137.....674
Tabla 138.....674
Tabla 139.....675
Tabla 140.....676
Tabla 141.....677
Tabla 142.....678

Tabla 143.....680
Tabla 144.....682
Tabla 145.....683
Tabla 146.....684
Tabla 147.....685
Tabla 148.....686
Tabla 149.....688
Tabla 150.....689
Tabla 151.....691
Tabla 152.....700
Tabla 153.....701
Tabla 154.....702
Tabla 155.....703
Tabla 156.....710
Tabla 157.....720
Tabla 158.....721
Tabla 159.....723
Tabla 160.....723
Tabla 161.....729
Tabla 162.....731
Tabla 163.....732
Tabla 164.....733
Tabla 165.....734
Tabla 166.....735
Tabla 167.....736
Tabla 168.....737
Tabla 169.....738
Tabla 170.....740
Tabla 171.....740
Tabla 172.....742
Tabla 173.....742
Tabla 174.....744
Tabla 175.....744
Tabla 176.....746
Tabla 177.....747
Tabla 178.....749

Tabla 179.....752
Tabla 180.....753
Tabla 181.....753
Tabla 182.....755
Tabla 183.....756
Tabla 184.....758
Tabla 185.....758
Tabla 186.....761

Resumen

El objetivo de estudio es el diseño de una metodología cibermétrica para medir el éxito de los contenidos publicados en un medio de comunicación online y su posible predicción, de manera que se pueda orientar la optimización de los futuros contenidos publicados por el medio.

Enmarcada en el ámbito del periodismo digital, responde a la necesidad de analizar el éxito de los contenidos web de manera que se pueda ayudar en la toma de decisiones del equipo editorial.

Para ello, se ha realizado un extenso estudio de las publicaciones académicas versadas en las diferentes disciplinas que tienen lugar en esta tesis: la comunicación de contenidos digitales, Twitter, la difusión de las noticias en Twitter, la analítica web, la cibermetría, la analítica en Twitter, el análisis de tendencias en Twitter y la publicidad web.

Con dicho marco, se ha obtenido información valiosa para la optimización futura de los contenidos digitales, ya sea procedente del análisis estadístico de los datos o de la posible predicción de los indicadores de éxito de mayor interés para el medio. De esta manera, se podría integrar de manera retroalimentada en la estrategia de contenidos y favorecer así su optimización iterativamente.

Para ello, se han tenido en cuenta los siguientes objetivos específicos: investigar el concepto de éxito en el periodismo digital, la red social Twitter, la analítica web y la publicidad en la web; diseñar la metodología y determinar qué herramientas y reportes son necesarios; extraer y procesar los datos para su análisis estadístico; realizar regresiones que permitan obtener ecuaciones de predicción de las variables de éxito seleccionadas; y validar las ecuaciones de predicción con datos de test y obtener su precisión, sirviendo esta como grado de confianza en la predicción.

El diseño de la metodología ha servido para observar una sobre dispersión significativa en los datos, así como demostrar que el éxito de un contenido web tiene un carácter fuertemente multifactorial, lo cual provoca una disminución en la variabilidad calculada mediante los indicadores propuestos por investigaciones previas.

Esta tesis sirve, entonces, como base para una línea de investigación sobre la optimización de contenido digital basándose en la predicción estadística de su éxito.

Palabras clave

periodismo digital, ciberperiodismo, cibermetría, altmetría, analítica web, analítica en Twitter, análisis de tendencias, predicción de tendencias, publicidad web, Google AdSense, optimización de contenidos digitales, aprendizaje automático

Resum

L'objectiu d'estudi és el disseny d'una metodologia cibermètrica per a mesurar l'èxit dels continguts publicats en un mitjà de comunicació en línia i la seua possible predicció, de manera que es puga orientar l'optimització dels futurs continguts publicats pel mitjà.

Emmarcada en l'àmbit del periodisme digital, respon a la necessitat d'analitzar l'èxit dels continguts web de manera que es puga ajudar en la presa de decisions de l'equip editorial.

Per a això, s'ha realitzat un extens estudi de les publicacions acadèmiques versades en les diferents disciplines que tenen lloc en aquesta tesi: la comunicació de continguts digitals, Twitter, la difusió de les notícies en Twitter, l'analítica web, la cibermetría, l'analítica en Twitter, l'anàlisi de tendències en Twitter i la publicitat web.

Amb aquest marc, s'ha obtingut informació valuosa per a l'optimització futura dels continguts digitals, ja siga procedent de l'anàlisi estadística de les dades o de la possible predicció dels indicadors d'èxit de major interés per al mitjà. D'aquesta manera, es podria integrar de manera retroalimentada en l'estratègia de continguts i afavorir així la seua optimització iterativament.

Per a això, s'han tingut en compte els següents objectius específics: investigar el concepte d'èxit en el periodisme digital, la xarxa social Twitter, l'analítica web i la publicitat en la web; dissenyar la metodologia i determinar quines eines i reportes són necessaris; extraure i processar les dades per a la seua anàlisi estadística; realitzar regressions que permeten obtindre equacions de predicció de les variables d'èxit seleccionades; i validar les equacions de predicció amb dades de test i obtindre la seua precisió, servint aquesta com a grau de confiança en la predicció.

El disseny de la metodologia ha servit per a observar una sobre dispersió significativa en les dades, així com demostrar que l'èxit d'un contingut web té un caràcter fortament multifactorial, la qual cosa provoca una disminució en la variabilitat calculada mitjançant els indicadors proposats per investigacions prèvies.

Aquesta tesi serveix, llavors, com a base per a una línia d'investigació sobre l'optimització de contingut digital basant-se en la predicció estadística del seu èxit.

Paraules clau

periodisme digital, ciberperiodisme, cibermetría, altmetría, analítica web, analítica en Twitter, anàlisi de tendències, predicció de tendències, publicitat web, Google AdSense, optimització de continguts digitals, aprenentatge automàtic

Abstract

The object of this study is the design of a cybermetric methodology whose objectives are to measure the success of the content published in an online media and the possible prediction of the selected success variables.

Framed in the field of digital journalism, it responds to the need to analyze the success of web content so that it can help in the decision-making of the editorial team of a digital medium. A line of research focused on the content itself, providing an innovative vision to that of previous research, and a methodology that serves as a basis for future scientific advances.

It is about the contribution of valuable information, either from the statistical analysis of the data or from the possible prediction of the success indicators of greatest interest to the environment. In this way, it could be integrated as a feedback into the content strategy and thus favor its iterative optimization.

The main objective, therefore, is the design of a cybermetric methodology for calculating the success of an online publication, having as specific objectives: to research the concept of success in digital journalism, the social network Twitter, web analytics and web advertising; design the methodology and determine what tools and reports are needed; extract and process data for statistical analysis; perform regressions that allow to obtain prediction equations of the selected success variables; and validate the prediction equations with test data and obtain their precision, serving this as a degree of confidence in the prediction.

The design of the methodology has served to observe a significant over-dispersion in the data, as well as to demonstrate that the success of a web content has a strongly multifactorial nature, which causes a decrease in the variability calculated using the indicators proposed by previous research.

This thesis serves, then, as the basis for a very interesting research framework both at an academic and business level: the prediction of the success of digital content.

Keywords

digital journalism, cyberjournalism, cybermetrics, altmetrics, web analytics, Twitter analytics, trend analysis, trend prediction, web advertising, Google AdSense, digital content optimization, machine learning

1. Introducción

1.1. Objeto de estudio

El objeto de estudio consiste en los contenidos web publicados por un medio digital, que serán analizados desde el punto de vista de la analítica web, la analítica de la publicidad web, la analítica de Twitter y el análisis de tendencias con el objetivo de estudio de tratar de predecir su éxito.

Esta tesis se enmarca, por tanto, en el campo del ciberperiodismo o periodismo digital, que consiste según Diaz Noci (2008) en la publicación de información de actualidad mediante medios online a través de textos periodísticos que, al fin y al cabo, pretenden representar la realidad, pero sacando partido a la hipertextualidad, multimedialidad, interactividad, frecuencia de actualización y forma del contenido de la red.

Esta propuesta surge de la necesidad de analizar el posible éxito de los contenidos que se podrían publicar, de manera que pueda ayudar en la toma de decisiones de la estrategia de creación de contenidos. Se basa tanto en los datos obtenidos de los contenidos ya publicados como en las tendencias que surgen de las necesidades de información del público objetivo. Con todo ello, el objetivo es tratar de pre-calcular el éxito de un contenido con un nivel de confianza suficiente que permita asignarle así la prioridad correspondiente de los recursos del medio online.

Este requerimiento es genérico en cualquier ámbito dentro del periodismo digital, puesto que afecta a la asignación de recursos que se produce continuamente en el transcurso del ritmo de publicación de un medio digital. Ya no solo por el interés en aumentar la audiencia y la calidad del consumo de información, como señala Schudson (2016, p. 98), sino también por la necesidad de conocer cómo se produce esta en un periodismo cada vez más variado.

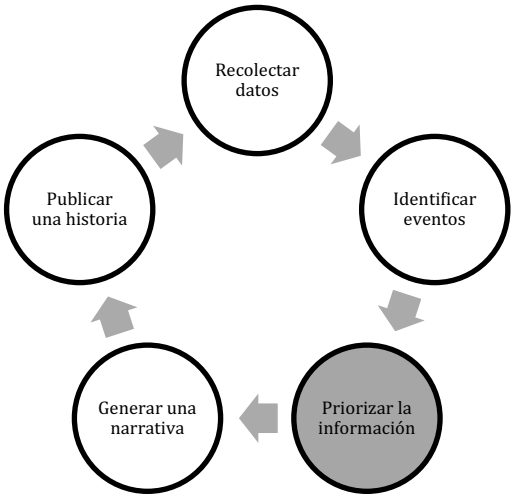


Figura 1. Modelo de generación de noticias. Esta tesis se enmarca en la priorización de la información (Graefe, 2016).

Para validar la metodología, el presente trabajo utiliza el caso de uso de un sitio web en el que se publican artículos y programas de radio online. De esta manera, se puede acceder a contenidos y servicios relacionados con un conjunto concreto de temáticas. Pretende cumplir con la necesidad del usuario de buscar información específica sobre una temática concreta y poder interactuar con otros usuarios que tengan los mismos intereses. Los contenidos se muestran de manera estructurada, organizados en categorías de artículos para facilitar la navegación y búsqueda.

Puesto que un medio online puede servir contenidos de unas temáticas específicas, también puede adaptarse a estas a la hora de elaborar una estrategia de marketing y conseguir fuentes de ingresos que faciliten la viabilidad del proyecto. Las fuentes de ingresos pueden ser de diversa índole: desde artículos patrocinados, anuncios en bloques de publicidad, enlaces de texto dentro de los contenidos, cortes de audio en los programas... e incluso algunos medios de comunicación online han optado por ofrecer determinados contenidos a los usuarios que tengan una cuenta de pago.

Las temáticas mencionadas anteriormente cuentan con una amplia gama de medios web que compiten por los primeros resultados en los buscadores, así como por ser los primeros en servir los contenidos más innovadores e inmediatos, ya que según King (2008, pp. 111-112) la mayor parte de las visitas son atraídas por las novedades de última hora, aumentando así su visibilidad y, probablemente, sus ingresos. El mismo autor asevera que este tipo de contenidos cuenta con la ventaja de responder ante tendencias y reacciones repetitivas en las redes sociales, aunque presentan un decrecimiento de atención muy rápido tras su publicación. Debido a este tipo de consumo, resulta primordial optimizar el

alcance de los contenidos web en los diferentes canales disponibles para que los usuarios accedan a ellos antes que en los competidores.

Para plantear la hipótesis en un caso de uso público y activo, se ha elegido a Twitter[1] como canal intermediario, a través del cual se pondrá en práctica la metodología cibermétrica que se propone en esta tesis. Twitter es una red social que permite una gran libertad a la hora de automatizar el proceso de publicación de contenidos, analizarlos y obtener reacciones inmediatas y cuantificables ante ellos. Laura Gómez, gerente de internacionalización de Twitter, especifica que no se trata de una red social sino de una red de información abierta, ya que lejos de mantener una relación mutua, se trata del consumo de información en una plataforma en la que en 2021 tiene 353 millones de usuarios activos mensualmente, de los cuales 187 acceden a Twitter diariamente (Dean, 2021).

Por ello, esta tesis trata de plantear una metodología cibermétrica para el cálculo del éxito de los contenidos web, de manera que se pueda optimizar la estrategia de contenidos según el posible éxito que se haya calculado de los diferentes contenidos a publicar, basándose en el éxito de los artículos ya publicados y de las temáticas relacionadas. Se sirve para ello de una serie de indicadores y fuentes estadísticas tanto de la analítica web como de la cuenta de Twitter de un medio digital, con el objetivo de cuantificar y optimizar el proceso de creación de contenidos.

La analítica web incluye datos extraídos de las visitas de la web en sí y su experiencia de navegación. Para optimizarlas, es necesario seguir un modelo de medición y de marketing digital, es decir, un conjunto de medidas objetivas con las cuales identificar el éxito o el fracaso. Dicho modelo consiste para King (2008, pp. 112-113) en definir una serie de objetivos y metas, medir unos indicadores clave de rendimiento, elegir los parámetros de éxito e identificar los segmentos a estudiar. Además, también es necesario experimentar con factores que puedan aumentar las ratios de conversión y aplicar el modelo de medición para comprobar si dichas modificaciones han sido beneficiosas o no.

La analítica en Twitter consiste en una serie de indicadores y métricas relacionados con la visibilidad, la influencia, la interacción, la fidelización y la popularidad en dicha red social, siendo esta un canal externo a la web. Según Castelló Martínez et al. (2014), pretende estudiar a través de ratios de conversación y amplificación el éxito de un contenido más allá de la navegación en sí de la web. Albishry et al. (2018) añade que su optimización

[1] https://twitter.com/

también se basa en un modelo de medición y experimentación, intentando asimismo mejorar el alcance a través de un estudio de la competencia y las fuentes de autoridad.

Este análisis estadístico pretende medir el éxito de una publicación para, así, definir un proceso de publicación de contenidos online de carácter retroalimentado. La experimentación y la creación de nuevos contenidos da lugar a un proceso de medición que incluye analítica interna y externa de la web del medio, que pretende analizar si dichos contenidos han tenido éxito o no, de manera que se pueda elaborar una estrategia de contenidos inteligente.

Así, se podría trazar un proceso de publicación basado en métricas que, al retroalimentarse del éxito de los contenidos publicados previamente, serviría para elegir una prioridad de publicación que maximice el éxito de los futuros contenidos.

Todo ello se basa en el coste de oportunidad. Los recursos económicos y temporales son limitados, por lo cual es necesario optimizar los esfuerzos de manera que se obtenga la mejor respuesta posible, gestionando la actualización de los contenidos y la importancia que tiene la inmediatez en la estrategia (Arrese, 2013). Se pretende por un lado disminuir el coste de la selección y elaboración de artículos y por el otro aumentar el alcance para extraer el máximo rendimiento de cada publicación aprovechando las características propias de cada canal.

El objeto de estudio está formado, entonces, por los contenidos web publicados por un medio digital, y está delimitado en función de los siguientes parámetros:

a. Temático

Los contenidos web que forman parte del objeto de estudio están ambientados en las temáticas a las que se dirige el medio digital analizado, para el cual se quiere optimizar su estrategia de contenidos.

El resultado de consultar dichas temáticas y las palabras clave relacionadas es la obtención de los contenidos web más exitosos de cada temática en el presente, así como la tendencia de las temáticas relacionadas con cada artículo publicado.

b. Idiomático

Los contenidos web que se han analizado, ya sea a través del medio online que se ha utilizado como caso de uso en esta tesis como el resto de los usuarios que han publicado en las mismas temáticas en Twitter, han sido en el idioma español.

Un buen porcentaje de los contenidos publicados en el caso de uso se basan en productos con sede en Estados Unidos. Por otro lado, puesto que el equipo del medio online proviene en su mayoría de España, también se publican contenidos sobre productos españoles con una mayor asiduidad de la de los demás países del mundo. Y la mayoría de los usuarios que visitan el sitio web son asimismo de España.

Sin embargo, puesto que los contenidos son de interés general dentro de las temáticas escogidas, no se ha concretado ninguna variedad dialectal salvo la oficial del idioma español, ni tampoco un parámetro geográfico que lo delimite.

c. Temporal

La temporalidad está sujeta a los periodos elegidos para poner en práctica la metodología cibermétrica en la estrategia de contenidos del caso de uso.

El periodo elegido en el que se han recogido datos de entrenamiento (fase 1) ha sido el de los artículos publicados entre el 23 de septiembre y el 22 de noviembre de 2020, sumando un total de dos meses naturales. El periodo en el que se han recogido los datos de test (fase 2) ha sido el de los artículos publicados entre el 23 de noviembre de 2020 y el 22 de diciembre de 2020, sumando un total de un mes natural.

El total es de tres meses, siguiendo temporalmente una proporción de 66,66% de datos de entrenamiento y 33,33% de datos de test, muy parecida a la que se suele utilizar según Recuero de los Santos (2020), de 70% y 30% respectivamente.

A la hora de analizar cada artículo y la tendencia de sus términos, se han capturado sus datos catorce días después del día de su publicación, por lo que los datos de entrenamiento terminaron de recogerse el día 6 de diciembre de 2020, y los datos de test el día 5 de enero de 2021.

Tabla 1

Resumen de fechas del objeto de estudio

Fase	Publicación (inicio)	Publicación (fin)	Toma de datos 14 días después (inicio)	Toma de datos 14 días después (fin)
Fase 1: Entrenamiento	23/09/2020	22/11/2020	07/10/2020	06/12/2020
Fase 2: Test	23/11/2020	22/12/2020	07/12/2020	05/01/2021

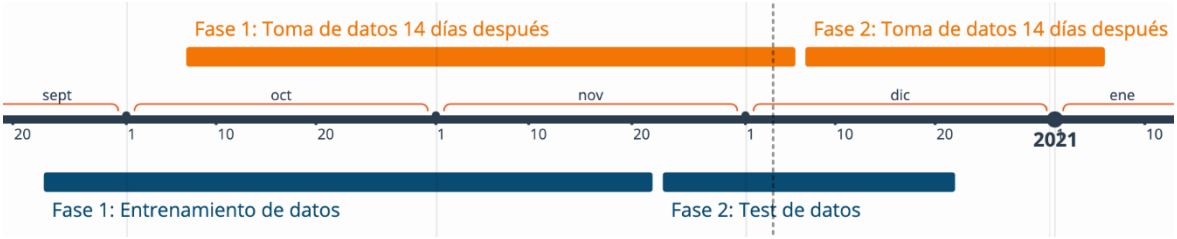


Figura 2. Línea de tiempo con el marco temporal del objeto de estudio

d. Unidad de observación

La unidad de observación básica es la URL, ya que cada una pertenece a un contenido web individual que a su vez es enlazado y accedido desde los canales utilizados por el medio digital.

Cada URL contiene un artículo que a su vez está relacionado con determinadas temáticas, ya sea por sus categorías o por sus etiquetas, para así poder estudiar no solo su éxito sino también el alcance al que opta en el momento actual.

Debido a que en un medio de comunicación online se pueden publicar diversas tipologías de artículos, para delimitar aún más esta unidad, se ha escogido el tipo de artículo “noticia”. Cuenta con una frecuencia de publicación mayor en un medio de información online y presenta un factor de novedad y de inmediatez, por lo que se podrá beneficiar en mayor medida de la metodología cibermétrica que se propone en esta tesis.

1.2. Objetivos

1.2.1. Objetivo principal

El objetivo principal que supuso el aliciente y la necesidad de la elaboración de esta investigación ha sido el diseño de una metodología cibermétrica de cálculo del éxito y su predicción de una publicación online, de manera que se pueda plantear como problema de investigación la optimización de la publicación de contenidos web en el ámbito del periodismo digital.

1.2.2. Objetivos específicos

Los objetivos específicos son de diversos tipos, puesto que tratan de acercarse al objetivo principal en las diferentes etapas de la tesis.

En primer lugar, se ha investigado las funcionalidades de Twitter como red social y qué se entiende como éxito dentro de esta, tanto a nivel de cuenta como a nivel de publicaciones singulares. También ha sido necesario hacer un análisis de las tendencias relacionadas con cada contenido y cuáles de estas podrían tener más éxito.

Con todo ello se ha propuesto una metodología mediante la cual calcular el éxito de una posible publicación. Se ha estudiado una lista de herramientas que puedan aportar los datos necesarios para realizar dicho cálculo en base a unos indicadores de éxito y, así, proponer ecuaciones de predicción de éstos.

En cuanto a la aplicación de la metodología de la tesis, se ha comprobado los resultados en un caso real, adaptando las temáticas a estudiar a los contenidos de este. De esta manera se han puesto a prueba las ecuaciones predictoras de las variables de éxito y se ha comprobado su precisión de manera que sirva como nivel de confianza en cada una.

Los objetivos específicos de esta tesis, planteados de manera concisa, han sido los siguientes:

1. Investigar el concepto de éxito en el periodismo digital, la red social Twitter, la analítica web y la publicidad en la web.
2. Diseñar la metodología y determinar qué procedimientos estadísticos, herramientas y reportes son necesarios.
3. Extraer los datos y procesarlos para obtener los indicadores.

4. Realizar regresiones que permitan obtener ecuaciones de predicción de las variables de éxito seleccionadas.
5. Validar las ecuaciones de predicción con datos de test y obtener su precisión.

1.3. Justificación de la investigación

Este proyecto ha nacido de la necesidad de optimizar los recursos y aumentar la difusión de los contenidos online, algo muy vigente en el ámbito actual debido a la creciente proliferación de proyectos web que utilizan los contenidos para alcanzar y retener usuarios, pero también una necesidad que ha existido desde los comienzos de internet.

Bill Gates escribió un artículo el 1 de marzo de 1996, citado por Bailey (2010), en el cual aseguraba que el contenido era el rey, puesto que era donde él esperaba encontrar dinero. Se basaba en el éxito del contenido en otros medios como la televisión, destacando que Internet permite la distribución de contenidos mundialmente a un coste marginalmente nulo para el publicador.

Sin embargo, ¿qué contenidos tienen más éxito y por qué? La motivación principal de esta tesis es la de proponer una metodología aplicada a un caso real, que permita analizar los diversos factores entendidos como éxito en el marketing de contenidos y su posible predicción.

Las principales utilidades que presenta esta tesis son las siguientes:

a. Metodológica

El diseño de la metodología cibermétrica que se aporta como posible respuesta al problema de investigación de la tesis actual ha servido como marco para estudiar el éxito en el marketing de contenidos basándose en los contenidos en sí y no en la web como conjunto, acercándose a las métricas singulares de cada temática y cada tendencia.

Así, se entiende el éxito de la optimización de contenidos web como un flujo continuo en el cual el proyecto online avanza aportando contenido que cumpla con las necesidades de información de los usuarios que lo consumen, y alejándose de la definición del éxito como una métrica fija del medio.

Al proponer al contenido como foco del análisis, también se obtiene una metodología que estudia con más profundidad el alcance y la evolución de este, lo cual puede servir de ayuda para llevarlo a otros ámbitos diferentes del cual pertenezca el medio publicador del contenido.

El desarrollo de esta metodología ayudará, entonces, a rediseñar y mejorar el análisis del marketing de contenidos en proyectos similares, utilizando un análisis y una recolección

adaptadas al medio y a las temáticas a las que se dirija, pero sirviéndose de las conclusiones que surjan de los nichos estudiados.

b. Práctica

Este trabajo tendrá una serie de implicaciones prácticas derivadas del análisis del sector del periodismo digital:

1. La metodología permitirá optimizar tanto los recursos como la toma de decisiones en la estrategia de contenidos de la web, lo cual puede llevar a la misma optimización en proyectos parecidos.

2. Se analizarán las tendencias en las temáticas relacionadas con el sitio web, así como la necesidad de información y el consumo de esta.

3. Servirá como estudio más en detalle de la audiencia y la relación entre los usuarios, así como los de estos con las temáticas, los competidores y los influencers o usuarios con más seguidores.

Todo ello aporta conclusiones relacionadas con las temáticas del caso de uso, pero que también pueden ser llevadas a otros contribuyendo a la mejora de la comunicación de entretenimiento en internet y un mejor estudio del marketing tanto de contenidos como de redes sociales.

Muy lejos están las ideas iniciales de los profesionales de la industria de que se trataba de una novedad pasajera. Internet ya era usado a principios de los 2000, según Kawamoto (2003, p. 13), como fuente de información para artículos, conferencias, libros y otros trabajos académicos, así como todo tipo de publicaciones de otros ámbitos.

De esta manera, se podrá extrapolar el presente estudio a otros marcos diferentes, en los cuales estudiar el éxito tanto a nivel de medio como de contenido. Esto permite que se pueda aportar conclusiones valiosas para el sector del marketing digital, derivadas de la identificación de los factores clave del éxito, su análisis en un proceso automático y su aplicación en el entorno siempre cambiante de Internet.

c. Académica

La presente tesis presenta una visión novedosa en el cálculo del éxito de contenidos en línea, que inicia una nueva vía de investigación basad en la analítica de contenidos, la

interacción de los usuarios, la necesidad de información de actualidad y el análisis de tendencias.

Los expertos tanto del marketing de contenidos como de redes sociales podrán beneficiarse de los resultados de esta metodología, pero también los académicos cuyo interés sea investigar cómo evoluciona la sociedad de la Información, cuyo estado actual se analizará en el apartado 2, el estado de la cuestión de esta tesis. Un análisis más pormenorizado y aplicado a otras redes sociales, temáticas o públicos objetivo puede llevar a todo tipo de estudios que aportarán conclusiones teóricas e incluso prácticas a la hora de entender el carácter de nuestra presente comunidad digital.

Todo ello puede servir no solo para probar entendimientos generales en la publicación de contenidos online, sino también para desmentir bulos sobre el comportamiento de los usuarios cuando están en línea y lo que consideran de interés en cada momento.

1.4. Tipo de investigación, metodología y fuentes

1.4.1. Tipo de investigación

Se trata de una investigación aplicada porque se centra en desarrollar una metodología o estrategia que permita encontrar un objetivo concreto, como es la optimización de los contenidos web, aplicándose en un ámbito muy específico y bien delimitado, utilizando un caso de uso con unas temáticas cerradas.

Debido al carácter innovador del problema investigado, ya que no se han encontrado investigaciones que pongan su foco en los contenidos web en sí, se trata de una investigación exploratoria, en la que se realiza un primer acercamiento a la problemática bajo unas circunstancias concretas y permite que otras investigaciones posteriores puedan dirigirse a otras temáticas y ámbitos del marketing de contenidos.

Así pues, trata de encontrar patrones significativos en los datos a través del cálculo e intento de predicción del éxito que sirve de base para la metodología que abarca esta tesis. Asimismo, también podría servir a futuras investigaciones que modifiquen o resuelvan las incógnitas a las que pueda dar pie este acercamiento.

El enfoque de la investigación ha sido cuantitativo, pues se ha servido de diversos procedimientos basados en la medición del éxito de las publicaciones en Twitter, el análisis de tendencias y la analítica web perteneciente al medio en sí.

Para ello, se ha servido de una investigación experimental, haciendo uso de un medio digital en un intervalo de tiempo concreto, en el cual se han dado lugar un flujo de tendencias y de intereses de cara a los usuarios que han dado pie a los resultados obtenidos. Así, se ha analizado qué efecto ha producido este flujo y qué variables tienen una mayor importancia en el cálculo del éxito que da origen a la metodología.

1.4.2. Estructura del trabajo

Ha sido necesaria la investigación longitudinal de un caso de uso real a lo largo de un periodo concreto de tiempo, para cuya puesta en práctica ha sido necesaria la elaboración de una metodología, proceso que ha constado de las siguientes fases:

1. Desarrollar una metodología que incluyera la extracción, recolección y tratamiento de los datos necesarios para recuperar la información de las publicaciones, tendencias e intereses del público objetivo.
2. Evaluar e interpretar los datos extraídos y detectar resultados anómalos.
3. Extraer de su análisis estadístico las ecuaciones de predicción de las variables de éxito de una publicación concreta.
4. Validar las ecuaciones de predicción durante un intervalo de tiempo, para calcular la precisión de las regresiones estadísticas.

1.5. Fuentes de datos

Las fuentes primarias de datos son los contenidos web publicados y los tuits, es decir, las publicaciones vía Twitter cuya analítica posibilitará la ejecución del experimento que confirma o descarta la hipótesis de este proyecto.

Por un lado, se ha hecho uso del conjunto de tuits publicados de manera automática por el medio digital que sirve como caso de uso de esta tesis. Estos tuits representan los contenidos web del medio y están relacionados con palabras clave que pertenecen a un grupo temático concreto. Gracias a éstos se pueden extraer métricas que miden el éxito de cada contenido web.

Debido a la necesidad de priorizar y evaluar estos tuits, ha sido necesario hacer uso de los tuits publicados en las tendencias relacionadas con las palabras clave que aparecen en los tuits del medio, derivando en un análisis de tendencias en Twitter.

Por último, para comprender mejor el carácter y las necesidades del público del medio, se han analizado los tuits publicados por los usuarios que siguen la cuenta y que estén relacionados con las temáticas de los contenidos publicados, de manera que faciliten el análisis de tendencias y el posible cálculo de éxito de las temáticas relacionadas.

Asimismo, los contenidos web en sí han sido una fuente primaria de datos. La analítica web de sus sesiones y características de estas ha proporcionado una valiosa información que ha permitido analizar el éxito de los contenidos según los indicadores seleccionados en la metodología.

1.6. Fuentes de información

Las fuentes de información que han ayudado a desarrollar el estado de la cuestión han sido los libros, los artículos publicados en revistas científicas, en periódicos y las páginas web.

Se han consultado libros y artículos de revistas científicas relacionados con el periodismo digital, el sector del marketing digital, el marketing de contenidos, el marketing de redes sociales y especialmente de Twitter, el análisis de tendencias en Twitter y la publicidad web. Como mecanismo para llegar hasta ellos se ha hecho uso principalmente de bases de datos como Google Scholar[2], Web of Science[3] y Scopus[4], que han servido a modo de fuentes secundarias de éstos y han ayudado a buscar y contrastar las fuentes bibliográficas. Para ello, se ha hecho uso de términos como: “digital journalism”, “Twitter”, “Twitter news”, “web analytics”, “cybermetrics”, “altmetrics”, “Twitter analytics”, “Twitter data mining”, “trend detection in Twitter”, “event detection in Twitter”, etc.

Otra fuente de información vital en esta investigación ha sido la de las páginas web, puesto que muchos de los mayores expertos en marketing digital y ciberperiodismo contaban con medios digitales a través de los cuales comparten sus conocimientos y experiencias. Y no solo a través de estas, sino también sirviéndose de las publicaciones en otros formatos como las diapositivas, infografías y vídeos explicativos.

En conjunto, las fuentes han hecho posible elaborar un Estado de de la cuestión que informa de los avances más importantes en la actualidad en lo relacionado con la tesis, que justifican tanto la necesidad de esta como su factor innovador.

[2] https://scholar.google.es/
[3] https://www.webofscience.com/
[4] https://www.scopus.com/

2. Estado de la cuestión

En el presente estado de la cuestión se muestran antecedentes de investigación de los temas relacionados con la tesis actual. La finalidad es la de conocer qué se ha investigado hasta la fecha y cuál es el conocimiento existente, destacar las vías de búsqueda abiertas hasta el presente y buscar pautas para plantear el problema de investigación que se estudia y en qué carencias se justifica su necesidad.

Este apartado enmarca la tesis doctoral tanto en el ámbito de la difusión online de información periodística como en el análisis de dicha difusión tanto en la web como en Twitter.

Para ello, el estado de la cuestión se ha dividido en dos grandes apartados: el del Periodismo digital (2.1) y el de la Medición web de éxito (2.2). En el primero se introduce la transformación digital del periodismo, en qué consiste el consumo de contenidos periodísticos en su formato online, qué es Twitter y cuáles son sus características principales, y cómo se produce la difusión de las noticias en Twitter. En el segundo apartado se profundiza en la analítica web, la cibermetría, la analítica en Twitter, el análisis de tendencias en Twitter y la publicidad web.

2.1. Periodismo digital

A continuación, se revisa el conocimiento existente relacionado con el periodismo digital, con el objetivo de enmarcar el análisis de contenido periodístico dentro del panorama actual. Esta búsqueda de antecedentes de investigación también incluye una revisión a través de los años tanto del periodismo digital como de Twitter para la difusión de las noticias de última hora en dicha red social.

2.1.1. Consumo de contenidos

En primer lugar, se revisa el origen de los conceptos principales del ámbito del periodismo digital y qué se ha estudiado en ese ámbito.

2.1.1.1. La transformación digital

El impacto de internet en las tecnologías de la comunicación en el periodismo lleva siendo analizado desde 1994, con la llegada de un nuevo tipo de comunicador: el periodista online (Deuze, 1998). Bardoel y Deuze (2001) señalaron que la convergencia, la interactividad y la hipertextualidad del periodismo en internet unidos a un uso generalizado del mismo, estaban poniendo a prueba los tipos y géneros del periodismo. Todo ello podría desembocar en lo que ellos definieron como “periodismo en red” (*network journalism*).

Pavlik (2000) describió el periodismo online como un nuevo sistema mediático que incluía todas las formas de comunicación humanas en un formato digital, en el cual ya no se podían aplicar las reglas del mundo analógico. De esta manera, destacaba el impacto de la tecnología en el periodismo hasta el punto de que estaba reescribiendo la organización y la estructura de la sala de prensa. Más adelante, Salaverría (2005, p. 21), que prefiere llamarlo ciberperiodismo, lo describiría como “la especialidad del periodismo que emplea el ciberespacio para investigar, producir y, sobre todo, difundir contenidos periodísticos”. Pero la innovación en el periodismo debería estar sustentada por cuatro principios. Según Pavlik (2013), estos serían inteligencia o investigación, libertad de expresión, dedicación a la búsqueda de la verdad y la autenticidad y la ética.

Cebrián Herreros (2009) habló de tres modalidades de cibermedios: cibermedios matriciales, cuyo origen se inició en papel como la prensa o difusión electrónica como la radio y la televisión; cibermedios nativos, nacidos en Internet pero con una mentalidad tradicional que luego incorporó características específicas de las nuevas tecnologías; y cibermedios sociales, basados en la comunicación directa entre los usuarios a través del diálogo o intercambio de información de manera independiente.

La rápida evolución de las nuevas tecnologías mediáticas obligaba a una reestructuración del mercado y los valores de las noticias, debido según Bierhoff et al (2000) a motivos sociales, culturales y político-económicos, lo cual llevó a la confusión tanto entre profesionales como entre los educadores de éstos últimos. Todo ello obligó a una redefinición del periodismo a través del diálogo y el intercambio de puntos de vista, y la necesidad de crear plataformas periodísticas nacionales que facilitaran la investigación y la adaptación de la educación a las características del periodismo digital. Wahl-Jorgensen (2017, p. 251) argumenta que esto se debía a que, más allá del discurso entre las nuevas tecnologías contra las viejas o lo innovador contra lo conservador, era necesario plantear cuestiones relacionadas con las prácticas en el periodismo digital, especialmente sobre las menos privilegiadas, menos utilizadas o más marginales, para añadir no solo un estudio del éxito sino también del fracaso y su inevitabilidad. Sin embargo, según Deuze (2017), los valores tradicionales que definieron al periodismo profesional han sobrevivido a las presiones de las empresas, los desafíos tecnológicos y los cambios en las subvenciones.

Zion y Craig (2015, p. 4) señalaron como necesaria una revisión del código ético periodístico sobre materias como el origen de los datos, la selección de éstos y su uso. Díaz-Campo y Segado-Boj (2015) investigaron 99 códigos éticos periodísticos de todo el mundo y, de los 31 que se habían escrito o revisado desde 2001, solo 9 de ellos habían incorporado referencias específicas a Internet y las tecnologías de la información y la comunicación (ICTs).

El periodismo online era así considerado el cuarto tipo tras la radio, televisión y el medio impreso, siendo definido por Bardoel y Deuze (2001) como “el acto de reunir y distribuir noticias de contenido original en Internet”. Sánchez Sánchez (2007) no lo consideraría una nueva forma de hacer periodismo, sino un retorno a la esencia del periodismo, ya que en un entorno online los periodistas se ven obligados a enfocarse hacia las bases históricas de su profesión: “la investigación, la claridad, la brevedad, la contextualización y el manejo de múltiples fuentes”.

Navarro Zamora (2000) estimó que Internet fue el precursor de la tercera crisis del periodismo (tras la radio en los años 30 y la televisión en los años 50), ya que obligaba al periodismo tradicional a adaptarse a este nuevo contexto. En 1995 había más de 120 diarios online en Estados Unidos, en 1996 había casi 800, de los cuales casi la mitad eran periódicos, y en enero de 2000 recopiló los datos que se pueden ver en las siguientes figuras, en los cuales se observa que el número se incrementó a un total de 11.428, de los cuales 4.322 eran periódicos. También se puede comprobar una fuerte predominancia de

Estados Unidos tanto a nivel de medios de comunicación online como de periódicos online en concreto, superando ampliamente en número al resto de países e incluso continentes.

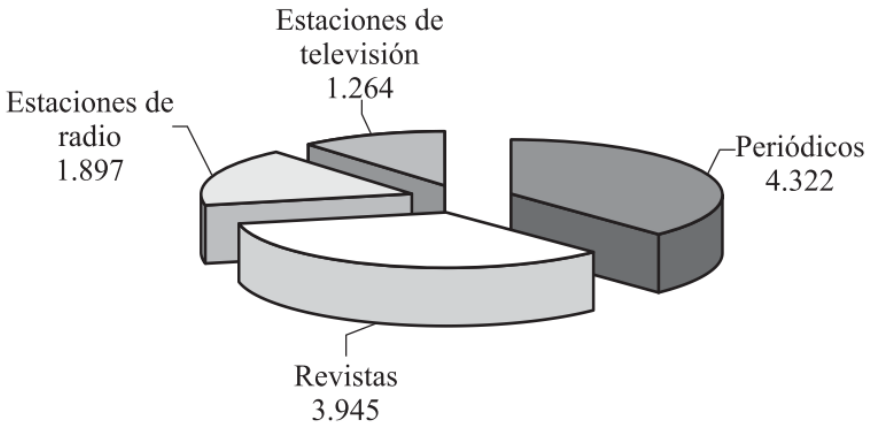


Figura 3. Distribución por medio de los medios de comunicación online en enero de 2000: 11.428 en total (Navarro Zamora, 2000).

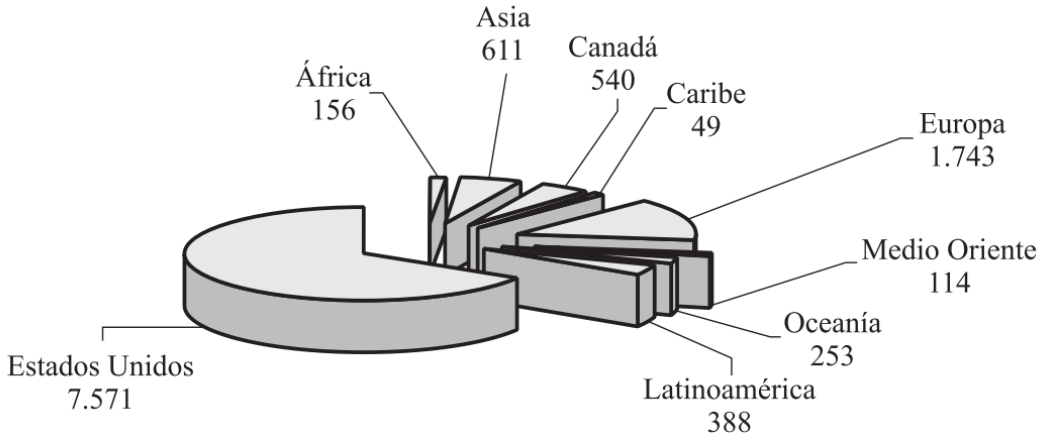


Figura 4. Distribución por zonas geográficas de los medios de comunicación online en enero de 2000 (Navarro Zamora, 2000).

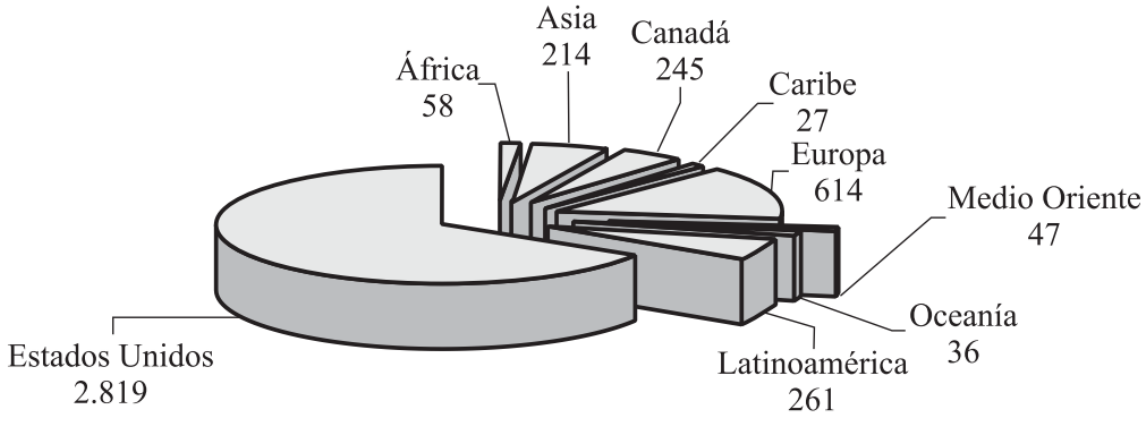


Figura 5. Distribución por continente y Estados Unidos de los periódicos online en enero de 2000: 4.322 en total (Navarro Zamora, 2000).

En cuanto a España, un estudio de Caminos Marvel et al. (2006) indicó que en 1994 ya existían publicaciones con contenidos en red de la revista valenciana El Tems o el Boletín Oficial del Estado, pero no sería hasta 1995 que la prensa española desembarcó en Internet, siendo La Vanguardia, El Periódico de Catalunya, Avui, Abc, El Diario Vasco y El Comercio entre los primeros medios en tener su versión online. En 1996 le siguieron El País, El Mundo y Marca. A finales de los 90 ya eran 60 los diarios españoles que tenían edición digital y, en 2006, la mayoría estaban presentes en la red.

La necesidad de que el contenido sea original se debe a la noción general de que la difusión online es un nuevo medio, y por tanto requiere, según Deuze (1998), de contenido original en vez de meras copias de sus versiones impresas. Si una publicación online no hace uso de las ventajas de este nuevo medio, no está haciendo uso de manera óptima de sus posibilidades. Aun así, dichas ventajas estaban supeditadas a los límites de velocidad de los módems, por lo que Li (1998) pronosticaba que el desarrollo de estos módems tendría un gran impacto en el diseño web y uso de gráficos por los periódicos en Internet. Y para hacer reales esas ventajas, Farías de Estany y María Prieto (2009) especificaron que se debía incorporar a la organización del cibermedio un equipo horizontal y multidisciplinar, ya que no solo son necesarios periodistas, sino informáticos, especialistas en telecomunicaciones y diseño gráfico, participando todos en la toma de decisiones, la concepción del contenido y los procesos que garantizan el correcto funcionamiento, como se puede observar en la Figura 6.

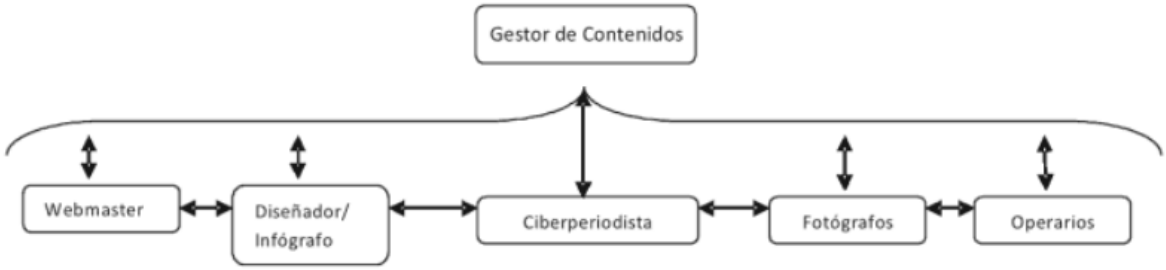


Figura 6. Estructura de funcionamiento del cibermedio (Farías de Estany & María Prieto, 2009).

La labor del ciberperiodista sería, según Farías de Estany y María Prieto (2009), la de ser también el gestor de contenidos, por lo que debe planificar, organizar y estructurar los contenidos informativos, investigar, redactar los contenidos, publicar la información en el entorno virtual, etc. Un proceso de producción del contenido que se puede ver en el esquema de la Figura 7.

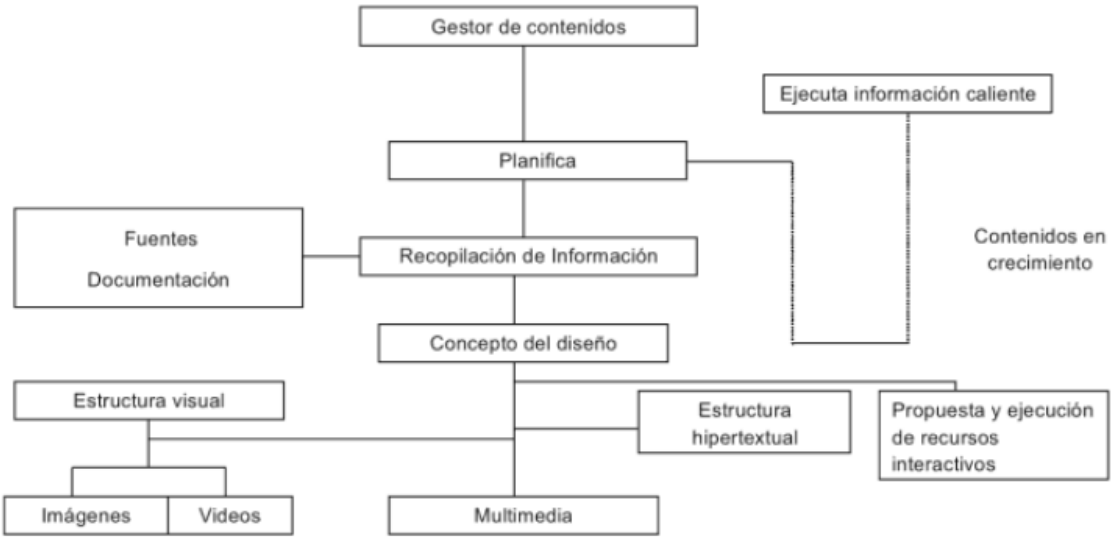


Figura 7. Proceso de producción del contenido (Farías de Estany & María Prieto, 2009).

Según Bardoel (1996, p. 178), la tecnología hacía menos necesaria la intervención periodística, por lo que el futuro de dicha profesión debía basarse en otros factores sociales, ya que Internet tenía el potencial de hacer que los periodistas fueran superfluos como intermediarios en la democracia del medio. Sin embargo, Singer (1997) llegó a la conclusión de que los periodistas no solo se consideraban profesionales que reunían información y le daban la forma de una historia, sino que también debían interpretarla para sus lectores. Ese énfasis en el rol interpretativo y en ser una fuente de información equitativa y creíble, es el que les hace imprescindibles para el medio. Según un estudio realizado por Singer (1997) a 27 reporteros y editores en 1995, las aptitudes necesarias para llevar a cabo un buen trabajo periodístico pueden ser aprendidas y mejoradas, pero serán solamente el medio por el cual ellos cumplen con el objetivo de dar sentido de manera creíble a la información.

El periodista online es, por tanto, un filtro que selecciona e interpreta lo que de otro modo hubiera sido un suministro caótico de información (Singer, 1997). Un componente esencial para que los medios sean de comunicación y no solo de información, puesto que además es, indica Sánchez Sánchez (2007), su responsabilidad ética la de aportar varios puntos de vista, un análisis ponderado y objetivo, una contextualización y la posibilidad de respuesta de los usuarios. El periodista online debe aprovechar la ventaja de que no haya un límite de espacio para cumplir con dicha responsabilidad en el formato de la pirámide invertida, mediante el cual se presentan los datos en el orden de mayor a menor trascendencia.

Los ciberperiodistas deben, por ello, tener un entrenamiento fundamental en habilidades de escritura y edición y un conocimiento básico de informática, señala Ibrahim (2011). Necesitan reorientar su escritura hacia un estilo no lineal y con una alta exigencia en fuentes de información, experimentación, investigación de datos empíricos, habilidades estratégicas y de recolección de inteligencia, así como nociones de diseño y creatividad. Ante el desafío de tantas necesidades y una sociedad globalmente conectada, puede que los periodistas no estén preparados para realizar todas las tareas claves del periodismo tradicional, por lo que según Haak et al (2012) necesitan especializarse en alguna de ellas. Esto posibilita que, en suma, se puedan optimizar recursos y generar una sinergia y creatividad producidas por el intercambio de experiencias.

El periodista digital, advierte Albertos (2001), debería estar en constante formación e innovación, para asegurar que tiene las habilidades técnicas necesarias para el desempeño de su trabajo. Y dicha evolución tecnológica permitiría una mayor personalización de los contenidos, evolucionando del interés general al particular, lo cual además fomentaría un nuevo modelo de sociedad cimentada en dicho conocimiento individualizado.

El periodista, según Ardèvol-Abreu (2015), tiene una influencia en el encuadre o *framing* de los procesos comunicativos, ya que cualquier texto necesita estructuras narrativas que construyen el discurso y la noticia es, al fin y al cabo, una ventana a través de la cual se descubre la realidad a la que se tiene acceso. Esto limita la percepción de la realidad a la que se comunica con ese fragmento de información, produciendo así un marco de interpretación para quien lo consuma. Es el periodista el que elige el encuadre en ese proceso, llamando la atención de unos aspectos de la realidad en lugar de otros. La teoría del *framing*, por tanto, otorga un rol importante en la presentación del contenido, aunque no hay que olvidar al receptor, ya que participa en el proceso de decodificación, interpretación y asimilación de la información. Todavía se encuentra en un estado de definición y reescritura, pero este campo de investigación ayuda a analizar y cuestionar no solo lo publicado en el ámbito periodístico sino incluso en las redes sociales, donde también se podría aplicar la misma teoría (Acevedo, 2019).

Salaverría (2009) definió el concepto de convergencia periodística como un proceso multidimensional en el que los medios de comunicación, afectados por las novedades tecnológicas digitales, deben integrar herramientas, espacios, métodos de trabajo y lenguajes que antes no se tenían en cuenta. Es por ello por lo que en el ámbito tecnológico habla de multiplataforma, en el ámbito empresarial de concentración, en el ámbito profesional de polivalencia y en el ámbito de los contenidos de multimedialidad. Es en

conjunto una transformación multidimensional que exigiría no solo un mayor enlace entre los grados de Periodismo y Comunicación Audiovisual, sino una gran cualificación en el uso profesional de dispositivos y aplicaciones digitales. Sin embargo, la educación superior no estaba preparada para una evolución tan rápida de sus exigencias laborales en el ámbito tecnológico. Montiel y Villalobos (2005) comprobaron casi una década después del nacimiento del periodismo digital que muchas de las escuelas de comunicación social no habían introducido todavía las tecnologías de la comunicación y la información en su plan de estudios.

Es necesario diferenciar a los medios tradicionales de los medios online. Para ello, Deuze (1999) se basó en tres características principales: interactividad, personalización y convergencia.

- Interactividad: ha sido considerada desde los inicios del periodismo digital como una de sus principales ventajas, ya que un contenido digital puede disponer de múltiples elementos interactivos con su audiencia de gran importancia: enlaces, formularios de comentarios...
- Personalización: consiste en la individualización de los contenidos, de manera que se pueda ofrecer al usuario lo que necesite.
- Convergencia: es, al mismo tiempo, la personalización del diseño y los bloques de navegación, a través de los cuales puede elegir sus temáticas preferidas y, así, qué y cómo consumir la información que busca en el periódico. La tecnología de la convergencia no solo suponía más opciones para el usuario, sino más exigencias para los periodistas, ya que Dimitrova y Neznanski (2006) señalaron que incluso en los comienzos del periodismo digital ya se hablaba de tecnologías verdaderamente inmersas y experiencias tridimensionales.

Fenton (2009, p. 601) describió las principales características de crear el máximo impacto en Internet como velocidad y espacio, multiplicidad y policentrismo, interacción y participación. La necesidad de fomentar esta participación apremia a los periodistas a ser activos en las redes sociales de manera perenne, por lo que Hedman y Djerf-Pierre (2013) estudiaron el uso que hacían los profesionales del periodismo de las redes sociales y los clasificó en tres tipos de usuarios: escépticos, conformistas pragmáticos y activistas entusiastas. Las diferencias entre estos no estaban asociadas solo con su edad sino también con su actitud profesional hacia la adaptación de la audiencia y la percepción de la marca.

Los lectores, por su parte, valoraban especialmente en el medio online la facilidad de acceso, la personalización de los contenidos, la constante actualización, la confianza cada vez mayor en la información de la red y la gratuidad (Rodriguez-Martinez & Pedraza-Jimenez, 2009). Sin embargo, prefieren los medios que ya tenían renombre en otras plataformas. En un ambiente con un número cada vez mayor de medios digitales de noticias, Nelson (2020), en un estudio realizado en 2019 para evaluar el consumo de noticias online en Estados Unidos durante un año, comprobó que la mayoría del consumo se produce en los medios online que provienen de la radio, la televisión o el papel.

2.1.1.2. Tipos de medios y contenidos periodísticos

Deuze (2001) propuso una clasificación de los tipos de medios periodísticos en Internet, localizados dentro de los dominios de su concentración a nivel de contenido editorial y su comunicación participativa con los usuarios, pudiéndose observar en la Figura 8:

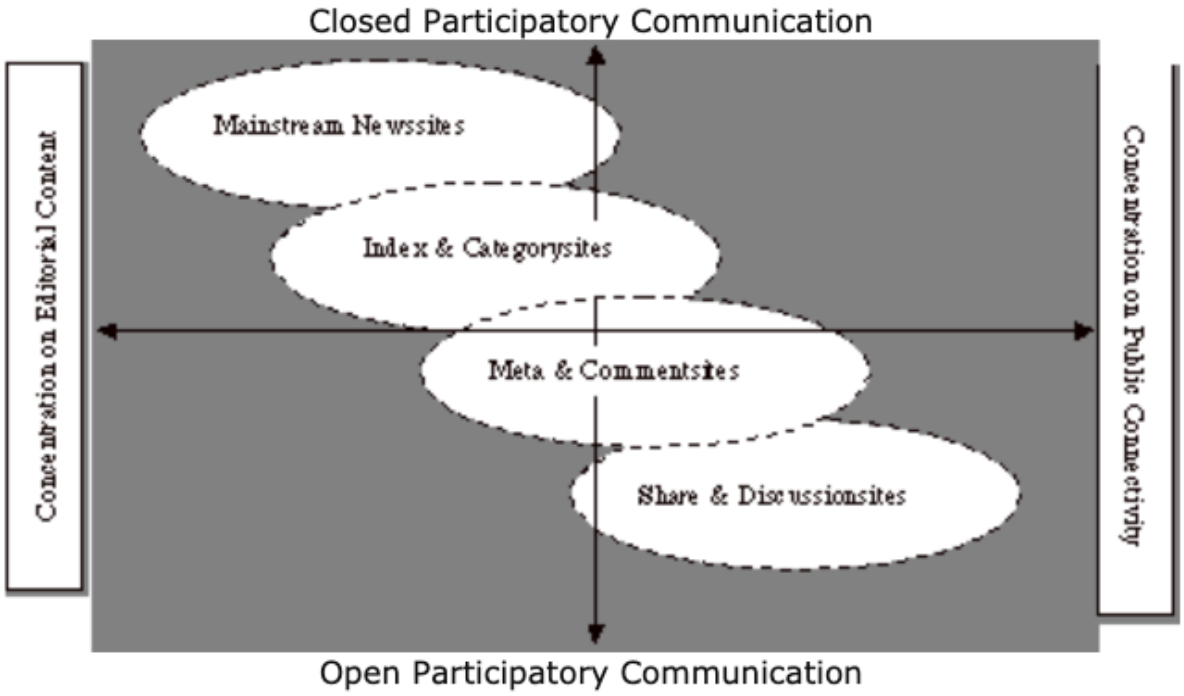


Figura 8. Tipos de periodismo online (Deuze, 2001).

Así pues, el tipo de periodismo online más predominante sería el de los sitios de noticias generales, que ofrecen una selección de contenido editorial y métodos moderados de comunicación participativa. El resto de las tipologías eran, según Deuze (2001), los directorios de enlaces, los directorios de contenidos y los foros, como una demostración de la democratización de la información en Internet.

Es por ello por lo que Deuze (2001) aporta tres estrategias para la producción de noticias online:

- Elaboración anotativa, en la cual se pretende informar con veracidad y añadiendo las fuentes pertinentes.
- Periodismo de fuente libre (*open source*), que aprovecha el potencial de información de expertos de todo el mundo en Internet convirtiéndolo en una infraestructura de comunicación en sí misma.
- Sitios de noticias hiperadaptativos, en los cuales predomina la digitalización de todos los formatos de información entre los dispositivos (TV, móviles...).

En los últimos años de la década de los noventa, los medios de comunicación anunciaron la llegada de internet con múltiples contenidos al respecto, que llegaron a provocar no solamente emoción sino también ansiedad en sus públicos. Estaban produciendo grandes avances en la información de manera constante, provocando una sensación de estar al margen para quien no se hubiera conectado ya a Internet. Dahlgren (1996) mencionó además el problema de la sobresaturación de información de la que disponía el usuario, así como de la rápida actualización de hardware y software que hacían anticuados los recursos informáticos en cuestión de meses. Al fin y al cabo, la periodicidad mínima de la radio y, en ocasiones, la televisión, se convirtió en instantánea en Internet para cualquier tipo de información (Edo, 2006).

También se discutió en un primer momento la desventaja de la lectura online frente a la tradicional en cuanto a limitaciones de comodidad, familiaridad, velocidad de lectura y comprensión de texto. Sin embargo, las nuevas generaciones, ya acostumbradas al soporte digital, no presentaron dichas limitaciones, e incluso mostraron en un estudio de Armentia Caminos et al. (2000) unas mejores ratios de comprensión en lectores digitales según un estudio de alumnos de 18 años de Periodismo y Publicidad.

El cambio de paradigma, de la vía tradicional a la digital, provocó según Regan (1997) una serie de diferencias asimismo en la manera de comunicar para el método convencional y para Internet. Al comienzo, periódicos como The New York Times y The Wall Street Journal copiaban los artículos de un medio a otro. Sin embargo, otros como The Chicago Tribute tenían un equipo de reporteros trabajando únicamente para la web, aportando imágenes digitales y vídeos a los artículos para maximizar su impacto e interactividad. Sin embargo, esta interactividad no era aprovechada de manera eficiente, tal y como analizó Schultz (1999) con una muestra de 100 periódicos estadounidenses. Esta afirmación no dejó de ser válida casi una década después, ya que Quandt (2008) hizo un estudio de 1603

artículos de medios europeos y estadounidenses y llegó a la conclusión de que había una falta de contenido multimedia, opciones de interacción directa con los periodistas, fuentes de autor e incluso un repertorio estandarizado de tipos de artículos.

Este cambio de paradigma fue acuñado en 2006 como “paradigma de la cultura convergente” por Henry Jenkins, citado por Elías Pérez (2009), siendo este la suma de la convergencia mediática, la cultura participativa y la inteligencia colectiva. La convergencia mediática no sería entre soportes mediáticos sino entre los mismos cerebros de los emisores y receptores, provocando un cambio cultural, en el cual además los usuarios se convertían en potenciales participantes activos. Esto último es lo que posibilita que se construya un conocimiento global formado por las experiencias de todos los usuarios.

En los primeros años del periodismo digital hubo muchas dudas de que el periodismo digital pudiera tener la habilidad de trascender al medio impreso, según Riley et al. (1998). Al comienzo, fue utilizado como un método para incrementar la cobertura de artículos locales y personales, pero sin facilitar la navegación de los usuarios de un periódico a otro, ya que tenían como principal objetivo mantener al usuario en la web, y alejándose así de la intención de búsqueda de otras opciones. De este modo, se llegó a llamar “colonización” a la ocupación del ciberespacio. Una ”colonización” que se produce, argumenta Londoño Russi (2013), gracias a la definición de procesos de interacción, unificación de recursos y conocimientos para alcanzar los objetivos relacionados con la cantidad de contenido y seguidores.

Tras los problemas producidos por los medios convencionales, se consideraba según Harper (1998, p. 22) que el público había perdido la confianza en ellos. Sin embargo, el periodismo digital contaba con varias ventajas para reducir esa brecha: daba la oportunidad al usuario de ver todos los documentos relacionados con un artículo, el usuario puede mandar un email a cualquier reportero y ser escuchado así por este y, en el caso de no confiar en los medios locales, podría buscar entre las fuentes de información de todo el país. Esta visión también se podría aplicar a nivel internacional, ya que un estudio realizado por Thurman (2013) se demostró que el canal online incrementó la audiencia internacional de los diarios entre 7 y 16 veces. Sin embargo, estas visitas eran breves y el efecto en su atención era mínimo.

2.1.1.3. Estructura de los contenidos en el periodismo digital

Se trata de una manera de comunicar muy compleja, ya que todo es considerado contenido y el periodismo online como el periodismo total, es decir, la integración de todas las formas de periodismo. Teniendo en cuenta una división inicial entre audio, vídeo y texto, Deuze (1998) propuso el siguiente esquema de la Figura 9:

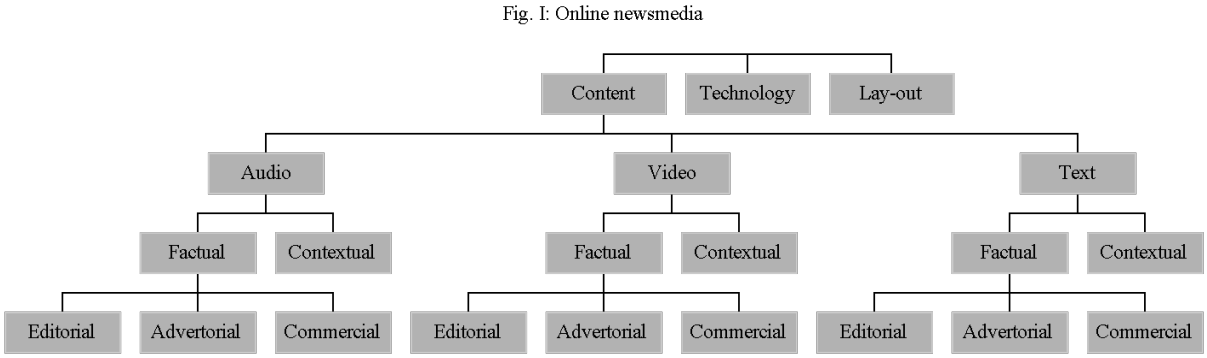


Figura 9. Estructura de análisis longitudinal de contenidos (Deuze, 1998).

Sin embargo, dicho esquema no tiene en cuenta la interrelación y la intercambiabilidad de los diferentes contenidos, ya que se presentan cuestiones como si un enlace es una tecnología o un contenido en sí mismo.

Li (1998) argumentó que los enlaces transforman el proceso de publicación tradicional unidireccional de uno a muchos a una comunicación de muchos a muchos, centrada en el usuario y facilitada por este, ya que al enlazar información más allá del propio contenido del artículo, se produce un cambio en la balanza de poder entre el emisor y el receptor.

En cuanto a la navegación interna, Fredin y David (1998) aportaron un modelo teórico y un marco metodológico que llamaron HIC (*Hypermedia Interaction Cycle*), argumentando que se trataba de un proceso cíclico que constaba de tres fases: preparación, exploración y consolidación. En cada ciclo, el usuario iría utilizando criterios de objetivo menos ambiciosos hasta conseguir llegar a la información deseada. De esta manera, los periodistas pueden estructurar la historia de manera diferente, permitiendo una lectura no lineal y que los lectores elijan su propio camino a través de la historia. Los enlaces, por tanto, se construyen como una asociación de ideas (Herbert, 1999, p. 2). Esta construcción discursiva permite una extensión de la intertextualidad, ya que según Rost (2002) ofrece nuevas vías de acceso a los contenidos, promueve la interacción y le permite al usuario construir la información que desea recibir.

El uso de los enlaces fue incrementando con los años. Tremayne (2004) analizó cerca de 1500 artículos de noticias online y descubrió que este se había triplicado entre 1997 y 2001, como puede verse en la Figura 10:

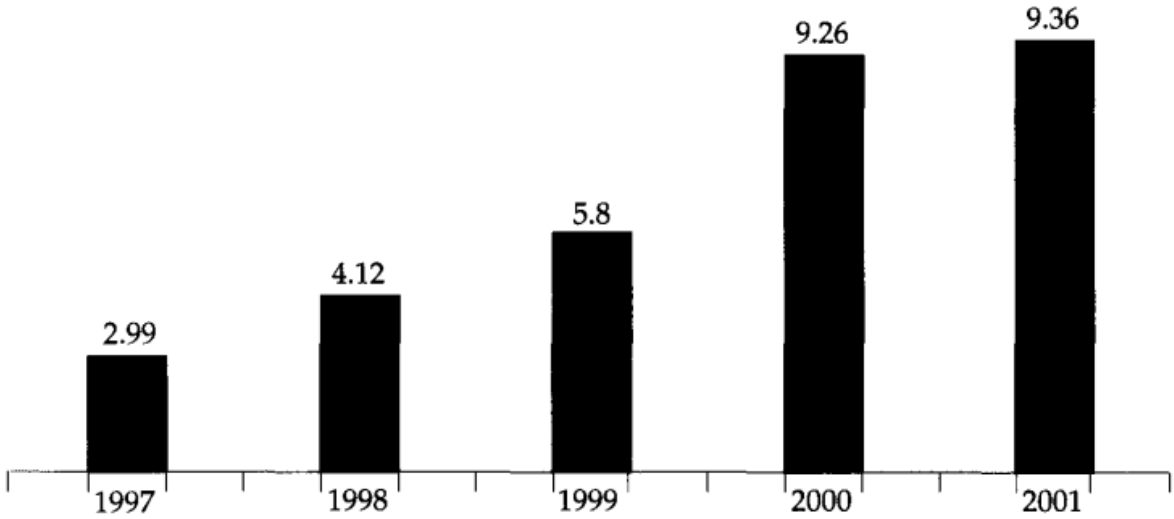


Figura 10. Número de enlaces a material relacionado por artículo y año (Tremayne, 2004).

Además, en 2001 los artículos online originales eran más cortos (con un promedio de 763 palabras) frente a los que se habían trasladado del medio impreso (909 palabras), pero usaban más enlaces (11,6 frente a los 6,6 de estos últimos).

Los enlaces siguen siendo a día de hoy un mecanismo habitual de navegación digital. Roos et al. (2020) comprobaron que, en promedio, los enlaces son beneficiosos tanto para aquellas webs que los contienen como aquellas que son enlazadas. Un enlace aumenta la posibilidad de que el usuario navegue a dicha web en un 0,14%, tres veces más probable que el promedio de los anuncios de publicidad web.

2.1.1.4. Modelos de comunicación

Los contenidos periodísticos se adaptan al modelo de comunicación elegido por el periódico digital que emite la información. Según Carlos Gutiérrez Argüello (2013), existen cuatro modelos de comunicación en los periódicos digitales:

- Modelo de Transmisión: el medio es una entidad transmisora y no generadora, transmitiendo información a modo de mensajes que son creados por los generadores de contenido para una audiencia específica. Estos medios realizan, además, un análisis de satisfacción de audiencia que a veces se produce gracias a una emisión semi-intencionada.

- Modelo Ritual o Expresivo: se prioriza el mantenimiento y la manifestación de los valores y creencias de la sociedad, basándose en la compartición de emociones y conceptos entre el medio, la sociedad y el autor en un ejercicio de interoperabilidad, es decir, la capacidad de compartir datos y hacer posible el intercambio de conocimiento entre ellos. Se produce dicho intercambio en el que el contexto es utilizado de manera directa, dependiendo de la satisfacción del usuario y haciendo por tanto más difícil la separación del contexto. En este modelo se incluyen conceptos como el de "periodismo ciudadano".

- Modelo Publicitario: se trata de lograr un objetivo económico gracias a la atención del público, buscando un beneficio directo en el aumento de la audiencia e indirecto en vender esa atención a los posibles anunciantes. La atención se convierte en medida de éxito y, para potenciarla, se subordina el contenido del mensaje a su presentación.

- Modelo de Recepción: se basa en el modelo de Codificación y Decodificación de audiencias de Hall et al. (1980) y en la Estructura y Discurso Significado, y consiste en codificar la estructura y el tipo de discurso con un propósito de una institución, de manera que el lector atribuya el significado al mensaje, que se convierte en polisémico, y pueda así ser consumido por una gran audiencia.

Según un estudio con datos de campo elaborado entre junio de 2004 y enero de 2005 por Albornoz (2006, p. 201), se comprobó que los contenidos informativos más publicados eran los de tipo "información-noticia", entre los cuales se contaban los publicados directamente para el medio digital, los derivados de la edición impresa y los provenientes de los suplementos impresos. Además, se publicaban las últimas noticias de manera permanente salvo de madrugada, cuando procedían a la publicación de las ediciones impresas en el portal online. Por otro lado, algunos medios ofrecían coberturas en tiempo real de acontecimientos deportivos, información bursátil y financiera y acontecimientos sociales relevantes. En 2005, solamente tres cabeceras digitales españolas contaban con versiones regionales o locales: Mundinteractivos, Prisacom y ABC.

Para considerar que un medio de comunicación en el ámbito del periodismo publica contenidos y utiliza recursos digitales de manera adecuada, Rodríguez-Martínez et al. (2012) consideran necesarios los siguientes supuestos: la adaptación de las herramientas y usos propios de la web para aumentar la visibilidad y accesibilidad, la creación de herramientas de tipo Web 2.0 para canalizar los contenidos y satisfacer la necesidad de sus usuarios, y la relación con otros sitios web que hayan sido aceptados ampliamente por

los usuarios. Para permitir la medición de esta adaptación, los mismos autores propusieron un método de evaluación que estableciera las dimensiones, parámetros e indicadores necesarios. En él, propusieron seis dimensiones: cooperación, participación, creación de contenido, acceso al contenido, socialización y comunicación. Los parámetros e indicadores serían los siguientes:

- Parámetro 1: interacción medio de comunicación-usuario. Indicadores:
    - Comunicación con el autor de la noticia.
    - Contacto con la redacción del medio de comunicación.
    - Comentar noticias publicadas por el medio de comunicación.
    - Votación de noticias publicadas por el medio de comunicación.
    - Comentar entradas publicadas en los blogs del medio de comunicación.
    - El usuario puede modificar o corregir contenido publicado por el medio.
- Parámetro 2: publicación de contenidos creados por los usuarios. Indicadores:
    - Creación de blogs por parte de los usuarios.
    - Publicación de textos escritos por los usuarios.
    - Publicación de fotos tomadas por los usuarios.
    - Publicación de vídeos realizados por los usuarios.
    - Sección exclusiva para contenido creado por los usuarios.
- Parámetro 3: registro del usuario. Indicadores:
    - Registro por parte del usuario en el medio.
    - Contacto con otros usuarios registrados.
- Parámetro 4: acceso a la información. Indicadores:
    - Acceso a la información a través de la portada.
    - Acceso a la información a través de secciones.
    - Acceso a la información a través de noticias relacionadas.
    - Acceso a la información a través del buscador.
    - Acceso a la información a través del mapa web.
    - Acceso a la información a través de la recomendación de los usuarios.
    - Acceso a la información a través de plataformas externas de la Web 2.0.
- Parámetro 5: personalización de la información. Indicadores:
    - Adaptación de la interfaz del sitio web del medio de comunicación en función de los contenidos de interés para el usuario.
    - Sindicación de contenidos del medio de comunicación a través del móvil o correo electrónico.
    - Suscripción de alertas o boletín electrónico.
- Parámetro 6: el medio ofrece distintas versiones de su información. Indicadores:

    - Versión impresa del medio.
    - Versión global.
    - Versión actualizada de forma constante.
    - Versión impresa adaptada a la Web 2.0.
- Parámetro 7: empleo de herramientas de la Web 2.0. Indicadores:
    - Compartir información con otros usuarios.
    - Blogs vinculados al medio de comunicación.
- Parámetro 8: plataformas de la Web 2.0 en las que tiene presencia el medio de comunicación. Indicadores:
    - Presencia del medio de comunicación en plataformas audiovisuales.
    - Presencia del medio de comunicación en plataformas de imágenes.
    - Empleo de redes sociales propias.
    - Presencia del medio de comunicación en redes sociales profesionales externas.
    - Presencia del medio de comunicación en redes sociales de amistad.
    - Presencia del medio de comunicación en plataformas de *microblogging*.
    - Vinculación entre el sitio web del medio de comunicación y las plataformas sociales.

Uno de los ámbitos del periodismo es el enfocado en la cultura, que comenzó en las últimas décadas del siglo pasado a contar con un marco sintagmático fijo donde encuadrarse, con secciones dedicadas a Cultura y Espectáculos, periodicidad diaria y ubicación fija, explica Armañanzas (1996). Eran acompañadas por suplementos dedicados a las Artes, elaborados por especialistas fuera del ámbito periodístico. Todo ello exigía al periodista convertirse en un especialista en cultura para hacer frente a las necesidades de información más específicas del sector, ya que su principal motivación sería saber en lugar de divulgar. Según Londoño Russi (2013), este periodismo de entretenimiento debe, por tanto, facilitar una conexión entre la creación artística y el público, colaborar con el artista en su superación gracias a críticas basadas en conocimientos y la aportación de un contexto que ayude al público a comprender y apreciar la expresión artística. Se pretende preparar a la audiencia y, al mismo tiempo, hacerle llegar en especial los dedicados a la gran masa. Gonzáles Mina (2004) distingue por tanto una diferencia entre el periodismo de contenido de entretenimiento y el periodismo cuya finalidad es entretener.

Albornoz (2006, p. 201) comprobó en 2005 que muchas editoras apostaron por incorporar "información-ocio", con el cual pretendían dirigirse a audiencias jóvenes aficionadas a las

nuevas tecnologías. En ese tipo de artículos, trataban temáticas como los juegos online, concursos de preguntas y respuestas y contenido de humor.

Milenta y Pestano (2009) advirtieron de la necesidad de los medios tradicionales de cumplir con las exigencias informativas de los jóvenes, que poco a poco iban acercándose a ellos, e hicieron un análisis de la presencia de la industria de los videojuegos, una de las más prósperas de entonces, en los periódicos españoles de mayor difusión. Llegaron a la conclusión de que El Mundo y sobre todo El País sí que incluían dicha temática en sus contenidos, pero sin embargo ABC no.

Tanto los críticos como los defensores de Internet veían a este como una gran oportunidad, argumenta Scott (2005). Los primeros esperaban que aliviase las condiciones de crisis del periodismo y le empujasen de nuevo hacia el rol de sirviente público. Los últimos, en cambio, veían un nuevo gran mercado en el cual el periodismo encontraría su lugar en el ciberespacio, ya que Internet tenía el potencial de cambiar la manera en que la información es producida y consumida.

Según Rasmussen (1997), esto último provoca una mejora en la heterogeneidad y crea ambientes de comunicación relativamente cerrados, en los cuales se tienen más en cuenta criterios sociales como la pertenencia a grupos sociales y las características inherentes de éstos, su estatus social... con lo cual se refuerzan las identidades existentes tanto en el plano privado como en el profesional y se incentiva la creación de nuevos roles gracias a una comunicación más especializada. Además, el periodismo digital promueve la discusión y la comunicación a través de intereses comunes, trascendiendo limitaciones de idioma, hábitos, competencias...

2.1.1.5. Periodismo automatizado y el procesamiento de la información

La información asistida por ordenadores ha ayudado a los periódicos a procesar grandes cantidades de datos y estudiar así patrones, lo cual ha hecho posibles artículos de investigación basados en datos que, incluso ya en el año 2000, habían producido varios ganadores de premios periodísticos. Esto, según Hicks Maynard (2000, p. 19), significaba que la tecnología digital de la información no solo hacía que las noticias se abaratasen, sino que también incrementaba su calidad.

Esta evolución en el procesamiento de datos permitió la creación del “periodismo automatizado”, acuñado por primera vez por Bygrave (2001), quien imaginó un posible tipo de periodismo basado en la búsqueda y combinación automáticas de datos de diversas fuentes digitales. En la última década, diversos medios han utilizado robots que procesan

información para identificar y publicar novedades de prensa de diferentes fuentes en modo de paquete y distribuirlas a través de redes específicas, como en el caso de las novedades del mercado financiero. Ejemplos como Forbes y The New York Times han utilizado esta técnica, que denominan “tecnología de la web semántica”. Esto abre la posibilidad de liberar a los periodistas de parte de la labor de recolección de datos (Haak, et al., 2012), gracias también al acceso más abierto de los conjuntos de datos gubernamentales, la mejora del software y la economía digital en desarrollo. El periodismo automatizado o computacional, además, permite según Flew et al. (2012) una mejora tanto en la producción de investigación periodística original como en la atracción y retención de los usuarios.

Existe otro subcampo llamado “generación de lenguaje natural” (Natural Language Generation, NLG), y consiste en compartir la información, extraída de fuentes de datos, con un lenguaje humano. Es capaz de realizar tareas institucionalizadas a un nivel técnico y con la posibilidad de generar contenidos en varios idiomas a un coste más bajo que el humano (Dörr, 2016). Y no solo se habla de una reducción de costes sino también de errores, una mayor rapidez y un volumen mucho mayor. El procedimiento por el cual un algoritmo crea noticias sería el siguiente según Graefe (2016): recolectar datos, identificar eventos interesantes, priorizar la información, generar una narrativa y publicar una historia, como se puede observar en la Figura 11.

Figura 11. Elaboración de noticias por parte de los algoritmos (Graefe, 2016).

Sin embargo, Clerwall (2014) hizo un estudio sobre la percepción de los usuarios del periodismo automatizado atendiendo a su redacción y factores como la cualidad, credibilidad, objetividad, etc. Éstos los encontraron descriptivos y aburridos, pero objetivos y no necesariamente discernibles de los escritos por periodistas.

Es por ello por lo que han surgido otros estudios como el de Linden (2017) que, a través de 31 entrevistas, llegó a la conclusión de que el procesamiento de información siempre necesitará la competencia humana de la toma de decisiones periodística. El motivo es que un algoritmo de noticias no puede funcionar sin una estructura, selección, evaluación y filtro de los datos, para que estos estén estandarizados, normalizados y validados. Y todos esos procesos deberían estar bajo decisiones periodísticas. Otro motivo de preocupación de los expertos es que el acceso a estos datos deja de ser democrático y se convierte en económico. Afortunadamente hay muchos productores privados de datos que permiten a los medios de comunicación acceder a sus datos.

Los medios periodísticos también han disfrutado de las ventajas de la infraestructura de Internet, puesto que les ha provisto de una distribución más fácil, barata y rápida, según Veglis y Pomportsis (2004). Esto es gracias al sistema de recuperación de la información centralizado que supone Internet, desde el cual se puede exportar dinámicamente y de

manera personalizada el contenido de los periódicos de acuerdo con las preferencias del usuario. Además, le aporta herramientas como la búsqueda y la posibilidad de descargar los contenidos. Salwen et al. (1995, p. 257) observaron que, dependiendo del ámbito del medio, había una diferencia en la aparición de estas herramientas. En el caso de los periódicos online nacionales, había una implementación más alta de motores de búsqueda, registros para envíos personales y actualizaciones inmediatas. Los periódicos online regionales ofrecían una implementación más alta de foros, chats, otros idiomas y encuestas. Los periódicos online locales ofrecían más enlaces con información relacionada, audios, vídeos y emails. También permitían compartir los artículos, ya sea enviando el título y un enlace a la página web o el texto completo en HTML. Albornoz (2006, p. 201) comprobó que ElPais.es y LaNación.com ofrecían además un servicio de almacenar hasta 20 elementos en el primer caso y hasta 35 en el segundo caso con tiempo ilimitado. Todas estas herramientas permitían al usuario convertirse asimismo en una fuente de información, dándoles una función más allá de la simple lectura personal de los contenidos (Canavilhas, 2007, p. 218). Costera Meijer y Groot Kormelink (2015), en cambio, demostraron en un estudio de los patrones de uso de noticias publicadas entre 2004 y 2014 que prácticas como enlazar, compartir, indicar que un artículo gusta, recomendarlo, comentarlo o votarlo no habían sido tan centrales para el consumo de noticias como se había asumido con frecuencia.

2.1.1.6. Medición de audiencias online

En la primera década del periodismo online había diferentes métodos de control de audiencias. Según Caminos Marcel et al. (2006), algunas empresas se basaban en la elección de muestras de usuarios (metodología centrada en el usuario), grabando la actividad de sus ordenadores para medir qué webs visitaban, qué herramientas utilizaban... y extrapolando esos datos para extraer tendencias generales. Otros se basaban en encuestas, como era el caso de la EGM que realizaba 43.540 entrevistas anuales en tres oleadas.

Otro método (centrado en el sitio web) se basaba en el control de las visitas por los servidores del periódico digital a través de los logs o con un sistema de etiquetado o *tags*, pero este último contaba con el problema de que, si más de una persona usaba el mismo dispositivo, dichos datos se cruzarían entre las diferentes sesiones.

Todo ello ayudaba a analizar una audiencia cada vez mayor, ya que, explican Caminos Marcel et al. (2006), en 2005 ya se consideraba que más de un tercio de la población española era usuaria de Internet, un dato muy alejado, sin embargo, de otros países de la

Unión Europa (Suecia con un 76,8%, Holanda con un 66%, Dinamarca con un 62,5% o Reino Unido con un 60,6%). Tal y como expone un estudio de Internet World Stats (2021), en el primer cuatrimestre de 2021, el 64,2% de la población mundial es usuaria de Internet, llegando al 89,9% en Norte América y al 87,1% en Europa. Sin embargo, el 53,6% de todos los usuarios online son de Asia.

2.1.1.7. La publicidad en el periodismo digital

La publicidad también estaba presente en el periodismo digital. Albornoz (2006, p. 201) observó diversos formatos publicitarios, entre los cuales se encuentran: banners; banners flotantes o *Fly Ad*; vídeo banners; máscaras de audios, vídeos y fotos; marcas de agua; coberturas especiales.

Los medios tradicionales habían estado perdiendo audiencia paulatinamente desde 1980, pero no así su financiación, que incluso aumentó en los primeros años de la Web gracias al empuje de la publicidad online (Grueskin, et al., 2011, p. 8). No obstante, tras la crisis de 2008, la publicidad que sustentaba los medios tradicionales no se veía capaz de invertir en la era digital, según Kaye y Quinn (2010, p. 1). Esto provocó una crisis de financiación en el sector que hizo que decreciera el número de periodistas a tiempo completo en un 13% entre 2001 y 2008. Todo ello hacía cada vez más difícil mantener la cualidad y cantidad que el público exige en un medio periodístico.

Al cambiar el paradigma del consumo de periodismo, cambió asimismo el paradigma del consumo de la publicidad en el periodismo. Antes, la publicidad se veía obligatoriamente al pasar la página, pero al evolucionar hacia una navegación no lineal, dicha visualización era más atomizada. Sin embargo, argumenta Grueskin et al. (2011, p. 13), el alcance era mayor, ya que cualquier usuario con acceso a internet y un enlace o conocimiento de la URL podría acceder a la web del periódico, y la información de su navegación podía ser informada a los publicistas en tiempo real.

Blumler (2010) señala que el periodismo, por tanto, se enfrentaba a una crisis de dos dimensiones: viabilidad y adecuación cívica. La primera era principalmente de carácter financiero y bastante reciente, pero la segunda, que se debía al empobrecimiento de las contribuciones del periodismo hacia los ciudadanos y la democracia de la información, ya tenía una trayectoria de varias décadas y se había visto perjudicada por causa de la primera.

Esta viabilidad se trata de conseguir con los siguientes modelos de negocio basados en contenidos online, presentados por Gutiérrez Argüello (2013):

- Modelo Gratuito: se trata de un modelo tradicional transformado a digital y basado únicamente en publicidad, el contenido es gratuito y el precio de la publicidad varía según el tamaño, la profundidad, los ingresos y el tráfico obtenido.

- *Real Time Bidding*: combinación de la conducta del usuario, el lugar del anuncio, el evento o promoción en sí y el momento en que se comparte. Utiliza medidas como el CTR (*Click Through Rate*, proporción por clics), CR (*Conversation Rate*, proporción por conversación), CPM (*Cost Per Millar,* precio por mil impresiones), ECPM (*Effective Cost Per Millar*, precio efectivo por mil impresiones) y CPS (*Cost Per Sale*, precio por venta). En cuanto a los modos de pago, utiliza PPC (*Pay Per Click*, pago por clic), PPS (*Pay Per Sale*, pago por venta), PPI (*Pay Per Installation*, pago por instalación), PPA (*Pay Per Action*, pago por acción), PPV (*Pay Per Visual*, pago por visualización) y PPD (*Pay Per Download*, pago por descarga). En la Figura 12 se puede ver la comparación del modelo tradicional y *Real Time Bidding* (RTB), y en la Figura 13 se compara ambas segmentaciones.

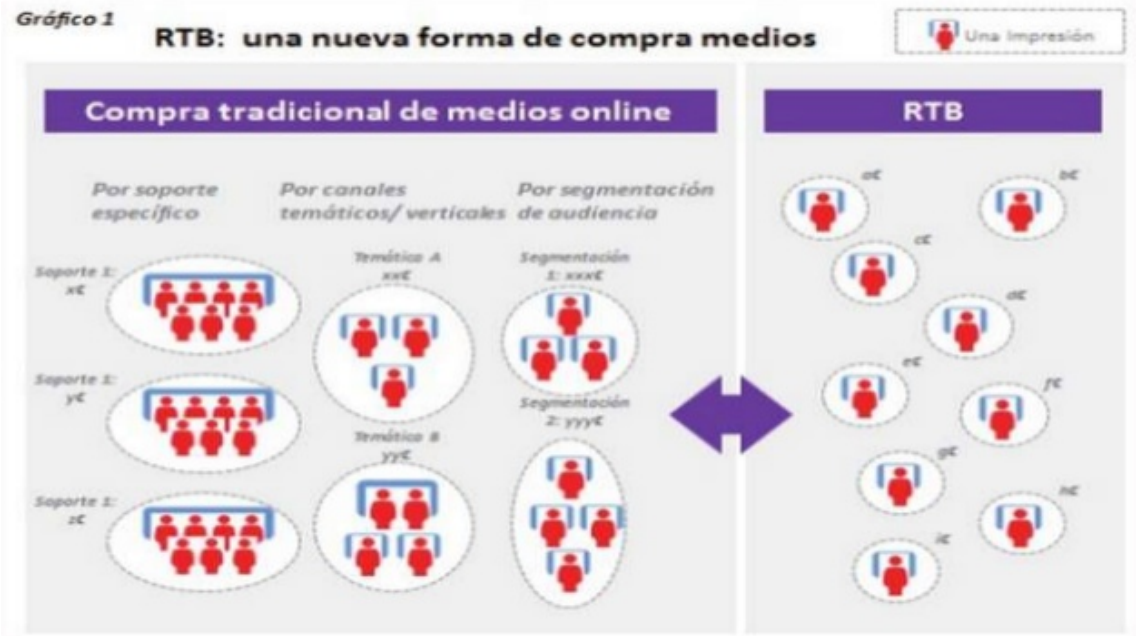


Figura 12. Comparación del modelo tradicional y RTB (Gutiérrez Argüello, 2013).

Figura 13. Comparación entre una segmentación normal y una basada en RTB (Gutiérrez Argüello, 2013).

- Suscripciones Gratuitas: los suscriptores dan sus datos para obtener el acceso gratuito, de manera que se pueda obtener referencias generales, segmentación de mercado y análisis de uso.

- *Pay Wall* y *Pay-Per-Use*: se cobra el consumo total de los contenidos de manera anual, mensual, semanal o, en el caso del *Pay-Per-Use*, a modo de micropagos semanales, diarios o por artículo.

- *Metered Model*: permite un consumo limitado de la información y, una vez consumido este, se solicita un pago para acceder al resto.

- *Freemium*: el contenido es gratuito salvo en el caso de algunos artículos considerados de mayor valor.

- Donaciones: se basa en donaciones voluntarias de los usuarios, siendo considerada por algunos autores como periodismo puro.

- Pago Total: los usuarios efectúan un pago anual para tener acceso a todos los contenidos.

- Pago por Horario: la información solamente está disponible en un horario concreto, normalmente en franjas horarias de menor tráfico, y el resto del tiempo es de pago.

- Venta de Información: la información es vendida a modo de noticias especiales, entrevistas, columnas de opinión, etc.

También ha habido casos de medios rescatados por los usuarios a través del *crowdfunding* o micromecenazgo. Revistas alternativas, coberturas independientes y otros proyectos periodísticos innovadores recurrieron a campañas mediante las cuales buscaban una financiación gracias a la solidaridad de sus usuarios. Sin embargo, solo el 1,7% de los encuestados en un estudio realizado por Sánchez-González y Palomo-Torres (2014) habían creado su empresa gracias a esta metodología. El emprendimiento en red todavía no estaba muy presente pese a que se cimentaba en la transparencia, la libertad informativa y una mayor conexión con la audiencia.

2.1.1.8. El periodismo y la Blogosfera

En la denominada por los medios como Web 2.0., el carácter social de Internet se hizo ver a través de los vídeos, las redes sociales, los podcasts y la Blogosfera, dando paso a una versión de Internet mucho más conectada socialmente (Anderson, 2007). Surgieron millones de blogs, suponiendo el mayor eco mediático de la información en la Red y formando uno de los espacios más dinámicos de esta. Sería por tanto la evolución natural de la función como fuente de información de los usuarios de la Red, desde la compartición de artículos periodísticos hasta la creación de sus propias bitácoras donde publicaban sus contenidos. Cuatro de las características principales de este entorno compartido por periódicos y blogs según Noguera (2009) serían: el diseño de sitios web con contenidos transversales, el aumento de cobertura informativa a través del *microblogging*, los relatos colaborativos y la gestión de identidades en los medios.

Todo esto provocó, según argumenta Noguera (2009), una fragmentación enorme entre una audiencia que no había crecido, ya que resultaba una utopía pensar que un usuario podría satisfacer todas sus necesidades informativas desde un único sitio. Se produjo un cambio de perspectiva, ya que en la era pre-internet toda información no publicada por los medios era despreciada, y con la llegada de la Blogosfera, los medios periodísticos tendrían que aceptar que son un eslabón más de la información sobre un tema debido a la multiplicidad de fuentes, la libertad de publicación y la inmediatez de los contenidos en Internet. De este modo, necesitaban asumir nuevas identidades y ocupar espacios donde la audiencia ya publicaba contenido: *microblogging*, vídeos, redes sociales...

En este entorno surgió una línea de actuación estratégica denominada *digital first* (primero digital), en la cual la edición en papel dejaba de ser una prioridad, dando paso a la versión online a nivel de negocio y servicios. Ejemplos de medios periodísticos que adoptaron esta metodología fueron The Guardian, The New York Times o The Washington Post. Una apuesta fuerte por el medio online que, según Guallar (2015), presionó en España a las

legislaciones vigentes a compensar económicamente en enero de 2015 con la denominada “tasa Google” o “canon Aede” a los medios periodísticos por el uso de contenidos por parte de agregadores e incluso blogs, motivando entre otros el cierre de Google News España en diciembre de 2014.

Los medios online tradicionales se plantearon asimismo incluir blogs como parte de su contenido, pero esto provocaba una tensión entre mantener una orientación unidireccional del periodismo y experimentar con las nuevas formas de comunicación dialéctica, según comprobaron Mitchelstein & Boczkowski (2009). Las principales críticas hacia ese tipo de contenidos eran que los usuarios que generaban contenidos y no eran periodistas, no hacían contribuciones necesariamente guiadas por normas editoriales, como la objetividad y la experiencia. Además, generaban otras diferencias organizacionales, como un acceso más limitado a recursos periodísticos y especialmente a nivel de producción: diferencias en contenido, procesos de trabajo, tono, valores y formato.

Los blogs, por su parte, representan la disponibilidad de noticias sin filtro para un público global, siendo descritos por Wright Brian como “la voz humana y la imaginación amplificadas por el poder de la Web”, citado por Quinn (2012, p. 30). El *blogging* es un modo de expresión que, si bien inicialmente se vinculó a principiantes, acabó convirtiéndose según Hernandez y Rue (2016, p. 3) en el método de cobertura más común de muchos nichos verticales. Entre sus principales características se encuentran un ritmo de publicación mayor, un número de palabras menor y un tono mucho más conversacional.

Ungria y Gaitan (2014) comprobaron que las publicaciones se habían ido adaptando cada vez más al marketing de contenidos y al SEO (Search Engine Optimization, optimización para buscadores), para por un lado dirigirse al público final con contenido de calidad y que, por el otro, esté optimizado para adquirir una buena posición según el algoritmo de Google y otros buscadores. Según el estudio “State of Search Marketing Report” de Econsultancy y SEMPO, citado por Ungria y Gaitan (2014), el 45% de las empresas de 25 países ya habían integrado el marketing de contenidos en sus estrategias SEO. La idea principal es generar contenido de calidad que fomente su consumo y su viralización.

2.1.2. Twitter

En esta tesis, parte del análisis se realizará ya sea de los contenidos publicados en Twitter por la cuenta del medio a estudiar, como de los contenidos de la propia red sobre una tendencia. Es, por lo tanto, necesario explicar en qué consiste dicha red social y cuáles son sus características principales.

2.1.2.1. La red social

Se trata de una plataforma de comunicación lanzada en octubre de 2006 por una *start-up* de 10 personas llamada Obvious (Honeycutt & Herring, 2009). Es considerado popularmente como un medio democrático según Small (2011), ya que permite la publicación de reportajes sobre el terreno, noticias de última hora y activismo democrático. Es portable, rápida, gratuita, flexible, amistosa y fácil de aprender y utilizar. Fortalece la escritura y facilita la síntesis, ya que obliga a transmitir ideas en textos muy breves, aunque Fainholc (2011) advierte de que esto último provoca un impacto disruptivo en la forma de expresión y comunicación.

Considerado *microblogging* o *moblogging*, Twitter combina los componentes de una red social, un blog y los mensajes de texto, y comenzó con un crecimiento acelerado, según Hamilton (2007). En marzo de 2007, medio año después de su lanzamiento, tenía 100.000 usuarios, el doble que un mes antes. Ojeda-Zapata (2008, p. 9) observó que en septiembre de 2008 ya era la red social con un crecimiento más rápido, registrando un 343% de crecimiento en los doce meses anteriores, desde los 533.000 usuarios a los 2.4 millones. Rápidamente se impuso al resto de alternativas de microblogging: en verano de 2008 ya tenía 12 veces más tráfico que Plurk y 24 veces más que FriendFeed. En mayo de 2009, Twitter ya tenía 18,2 millones de usuarios, con una ratio de crecimiento de 1448% desde mayo de 2008 (Marwick, 2010), y en 2010 contaba con 75 millones de usuarios (Murthy, 2011). Según Dean (2021), en febrero de 2021, Twitter contaba con 353 millones de usuarios activos mensuales, suponiendo el 10% de todos los usuarios de redes sociales, aunque el 80% de los tuits son publicados por el 10% de los usuarios.

Twitter se ha convertido, según Peñafiel Sáiz (2016), en uno de los medios de comunicación más poderosos de la historia gracias a su inmediatez y sus características transmedia, revolucionando la comunicación y la manera en que se transmite la información. La narrativa transmedia combina diferentes formatos de contenidos con un mensaje como núcleo, lo cual le permite ser muy amplia sin repetir mecanismos, algo de lo que pecaba la narración multimedia tradicional.

Twitter fue recibida con mayor entusiasmo por el público joven. Según Fox y Lenhart (2009), el 19% de los jóvenes de entre 18 y 24 años había usado Twitter y el 20% de entre 25 y 34 años, bajando a un 10% de entre 35 y 44 y un 5% entre 45 y 54 años. Esta declinación persevera con el aumento de la edad, con un 4% en el rango de 55-64 años y un 2% de 65 o más años. También había una mayor afluencia entre aquellos usuarios con menos ingresos y con acceso inalámbrico a Internet. El uso de Twitter, a su vez, también está relacionado con el uso de otras redes sociales, ya que el 23% de los usuarios de redes sociales admitían haber usado Twitter, mientras que solo un 4% de los que no usaban habían utilizado Twitter en el pasado. Los creadores de blogs también apostaban por Twitter en un 27%, frente al 10% de los que no tenían un blog.

Basada en la web, que combina el mensaje instantáneo y el SMS, los usuarios envíaban en un principio actualizaciones de estado de 140 caracteres o menos a sus seguidores como respuesta a la pregunta "¿Qué estás haciendo?", mensaje que se mostraba al usuario a modo de guía para motivar su participación. La limitación de caracteres y esta pregunta eran considerados fundamentales por sus creadores, ya que se trataba de una simple pregunta respondida por la comunidad de manera global. Cada usuario disponía de un perfil en el cual se podía ver el número de seguidores, cuentas a las que sigue, fecha de registro, descripción, localización y enlace a su web o blog personal. Los usos más frecuentes en esta fase inicial, según Mischaud (2007), eran enviar mensajes a otras personas conocidas por el usuario, publicar opiniones y pensamientos personales y compartir noticias con otros. Esta compartición de información personal y diaria podía incluso hacer factible la extracción automática de líneas de tiempo cronológicas de usuarios de Twitter desde su colección de tuits (Li & Cardie, 2014).

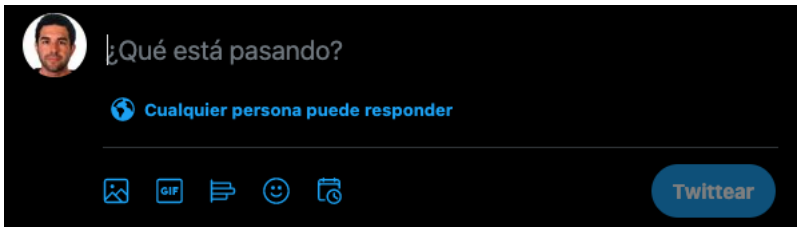


Figura 14. Muestra del formulario de envío de tuits en 2021, con la pregunta actual de "¿Qué está pasando?"

Las conexiones entre los usuarios de Twitter no son obligatoriamente bidireccionales, explica en su estudio Huberman et al. (2009). Cuando un usuario sigue a otro este último recibe una notificación de aviso, pero no está obligado a seguir al primero. Los mismos usuarios admiten que solo se mantienen en contacto con un pequeño número de sus seguidores, lo cual sugiere que la red social es menos densa de lo que parecería al analizar las conexiones simplemente por número de seguidores, y que muchos de los enlaces entre

los usuarios no son de la suficiente significancia a la hora de analizar las interacciones entre éstos. Por otro lado, según Kwak et al. (2010), solamente el 22,1% de los usuarios tenían una relación recíproca de seguimiento, y un 67,6% no eran seguidos por ninguno de los usuarios a los que ellos seguían, siendo para ellos Twitter probablemente una fuente de información más que una red social. En 2019, Estados Unidos, la mayor finalidad de uso de Twitter ha seguido siendo la de obtener noticias, junto con la del entretenimiento, ambas elegidas por el 48% de los participantes de una encuesta publicada por Tankovska (2021).

Por ello se puede delimitar el tipo de usuarios según su actividad e interacción con los demás, tal y como indican Krishnamurthy et al. (2008). Los usuarios con una gran diferencia entre el número de seguidores y de cuentas que siguen son considerados emisores de tuits, como radios y periódicos. Los usuarios que tienden a una reciprocidad en sus relaciones suelen ser usuarios que usan Twitter con sus conocidos. Los que en cambio siguen a muchos más usuarios que su número de seguidores son considerados *spammers*, ya que tratan de seguir gente para que les correspondan.

El 2 de noviembre de 2009, Twitter lanzó la herramienta de las Listas, en las que los usuarios pueden organizar usuarios según categorías de temática u otras, y decidir si dicha lista podrá ser accedida de manera pública o privada (solo el creador podrá usarla). Esta funcionalidad permite categorizar mejor a los usuarios y tratar de encontrar a la élite de usuarios que representan una temática concreta. Un estudio de Wu et al. (2011) demostró que las audiencias estaban volviéndose más fragmentadas, y que los medios de comunicación eran los usuarios más activos de Twitter.

2.1.2.2. Uso y tipos de contenidos

Según un reciente estudio de Wojcik y Hughes (2019) sobre el uso de Twitter en Estados Unidos, tal y como se puede observar en la Figura 15, el 73% de los usuarios de Twitter son menores de 50 años (el 44% en una edad comprendida entre los 30 y los 49) y el 42% de los usuarios adultos de Twitter tienen una carrera. Hay también una gran diferencia de actividad entre unos usuarios y otros en Twitter. El 80% de los tuits son publicados por el 10% de los usuarios, que producen un promedio de 138 tuits mensuales, mientras que el promedio del 90% restante es de solo 2 tuits al mes. En cuanto al uso de la plataforma, el 81% del primer grupo accede a la plataforma diariamente, mientras que del segundo grupo solo el 47% lo hacía con esa regularidad.

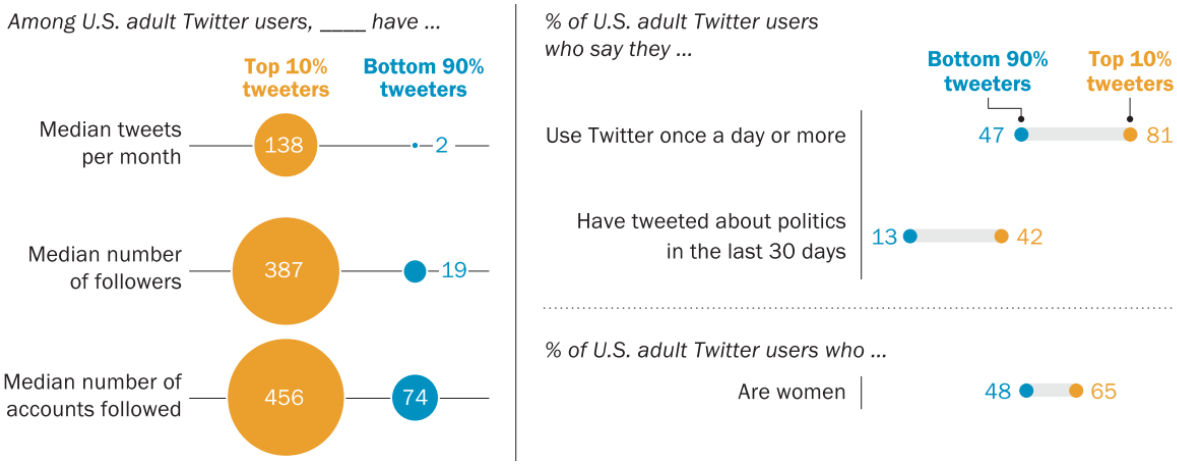


Figura 15. Uso de los usuarios adultos de Twitter a finales de 2018 (Wojcik & Hughes, 2019)

Los mensajes publicados, además, pueden ser directos o indirectos, según explican Huberman et al. (2009). Los mensajes directos están dirigidos a una persona específica y representan un 25,4% del total a finales de 2008, mientras que los indirectos son actualizaciones a nivel general. El número de mensajes, en cambio, crece junto con el número de seguidores hasta un punto en el que se satura, como nos indica el gráfico de la Figura 16:

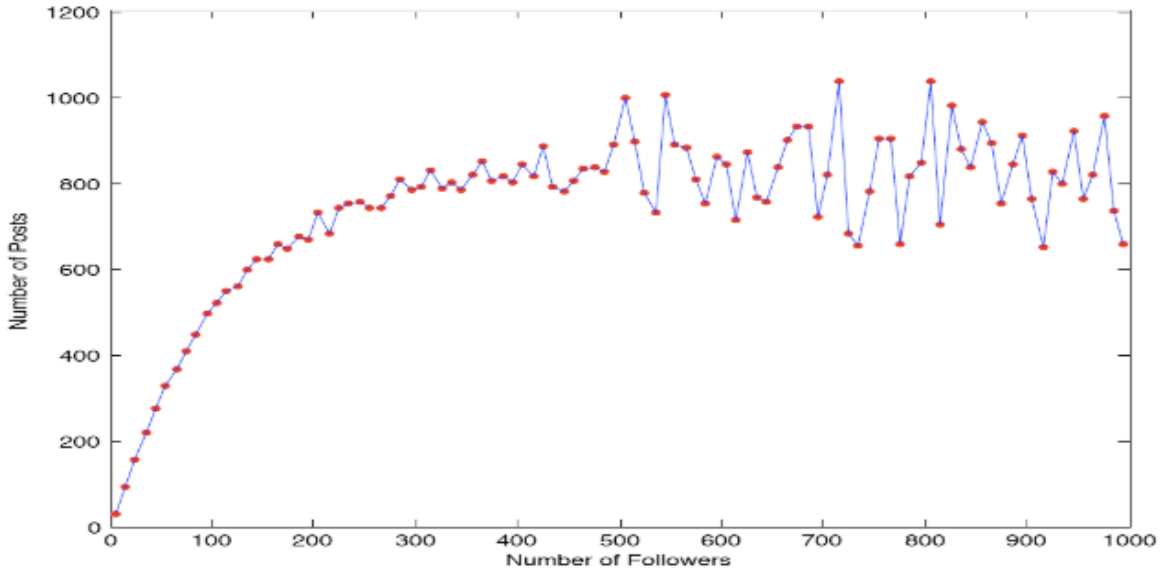


Figura 16. Correlación entre el número de mensajes y el número de seguidores (Huberman, et al., 2009)

Los mensajes dirigidos a otros usuarios van precedidos por el carácter "@" y el nombre de usuario, y aunque Twitter no fue diseñada inicialmente para ello, se convirtieron en una de las principales maneras de comunicarse en la red social a través de la interacción entre los usuarios e incluso, en ocasiones, con una extensión remarcable, según Honeycutt y Herring (2009). Los usuarios utilizaron esta función para conversaciones diarias, compartir

información o URLs, compartir noticias y crear conversación en general. La necesidad del carácter "@" se producía a la hora de crear una coherencia online en los intercambios de una persona a otra, en la que seleccionar a quién se envía dicho mensaje era necesario para evitar una disrupción de la información en un entorno tan rápido con múltiples participantes. Además de los dirigidos a otras personas a través del formato de respuesta "@", también se pueden enviar mensajes directos (DMs), aunque la comunicación predominante es pública (Marwick, 2010).

Los retuits o comparticiones de un tuit se designan como RT y los hashtags con el carácter "#", elegidos por consenso y comportamiento de la comunidad, en un esquema multidimensional de clasificación de contenidos en Twitter, tal y como indica Darn (2010).

Los hashtags sirven para focalizar conversaciones en una temática concreta, y consisten en el carácter "#" y una palabra clave. Los usuarios los utilizan a modo de etiquetar los tuits emulando a otros sistemas como los blogs (Boyd, et al., 2010). Twitter los incluyó a partir de febrero de 2008, ya que no estaban inicialmente debido a la naturaleza conversacional y no organizacional de la plataforma. De esta manera, se permite a los usuarios filtrar, dirigir contenido hacia determinados hilos, utilizarlos como un impulso de la interacción e incluso crear sus propios hashtags para englobar un evento o fenómeno que dejarán de ser usados cuando este finalice. Según Huang et al. (2010), no es común que un tuit contenga más de un hashtag, y el hecho de tener uno incrementa las posibilidades de que el tuit aparezca agrupado bajo un "*trending topic*" o "TT", es decir, una de las temáticas más utilizadas en el momento actual. Davis (2020) especificó que la participación puede aumentar entre un 50% y un 90% al usar los hashtags correctos, aunque la página oficial de ayuda de Twitter recomienda usar un máximo de dos hashtags por tuit (Twitter, 2020).

Estos métodos de indexación saltaron a la fama sobre todo a partir de determinados eventos sociales como #ArabSpring, #OccupyWallStreet o #SpanishRevolution en 2011, y establecen una nueva fase en la evolución de la sociedad digital en la que dicha indexación permite a los usuarios agruparse y conectarse más fácilmente por afinidades sociales, ideológicas o culturales. Es así como surgió el concepto de Generación # o Generación hiperdigital con respecto a la Generación @ o Generación digital que había existido hasta entonces, y que Feixa y Fernández-Planells (2014, p. 35) comparan en su estudio. Una generación que participa en una conversación global de conexión constante y descentralizada con un carácter colaborativo. Las informaciones dejan de expandirse de manera secuencial y lo hacen de manera viral, multiplicando su visibilidad de manera exponencial y en oleadas.

Los retuits son publicaciones de tuits de otros usuarios en tu propio perfil. Boyd et al. (2010) explican que se pueden considerar como un acto de copia y redistribución de información, pero que promueve una ecología conversacional en la que se produce un sentimiento de contexto conversacional compartido por voces de manera pública. Ayudan a traer a nuevas personas a un hilo particular, invitándoles a participar sin tener que interactuar de manera directa con ellos, por lo que puede cumplir al mismo tiempo una función de difusión de información y de participación en la difusión de la conversación. Además, ayudan a mantener la referencia de los mensajes publicados en un entorno de contenidos distribuidos por toda la red. El prototipo inicial consistía en copiar el mensaje precedido por "RT" e incluir una mención al autor original.

Esta práctica derivó en la posibilidad de cuestionar la autenticidad de la información que se comparte, puesto que podría servir como mecanismo para difundir falsos rumores frente a los datos verídicos. No obstante, según Mendoza et al. (2010), sería posible detectar si la información de un tuit está siendo cuestionada por muchos otros usuarios y comprobar si dicha información es fiable, ya que los falsos rumores suelen ser mucho más cuestionados que las noticias verídicas.

Wu et al. (2011) especifican que Twitter es un subconjunto de un ecosistema mediático mucho más grande, en el que el contenido existe y es descubierto repetidamente por los usuarios de Twitter. Algunos contenidos tienen un corto periodo de relevancia, como las noticias diarias, tras el cual es improbable que sean reintroducidos o redifundidos. Otros, como vídeos de música o artículos de revista, pueden tener un periodo de relevancia mucho mayor que permitan su redescubrimiento por los usuarios de Twitter, sin perder relevancia.

Según Hansen et al. (2011), el tipo de contenido y el tono de este afecta además a la viralidad. Los sentimientos negativos en el segmento de las noticias incrementan la viralidad, mientras que son los positivos los que consiguen dicho incremento en los contenidos que no son noticias.

2.1.2.3. Escucha activa

Para Ojeda-Zapata (2008, p. 8), una de las principales funciones de Twitter para las empresas es la de escuchar a los usuarios. Esta información es recibida de manera pasiva, absorbiendo la información en su toma de decisiones, o activa, ofreciendo a su vez información en Twitter e interactuando con sus clientes. Se trata de datos muy valiosos, ya que la rapidez y comodidad de Twitter contribuye a que los usuarios la utilicen para

desahogar sus frustraciones y, si sus tuits contienen desinformación o confusión, pueden desembocar según el mismo autor en falsos rumores que provoquen una crisis de marca.

Muchas marcas utilizan Twitter como canal de comunicación corporativa y publicitaria para participar en una era de marketing en tiempo real. Según Castelló Martínez et al. (2014), atendiendo a variables como el perfil de la empresa, la frecuencia de publicación, el tipo de publicaciones y la respuesta e interacción de los demás usuarios, se puede comprobar la capacidad de difundir contenidos y generar conversación, así como la reputación obtenida en Twitter. Es necesario además un seguimiento en tiempo real, transmitir una sensación de transparencia y una escucha activa.

Marwick (2010) señala como necesario, por tanto, un concepto de público para el punto de vista del emisor de los mensajes, para así crear una narrativa, ideología e imagen de uno mismo y captar la atención del público. La audiencia ideal es según este autor, a menudo, una imagen reflejada del mismo usuario, que necesita cumplir con las expectativas de autenticidad de cara a un concepto abstracto de la audiencia online que varía entre los usuarios de Twitter. Es por ello necesario navegar entre las diferentes audiencias gracias a unos objetivos e incluso usuarios diferentes.

Todo lo anterior se ve beneficiado debido a que, según comprobó Ryan et al. (2008) el 85-90% de los usuarios deja su perfil de Twitter público, lo cual permite que otros servicios puedan acceder a esos datos y procesarlos gracias a la API (*Application Programming Interface*) abierta de Twitter. Las API abiertas son un conjunto de funciones que permiten a los desarrolladores crear servicios que ayuden funcional y estéticamente a la comunidad. Se trata de un concepto en el que se profundizará más adelante en la tesis, ya que el estudio aprovecha esa escucha activa para en este caso tratar de predecir qué contenido tendrá una mayor acogida. Gracias a esta característica de las API abiertas, los desarrolladores externos pueden realizar el mantenimiento y la gestión de datos y cuentas a través de programas basados en web, aplicaciones, scripts automatizados y móviles.

Esto provoca una ampliación en la gama de funcionalidades potenciales de Twitter, teniendo repercusiones incluso en la vida real y su correspondiente realimentación con la participación en Twitter. Ofrece a la comunidad una manera de controlar su propia experiencia, así como promover la escucha activa de los usuarios. Algunos ejemplos podrían ser los dispositivos de muestra de actividad en Twitter. En junio de 2007 ya había 100 clientes externos, como Twitteriffic y Twitteroo que permitían enviar y recibir tuits en aplicaciones de escritorio (Honeycutt & Herring, 2009). Sin embargo, las API incentivan la creación y la existencia de robots en una plataforma. Según Gilani et al. (2017), en 2014,

Twitter reportó que 13,5 millones de cuentas (un 5% del total) eran falsas o spam. Su comportamiento se basa más en tuitear URLs, retuitear y un uso de tendencias menos intuitivas. Sin embargo, su comportamiento puede tener un gran efecto en los entornos de redes sociales.

Una vez un usuario ha establecido una red de contactos de profesionales, amigos o autoridades relacionadas con sus intereses, Twitter puede adquirir la funcionalidad de servir como plataforma para realizar preguntas sobre servicios, aplicaciones, problemas técnicos o dudas sobre temáticas especializadas, tal y como asevera Wilson (2008).

2.1.3. Difusión de las noticias en Twitter

Puesto que en esta tesis se analizan principalmente noticias de última hora, tras haber comprobado la evolución del periodismo digital desde sus orígenes, es interesante comprobar también su evolución en Twitter y cómo se ha utilizado esta plataforma en la difusión de contenidos informativos.

2.1.3.1. La adaptación del periodismo a Twitter

Las redes sociales forman una estructura vital para la difusión de las noticias y que estas sean significativas para los usuarios. Heikkilä y Ahva (2015) estudiaron durante un año la relación entre estas y las concepciones de la audiencia de los periódicos, partiendo de las primeras a través de las rutinas, la interpretación y la acción pública de los usuarios hasta llegar al punto de vista de los diarios. Este análisis conceptual proponía de esta manera una dirección contraria a la habitual, utilizando el momento de la lectura como punto de unión de ambos ambientes, como se observa en la Figura 17. De esta manera pudieron observar que las prácticas rutinarias de los usuarios en el consumo de noticias se replicaban también en las redes sociales y que, pese a la divergencia de fuentes de información, las noticias seguían siendo una fuente importante y un punto central de referencia para la comunicación interpersonal.

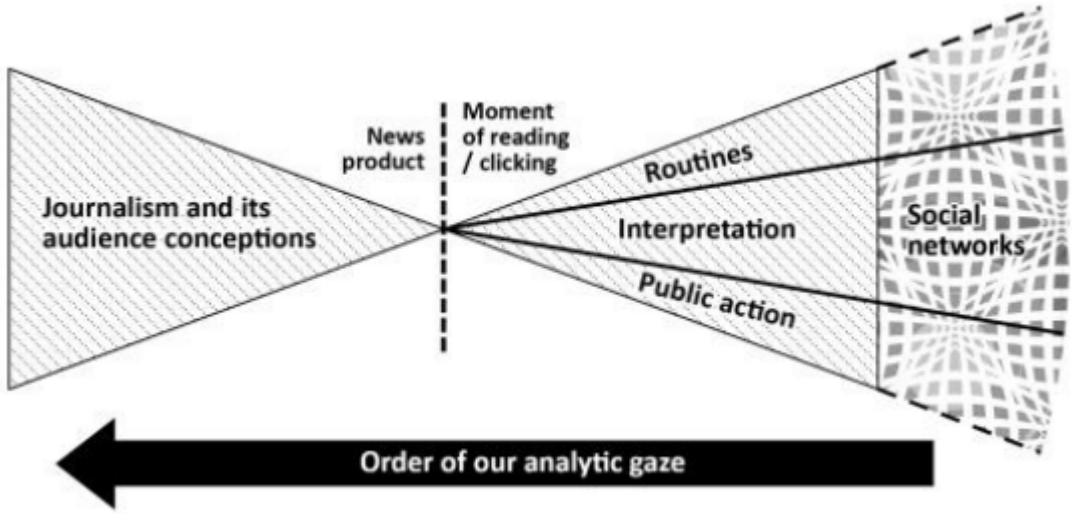


Figura 17. Marco teórico de entendimiento de la relevancia de las noticias en las redes sociales (Heikkilä & Ahva, 2015).

En cuanto al uso de las redes sociales por parte de los medios periodísticos, Bastos (2015) estudió el uso de periódicos como The Guardian y The New York Times y encontró que Twitter era la única red social con unas correlaciones estadísticas significativas entre la distribución de las noticias por secciones desde el momento en que las noticias se publican

en internet y son tuiteadas. Esto indicaba que Twitter es el canal de preferencia para difundir una gran variedad de contenido de noticias, a la cual también se añadiría la variedad de información compartida por los usuarios de Twitter en sí. También encontró una relación lineal entre el perfil editorial de los periódicos online y Twitter, y una proporción muy parecida entre los artículos de noticias publicados por el medio en su web y en sus perfiles sociales. En otro estudio realizado por Osborne y Dredze (2014) se comparó Twitter con Facebook[5] y Google+, y comprobaron que Twitter era la red social más rápida en disponer de información sobre eventos y noticias de última hora, aunque sin superar nunca a los medios tradicionales en su estudio. Hay incluso una significativa asociación entre los valores de sentimiento (tono, desacuerdo, etc.) y los valores de mercado (volatilidad, volumen comercial, etc.) según otro estudio de Sprenger et al. (2014), lo cual indica que Twitter puede ser considerado un intermediario de información que crea también nuevos contenidos, compila los existentes y los disemina en tiempo real, y esto exige a los usuarios una detección de tendencias y priorización de señales.

Diversos autores como Hermida (2013), Ahmad (2010) y Barnard (2012) parecen estar de acuerdo en que Twitter es una herramienta periodística muy útil, y esta se ha hecho presente también en España, tal y como indican Arrabal-Sánchez y De-Aquilera-Moyano (2016). En 2016, el 61% de los comunicadores españoles usaban Twitter, una cifra inferior a la esperada, y sin embargo su influencia está por encima de la media, medida con el índice Klout (índice de la empresa Kout muy utilizado hasta 2018), número de seguidores, número de tuits diarios, tuits favoritos, retuits, presencia en listas, etc.

Con la adaptación a esta red social, se produjo una serie de críticas debido a que, si la longitud de los artículos estaba decreciendo con el tiempo, Twitter podría impulsar esta tendencia hasta un límite muy alejado de lo que se considera periodismo. Dickerson (2008) habló por tanto de periodismo Twitter o microperiodismo como una reducción al absurdo del periodismo en este sentido, ya que el autor temía una reducción de la involucración del público con las noticias al dedicar demasiada atención a leer y responder tuits, sin profundizar en la información recibida. Sin embargo, también se podría interpretar que las publicaciones de Twitter construían a una comunidad de lectores que encuentra el camino hacia artículos más largos gracias a esta red social.

Hermida (2010) llamó “periodismo de ambiente” al basado en fragmentos digitales de información y noticias diseminados en tiempo real en la periferia de la conciencia de los usuarios, motivado por redes sociales como Twitter en las que los periodistas comparten

[5] https://www.facebook.com/

la jurisdicción con su público, ya que participan en la observación, selección, filtro, distribución e interpretación de los eventos. Todo ello provoca una omnipresencia de las noticias en la era digital a modo de experiencia social, en la que se convierten en una actividad participativa. Los usuarios aportan sus propios puntos de vista, experiencias y reacciones a los eventos que se publican.

2.1.3.2. Twitter y la comunicación con la audiencia

Según Herrera y Requejo (2012), los medios de comunicación que comprenden el potencial de Twitter utilizan una voz humana, los retuits, menciones a usuarios no relacionados con el medio y los enlaces externos para enriquecer sus contribuciones, así como escuchar y hablar con sus usuarios y realizar encuestas. Promueven el contenido más relevante de una manera atractiva con hashtags, multimedia como imágenes, vídeos, audios o gráficos e incluso enlaces a otras redes sociales donde estuvieran también presentes. Incluso utilizan Twitter como fuente de noticias a la que citan para apoyar o ilustrar una historia, llegando en ocasiones a realizar coberturas de noticias basadas en tuits (Broersma & Graham, 2013).

Para Al-Rawi (2019), uno de los principales objetivos es hacerse eco de noticias virales, es decir, noticias que se difunden en las redes sociales de una manera más rápida y amplia que las demás. Es difícil saber con anterioridad qué obtendrá una gran popularidad de una manera súbita, pero el mismo autor argumenta que se suele buscar ser relevante para todos y con un uso práctico. Los elementos que suelen fomentar esta viralidad son lo inesperado, extraño o sorprendente y con una significancia social (eventos económicos, culturales, públicos, políticos...). Al-Rawi (2019) comprobó también en su estudio que el tono de los tuits de noticias era positivo en un 77,1%. Temáticas como las noticias sensacionalistas y sobre celebridades también suelen tener efectos virales (Kalsnes & Larsson, 2018).

Sin embargo, Albero-Gabriel (2014) argumenta que los medios en general utilizan Twitter de una manera desigual y para usos muy concretos: noticias con contenidos audiovisuales y el mismo Twitter como fuente de consulta para aumentar el volumen de las noticias publicadas al incluir el análisis de la repercusión que han tenido los acontecimientos en las redes sociales y mantener el foco informativo sobre el suceso. Según Hochuli (2015), todo ello con la intención de incrementar el conocimiento de marca del medio de comunicación y sus contenidos en una audiencia nueva, como se observa en el embudo de conversión de la Figura 18.

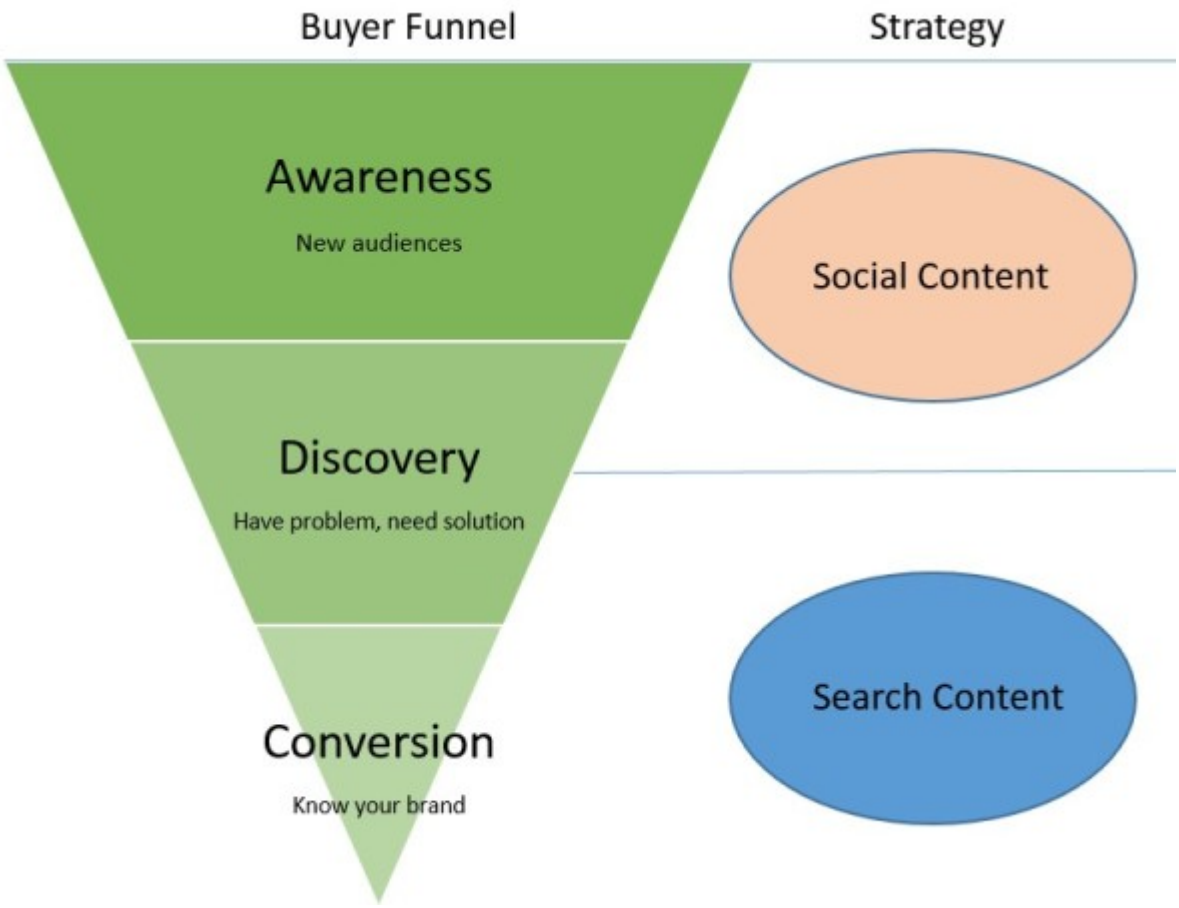


Figura 18. Embudo de conversión a través del contenido social y de búsqueda (Hochuli, 2015).

Pese al potencial que tienen las redes sociales, especialmente Twitter según Ju et al. (2014) puesto que es más efectivo que Facebook en alcance de audiencia, como canales de difusión de noticias, en 2014 no se había producido todavía una inclusión en el modelo de negocio de los medios de una manera clara. Tanto su contribución al tráfico web como las ganancias por publicidad habían sido decepcionantes comparadas con el resto de las canales, ya que los suscriptores de redes sociales eran solo una fracción pequeña de los que se habían suscrito a la web y a la versión impresa. El 0.8% de los tuits están relacionados con noticias, según un estudio de Malik y Pfeffer (2016), y se centran en hashtags diferentes de los utilizados por Twitter en general. Las cuentas de medios de noticias utilizan por tanto Twitter como un medio de comunicación profesionalizado y unidireccional en el que promover sus contenidos. Es por ello por lo que autores como Arrabal-Sánchez y De-Aguilera-Moyano (2016) consideran que con ello no se ha modificado el paradigma de la comunicación por falta de suficientes evidencias empíricas. Twitter solo ha afectado al medio y el formato de los mensajes, pero no al origen, proceso de fabricación ni destino de la información.

El uso de Twitter por los profesionales del periodismo no solamente podría indicar que éstos ven que sus ventajas eran superiores a sus desventajas, sino que quizá los propios medios de comunicación piden a sus reporteros que añadieran Twitter a sus rutinas diarias. De esta manera pueden proveer a la red social de cómo se han creado las historias, ofrecer notas personales de los eventos de las noticias y proveer de contexto para el desarrollo de la cobertura de noticias. Tratan de establecer así valores y normas periodísticas en las redes sociales como Twitter y, de paso, incrementar la sensación de transparencia y confianza en sus lectores. Todo ello ayuda a aumentar la influencia de los medios, así como la habilidad de filtrar noticias e información sin tener que recurrir a figuras como los editores de sección. Así, según Lasorsa et al. (2010) pueden divergir de sus papeles profesionales tradicionales y adoptar nuevas actitudes como las opiniones (16% de los tuits de periodistas ofrecían opiniones en 2010), incluir publicaciones de otros usuarios y usar enlaces, tuits y retuits a información externa que aporte contexto a sus artículos. Twitter ofrece a los periodistas la oportunidad de comunicar más allá de las limitaciones de las normas organizacionales o barreras de sus amistades en las redes sociales, y Bruns (2012) argumenta que eso le otorga una personalidad individual más que institucional y en algunos casos incluso una visibilidad mayor que la de los propios medios de comunicación que les emplean.

Vis (2013) estudió el uso de Twitter durante las revueltas de Reino Unido por los periodistas Paul Lewis y Ravi Somaiya y demostró que éstos utilizaron Twitter de manera extensa, aumentando de manera significativa su número de seguidores, lo cual demuestra que Twitter fue un buen instrumento de información por periodistas a nivel individual durante un evento importante. La red social, por tanto, es una importante plataforma para expertos sobre temáticas que no suelen tratarse con tanta asiduidad en los medios convencionales, como los problemas ambientales o la igualdad de género (Rogstad, 2016).

Twitter se convirtió, por tanto, en un campo de experimentación, innovación y participación para el periodismo. Prueba de ello es el uso del humor en sus publicaciones, que es más pronunciado cuanto más lo es la actividad del periodista en Twitter, sobre todo si este pertenece a un medio de comunicación menos selecto, según un estudio realizado por Holton y Lewis (2011). Estos autores argumentan que el humor ayuda también a provocar una sensación de conexión entre el medio y sus lectores, ayudándose de la narración de historias, la sátira, la ironía, la parodia y la comedia. Todo ello promueve una aceptación social, relajación y conectividad de un modo que, tal y como es la naturaleza del humor en sí mismo, es difícil de medir debido a su subjetividad. Los periodistas también publican contenido sobre ellos mismos buscando construir una marca personal, promoción de sus

artículos y una relación más estrecha con su audiencia, así como interpretaciones y análisis sutiles antes que opiniones más fuertes, según Molyneux (2014). Retuitean publicaciones de otros periodistas y fragmentos de opiniones que no planean incluir en sus propias historias, con lo que contribuyen en la conversación pública de un tema desde un punto de vista más profesional y cercano a las fuentes de información. Pero es más probable que promuevan contenido de las webs de sus medios que de otras fuentes, observaron Russell et al. (2015), y en el caso de hacer esto último, prefieren enlazar medios tradicionales que medios online. Por otro lado, utilizan más las herramientas de interacción como los retuits o las menciones con otros periodistas que con el público.

2.1.3.3. Twitter como fuente de noticias en tiempo real

Twitter es visto como una fuente de noticias en tiempo real, y el número de usuarios americano que acuden a esta red social para buscar noticias de última hora es el doble que en Facebook, según indica Hermida (2016, p. 81). Esta percepción de Twitter como plataforma de noticias es más evidente en tiempos de crisis, ya que es el medio preferido por los periodistas profesionales, organizaciones de noticias, grupos e individuos para generar hilos de noticias e informaciones híbridas en un proceso de transmisión, redacción y difusión. Twitter es una de las herramientas tecnológicas más usadas para intercambiar información y la construcción de fuentes informativas por parte del periodismo actual. Peñafiel Sáiz (2016) asegura que es, por tanto, la red social que más impacto ha generado en el Periodismo 2.0., ya que no solo involucra a la audiencia sino a la medición del éxito de los contenidos informativos. Las nuevas tecnologías de redes sociales como Twitter permiten, entonces, una diseminación digital instantánea de cortos fragmentos de información, y este nuevo modelo de flujo de noticias hace necesario según Hermida (2010) un ejercicio de recolección y transmisión de noticias por parte del usuario, por lo que podría ser necesario regular y negociar cómo se produce y mantiene esa conciencia de información.

Por tanto, Twitter puede ser utilizado por los usuarios como una manera de estar al día de todas las novedades periodísticas. Una vez el usuario realiza una selección de periodistas, blogueros, podcasters, personalidades mediáticas y emprendedores, este puede seguir sus actualizaciones y compartirlas con sus seguidores. Muchos de estos medios utilizan la red social para publicar sus artículos y podcasts, por lo que Twitter se convierte en una alternativa eficiente a los lectores de RSS (Wilson, 2008). Twitter es además una buena fuente de temas relacionados con entidades, como marcas y celebridades, cuya cobertura es menor en los medios de noticias tradicionales según un estudio de Zhao et al. (2011), y además ayuda de manera activa en la difusión de noticias de los eventos más importantes

a nivel mundial, pese a que los usuarios de Twitter no tienen tanto interés en las noticias de cobertura mundial.

Es por ello por lo que, para muchos usuarios de Twitter, la principal funcionalidad de la red social es la de consumir noticias, puesto que según comprobaron Fox y Lenhart (2009) éstos tienen una especial predilección por leer los periódicos online frente a los impresos (76% frente al 60% de los no usuarios de Twitter) y utilizar el móvil para ello (8% frente al 1% de los no usuarios de Twitter). Además, leen mucho más contenido de blog (57% han leído alguna vez un blog frente al 29% de los que no están en Twitter):

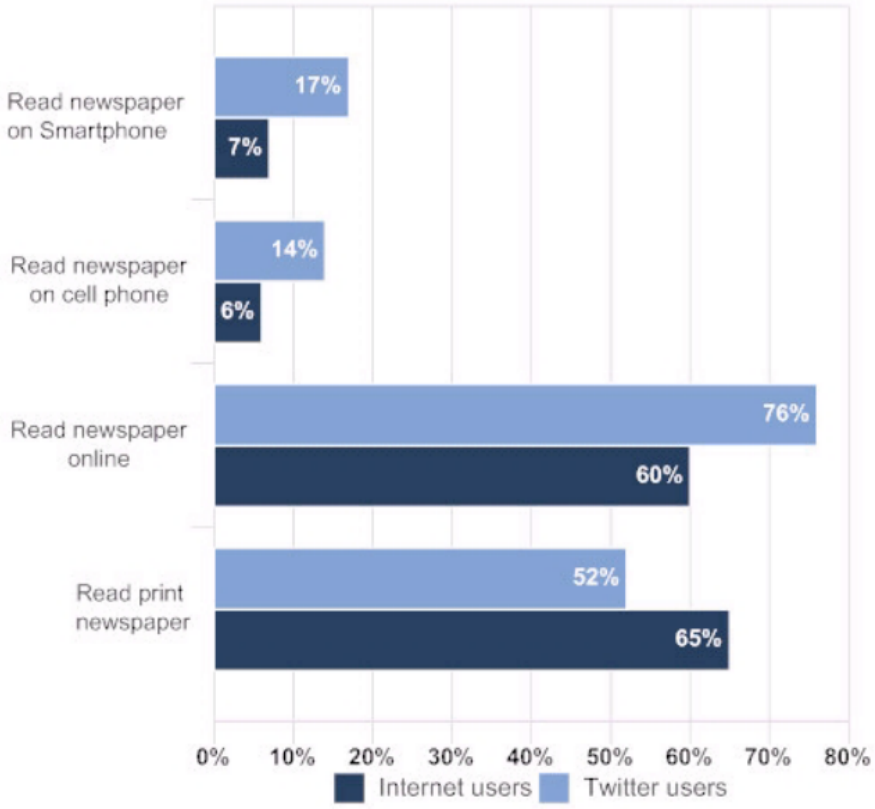


Figura 19. Lectura de periódicos según el medio de lectura y si son usuarios de Twitter o no (Fox & Lenhart, 2009)

Johnson (2016) estudió el consumo de noticias de última hora en el segmento de los estudiantes de universidad, y llegó a la conclusión de que éstos obtenían en primer lugar las noticias de última hora en Twitter, pero que sin embargo esta red social también servía como origen hacia otras fuentes de información más tradicionales. Es por ello por lo que Twitter también servía para aumentar el consumo de contenidos en las webs de noticias online.

A la hora de compartir las noticias a modo de retuit, los usuarios suelen basarse en criterios de contenido (es decir, un alto valor informacional) y criterios de contexto, ambos mediados

por la importancia percibida de los tuits. Rudat y Buder (2015) estudiaron la influencia de estos criterios en la compartición de noticias en Twitter y consideraron que el valor informacional sería considerado alto si afecta a una gran audiencia y tiene el potencial de provocar un cambio de comportamiento o reestructuración mental en los que lo reciben. El contexto sería el número de usuarios que han retuiteado y por tanto participado también en el proceso de difusión de la noticia. La influencia del valor informacional se ve aumentada si el autor no forma parte del grupo de sus contactos, al contrario de lo que se produce con los criterios de contexto. Los motivos principales que llevan a un usuario a querer compartir un contenido pueden ser, según Bright (2016), el aumento de estatus social al compartir noticias que consideran importantes, hacer denotar los intereses del usuario a través de las temáticas del contenido, así como evitar los que podrían perjudicar su imagen, y la tendencia a compartir ciertas temáticas mientras que otras solo se suelen leer (como es el caso de las noticias sobre crímenes).

En cuanto a la diversidad de este consumo de noticias, Morgan et al. (2013) estudiaron si la compartición de noticias en Twitter estaba ideológicamente sesgada, y descubrieron que, con el aumento de tuits enviados, aumentaba también la diversidad de la información que compartían, incluyendo no solo la que percibían como de acuerdo con su ideología (como la teoría de exposición selectiva predice), sino incluso la de puntos de vista opuestos.

Lehmann et al. (2013) comprobaron que algunos usuarios dedicaban mucho esfuerzo y cuidado a la labor de compartir noticias en Twitter, monitorizando una gran cantidad de fuentes de una temática o historia particular y seleccionando el contenido interesante para compartirlo con su audiencia, que podía alcanzar los millones de personas. Llamados "*news curators*", estos usuarios podían realizar este trabajo junto con sus comentarios personales y con una escritura personalizada o, por el contrario, automatizada compartiendo titulares y URLs.

Conforme Twitter gana más significancia como fuente de noticias válidas, sobre todo en el caso de situaciones de emergencia y eventos importantes, resulta relevante poder medir la credibilidad de la información online publicada en la red social. Algunas características que proveen de credibilidad son, según Castillo et al. (2011), la inclusión de URLs, tener árboles de propagación profundos, el número de mensajes de los autores, el número de retuits y un origen en común entre unos pocos usuarios de la red. No obstante, en un estudio de Vosoughi et al. (2018) se observó que las noticias falsas suelen difundirse de manera más rápida, profunda y amplia, quizá por una sensación de novedad y porque suelen buscar la provocación de miedo, disgusto y sorpresa en sus respuestas. Los

rumores se difunden a través de una o varias cascadas de comparticiones del contenido original, produciéndose incluso de manera más generalizada que en el caso de las noticias reales. Curiosamente, en algunos casos se producen más retuits que clics en el enlace de la información, lo cual demuestra que muchos usuarios retuitean sin leer siquiera la información antes y puede ayudar en la difusión de noticias falsas (Holmström, et al., 2019).

2.2. Medición web de éxito

Puesto que esta tesis consiste en tratar de obtener la predicción de ciertos valores entendidos como éxito por el editor de un medio, resulta necesario enmarcar qué se entiende cómo éxito en la analítica web, la cibermetría, Twitter, el análisis de tendencias en Twitter y la publicidad digital.

2.2.1. Analítica web

El primer apartado de indicadores de esta tesis procede principalmente de la analítica web, por lo que se revisará a continuación los antecedentes de investigación relacionados con ese ámbito, el periodismo digital y qué se entiende cómo éxito a la hora de analizar el contenido de una web.

2.2.1.1. Periodismo medible

Los medios de comunicación han tratado de adaptar las herramientas de Internet al contenido de sus webs para aumentar su visibilidad y accesibilidad, según Rodríguez-Martínez et al. (2012). En ocasiones han creado sus propias herramientas para cumplir con sus necesidades particulares de comunicación y han establecido relaciones con sitios web que reúnan las características propias de la Web y la aceptación del público online. De hecho, respecto al papel del reportero profesional, este ha ido perdiendo parte de su autonomía en la toma de decisiones de las noticias debido a que se ha vuelto más dependiente de las métricas de audiencia (Anderson, 2011). Estas métricas también tienen un efecto en los editores tanto en su labor como en su formación, lo cual le da un énfasis mayor a la audiencia a la hora de tomar decisiones relacionadas con la línea editorial de las noticias (Vu, 2014). Esto resulta en un proceso de selección que tiene menos en cuenta aquello a lo que la audiencia no responde tan bien, puesto que según Tandoc (2014) esta es entendida como una forma de capital para el medio, lo cual lleva a un debate moral sobre la distinción entre el interés público y lo que le interesa al público, así como si esta preferencia por el público puede hacer decaer en demasía la autonomía del periodismo (Tandoc & Thomas, 2015). También ha tenido como consecuencia la aparición de puestos de trabajo orientados a la audiencia que aportan datos relacionados con el éxito de artículos anteriormente publicados y sugieren líneas de acción para mantener o mejorar esas tendencias, teniendo un papel más directo en el proceso editorial (Ferrer-Conill & Tandoc, 2018).

Sin embargo, otros estudios como el de Brolin et al. (2017) indican que los periodistas digitales no describen la analítica web, es decir, el análisis de datos desde los sitios web para mejorar la experiencia del usuario, como una herramienta obligatoria ni necesaria para el proceso de selección de noticias, e incluso la academia asegura que no hay suficiente conocimiento que pueda demostrarlo así. Sí que contribuye, en cambio, en la percepción de la inestabilidad económica a un nivel organizacional y en el deseo de mantener a los lectores produciendo contenido que les interese. De la analítica web se hablará extensamente en este capítulo, pero esto último es, según los responsables entrevistados en un estudio de Belair-Gagnon y Holton (2018), desde un punto de vista más de viabilidad económica que de valores en la producción de noticias. Es decir, que mejora la visibilidad del buen contenido y puede ayudar a formar un modelo de negocio sostenible.

Esta información sirve, por tanto, para comprender mejor lo que pasa por la mente de la audiencia a la hora de consumir los contenidos, gracias a las métricas de su navegación por la web. Lee et al. (2014) realizaron un estudio basado en un análisis con desfase de tiempo y descubrieron que los clics de la audiencia afectan al posicionamiento de las noticias en la web ya que se busca una reciprocidad con ello, un efecto que se intensifica con el transcurso del día y que no se produce en el sentido contrario.

La analítica web está fomentando una diferenciación funcional dentro del periodismo que, según Hanusch (2017), se segmenta en tres tipos: contexto organizacional, orientación de mercado y según la plataforma de distribución. Todo ello podría producir una mayor divergencia en la profesión del periodismo, así como una mayor variedad de prácticas, normas y valores atendiendo a lo que la analítica sugiera. Es, también, necesario analizar los canales de Twitter, Facebook y diferentes dispositivos para completar el estudio, así como la hora del día ya que se está convirtiendo en un factor mediante el cual se modifica aquello que se muestra al público. Hanusch y Tandoc (2019) añaden que los comentarios de los lectores son también importantes a la hora de incrementar la importancia percibida tanto de la orientación de la comunicación al consumidor como al ciudadano, mientras que la analítica web ayuda a mejorar la importancia percibida por la orientación al consumidor.

Esta nueva forma de periodismo, que Carlson (2018) llama “periodismo medible”, incluye un cambio cultural y material en las plataformas digitales ya que estas pueden medir en tiempo real, individualizado y de manera cuantitativa los datos de audiencia y el consumo de esta. Adquiere, de este modo, las siguientes nuevas dimensiones según este mismo autor:

a) Material: Los programas utilizados para la analítica digital y la infraestructura sobre la que se sustenta, así como las plataformas de noticias digitales.

b) Organizacional: Nuevos puestos de trabajo en las salas de prensa, consistentes en el monitoreo de las métricas y la reacción ante estas.

c) Práctica: Uso de los datos para las decisiones de contenidos, tanto en la asignación como en el posicionamiento de éstos, y la elaboración de nuevos modelos de participación con la audiencia.

d) Profesional: Apoyar o resistir el uso de las métricas de audiencia en las decisiones periodísticas, la preocupación sobre la autonomía profesional y la intención de conectar más con la audiencia.

e) Económica: Cambios en la monetización del contenido, en la selección de recursos, personal y anunciantes según los datos, y fomentar el clic en los anuncios.

f) Consumo: Contenido personalizado y recomendaciones de contenido.

g) Cultural: Preocupación sobre la influencia de la popularidad sobre las decisiones, así como el debate entre las noticias individualizadas y las colectivas.

h) Políticas públicas: Preocupación sobre la privacidad de la audiencia.

2.2.1.2. Minería de uso de la web

La minería de uso de la web, entendida como la comprensión de la experiencia online para que pueda ser mejorada, ha tenido muchas definiciones. Guy Creese la definió como el monitoreo y reporte del uso de la web de manera que se puedan comprender las complejas interacciones entre las acciones de los visitantes y lo que la web ofrece, así como el uso de esa información para el incremento de la lealtad y las ventas de los clientes. Sarner y Janowski hablaron de una variedad de datos y fuentes que evalúan el rendimiento de una web y la experiencia de usuario con el objetivo de mejorar ambos desde una perspectiva combinada de técnica y contenidos e identificar oportunidades y riesgos. En el caso de Peterson (2004, p. 6) en su libro *Web Analytics Demystified*, citó las definiciones anteriores y la definió como “la evaluación de una variedad de datos que incluyen el tráfico de la web, transacciones basadas en la web, el rendimiento del servidor de la web, estudios de usabilidad, la información aportada por el usuario y las fuentes relacionadas que ayudan a crear un entendimiento generalizado de la experiencia del visitante online”.

La minería de uso de la web es, por tanto, la minería de datos aplicada al descubrimiento de patrones de uso en los datos de la web, de manera que sirvan para entender y ayudar a satisfacer las necesidades de las aplicaciones web. Algunos autores como Srivastava et al. (2000) la definen como un procedimiento consistente en tres fases: preprocesamiento, descubrimiento de patrones y análisis de patrones, como se observa en el esquema de la Figura 20:

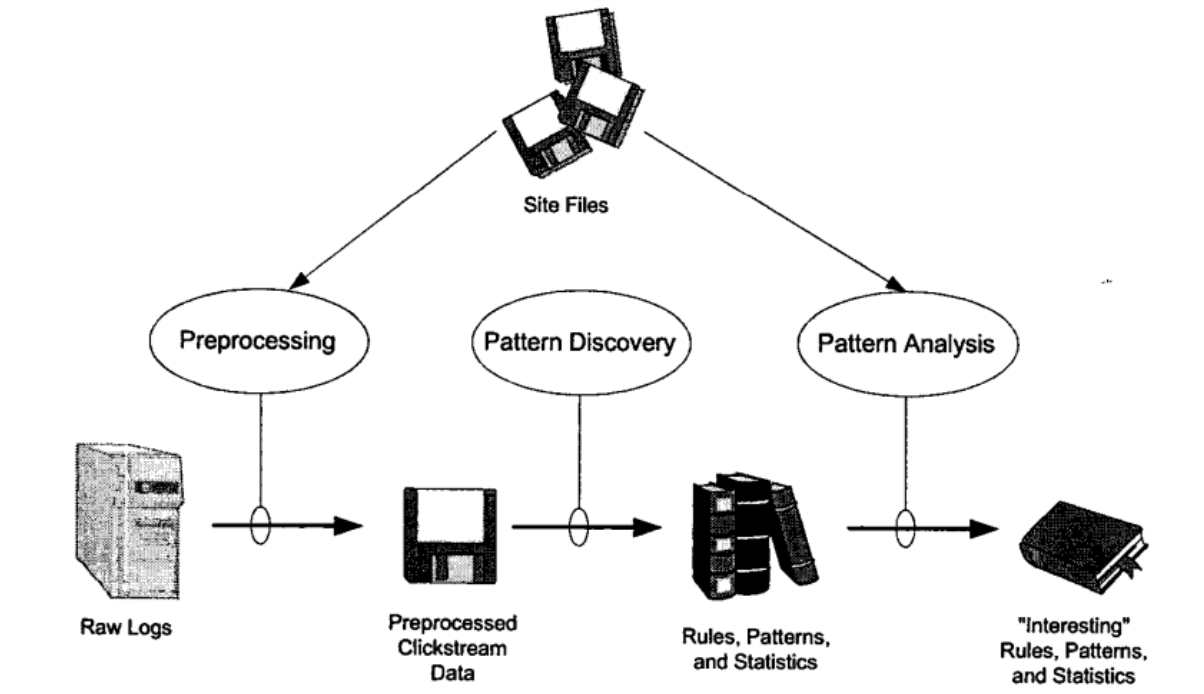


Figura 20. Proceso de minería de uso de la web (Srivastava, et al., 2000).

- Preprocesamiento: consiste en convertir el uso, contenido y estructura de la información en los datos necesarios para el descubrimiento de patrones. Se encuentra con desafíos como la relación entre el número de IPs y el número de sesiones, ya que no se trata de sinónimos, sino que se pueden producir en cualquier tipo de relaciones (una IP con una o varias sesiones, varias IPs con una o varias sesiones); la clasificación de las páginas vistas según temáticas u objetivos, puesto que muchas páginas tienen un contenido dinámico; la construcción de la estructura, ya que las páginas también pueden tener enlaces dinámicos. La fase de preprocesamiento consistiría por tanto en una combinación de los datos de las diferentes fuentes como pueden ser varias webs o herramientas, tras la cual se realiza una limpieza de los datos que no provean de información útil en la analítica web. Mobasher (2005, p. 1216) especifica que entre los datos obtenidos se encuentran los siguientes:

- Las páginas vistas son conceptualmente una colección de objetos o recursos que representan un evento del usuario. Normalmente se trata de la visualización del resultado de una combinación de plantillas estáticas con contenido generado por aplicaciones online y que resultan en la página que visualiza el usuario.

- Los usuarios se identifican por cookies en el lado del cliente, según Mobasher (2005, p. 1216), de manera que se pueda medir si realizan más de una sesión en la misma web, es decir, si la visitan más de una vez. Google, en cambio, ha estado trabajando en 2021 en una Sandbox de privacidad en la que destaca FloC (Federated Learning of Cohorts), que persigue al usuario en cada web y app que visita y usa sin usar cookies, tal y como explica Millán (2021) en su artículo para el medio Hipertextual.

- Las sesiones son las visitas a la web, una manera de segmentar las actividades de los usuarios para reconstruir, según los datos de los clics, la secuencia de acciones realizadas por el usuario durante su visita.

- La navegación se mide gracias a las referencias de acceso a cada página, resultando en un conjunto de sesiones que, a su vez, son un conjunto de páginas vistas.

- Descubrimiento de patrones: serie de métodos y algoritmos como la estadística descriptiva, la minería de datos, *machine learning* y reconocimiento de patrones, siendo las primeras las más habituales ya que analizan las sesiones a través de variables como páginas vistas, duración de la visita y profundidad de la navegación. Los patrones se descubren tanto por las sesiones como por reglas de asociación entre páginas que suelen estar en la misma sesión, agrupaciones de usuarios con patrones similares de navegación, agrupaciones de páginas con contenido relacionado, clasificación de usuarios, patrones secuenciales de navegación y modelos de dependencia (como pueden ser los pasos del carrito en un proceso de compra).

- Análisis de patrones: proceso de filtración de las normas y patrones del conjunto encontrado en el paso anterior, según los intereses del estudio que se quiera realizar.

Esos intereses son los que dividen la minería de uso de la web en tres clases, según indica Mobasher (2005, p. 1216): minería de contenido, minería de uso y minería de estructura.

El contenido es el texto y los gráficos que se sirven en una página concreta; la estructura está formada por los datos que describen el árbol de archivos que parten del HTML original hacia el resto de páginas a través de los enlaces; el uso está formado por los datos que describen el patrón de uso de las webs gracias a direcciones IP, referencias de páginas y fechas y horas de accesos; el perfil del usuario aporta además datos demográficos, como la fecha de registro y la información de su perfil.

Esta minería de uso de la web aporta los datos necesarios para la analítica web, que consiste según la Web Analytics Association (2009) en "la medición, recolección, análisis y presentación de informes para entender y optimizar el uso de una web", aunque Zumstein y Kaufmann (2009), al citar esta definición, remarcan que la analítica web no solo se utiliza para la optimización de webs, sino que puede usarse para optimizar:

- Calidad de la web: navegación, estructura, contenido, diseño, funcionalidad y usabilidad.
- Marketing digital: conciencia, imagen, campañas, publicidad de palabras clave y de banners.
- CRM online: *Customer Relationship Management* o gestión de las relaciones con clientes, y adquisición o retencion de clientes.
- Marketing individual: recomendaciones y contenidos personalizados, y personalización en masa.
- Segmentación del tráfico, usuarios y clientes online.
- Procesos internos y comunicación.
- SEO: *Search Engine Optimization* u optimización para buscadores, visibilidad y alcance.
- Tráfico: páginas vistas, sesiones y usuarios.
- Rentabilidad de negocios digitales: eficiencia y efectividad de la web.

La analítica web serviría para ejecutar un proceso de mejora continua basado en cinco pasos, según Peterson (2004, pp. 12-14): definir las actividades y acciones que se quieren mejorar y extraer los objetivos a conseguir; tomar las mediciones necesarias; priorizar las soluciones y llevarlas a cabo; verificar los cambios realizados en una muestra pequeña o con test A/B; y analizar en qué medida han resultado un éxito esos cambios.

Mulvenna et al. (2000) explican que la personalización de la web gracias a la analítica web se produce debido a la necesidad de reducir la distancia entre las verdaderas necesidades del usuario y la percepción de estas según el diseñador web. De esta manera, los usuarios pueden aprovechar la optimización producida por las similitudes con otros usuarios y el

comportamiento de estos. La analítica web permite medir la efectividad de las actividades de marketing digital, lo cual es necesario según Graham (2014) para mejorar el *engagement* con los clientes, optimizar la inversión en medios online y, gracias a ello, la rentabilidad de la empresa. El *engagement* es entendido en el sector del marketing como un factor que mide el nivel de compromiso, teniendo en cuenta la confianza, los valores, las percepciones y los mensajes que se producen a través de una interacción constante (Mafra, 2020).

Los datos se obtienen del servidor, el navegador, servidores proxy o bases de datos de organizaciones externas, y sirven para múltiples aplicaciones: la personalización de la experiencia de usuario, la mejora del sistema gracias a la detección de fraudes e intrusiones, la personalización del diseño y la estructura de la web y mejorar la inteligencia de negocio gracias al mayor conocimiento de la experiencia de usuario. Hay cuatro diferentes estrategias de obtención de los datos, descritas por Kaushik (2014):

- El propio banco de datos de la web: su principal ventaja es medir la ratio de conversión interno, ya que no se puede fiar y menos aún mezclar las ratios de conversión de fuentes externas. *Social Media*, tráfico directo, tráfico orgánico… Es recomendable hacer informes generales y luego otros dedicados a las páginas más importantes.

- Bancos de datos de analistas industriales: encuestas, datos de los resultados de acciones externas…

- Bancos de datos de competidores: SimilarWeb y Compete (este último solo para proyectos en Estados Unidos o con cierto mercado allí). Es necesario que tengan más de 100.000 usuarios únicos por mes para que los datos sean concluyentes, pero hay que ajustarse a las posibilidades de la competencia. Comparando los datos se pueden extraer conclusiones sobre en qué se está por debajo y por encima del resto de los competidores y analizar, de esta manera, sus acciones.

- Bancos de datos de los vendedores: los datos de conversión facilitados por las herramientas como MailChimp, iPerceptions… así como los reportes de benchmarking de Google Analytics. Con estos últimos se puede estudiar en qué segmentos de la audiencia se está consiguiendo una mejor conversión en cuanto a número de sesiones, de nuevos usuarios, de porcentaje de rebote… comparando los datos con la media del sector de webs. De esta manera se puede saber si se está por debajo o por encima de la media.

Con el crecimiento en el uso de programas en el lado del cliente basados en JavaScript y AJAX, la minería basada en el lado del servidor es menos capaz de capturar la experiencia de usuario, más todavía debido al incremento asimismo de la complejidad de las webs y las aplicaciones web, explican Edmonds et al (2007). En el caso de las aplicaciones de analítica web basadas en JavaScript, implementan funciones que escuchan las acciones del usuario con los elementos de la web y extraen de ellas eventos que se envían y registran en la base de datos de la herramienta. De esta manera, han surgido muchas herramientas que sirven para realizar analítica web, ya sea cuantitativa como Google Analytics, Adobe Analytics o Matomo; o cualitativa como Clicktale, Crazyegg o Heap (Mantilla Díez, 2018).

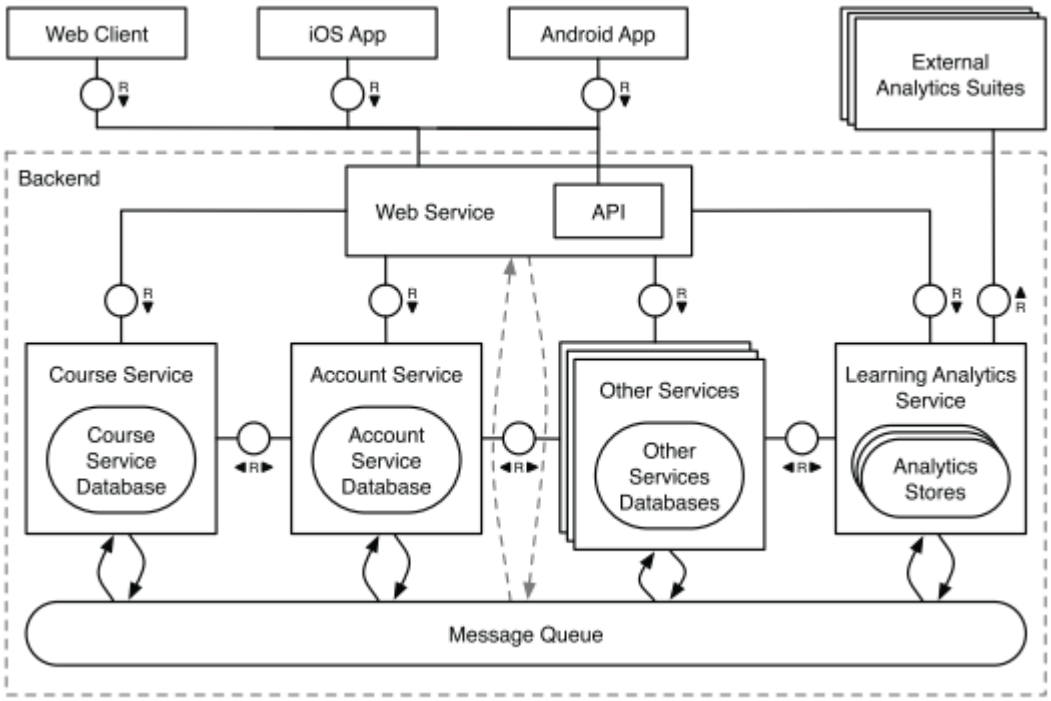


Figura 21. Arquitectura de una plataforma web de aprendizaje con almacenamientos analíticos internos y externos (Rohloff, et al., 2019)

2.2.1.3. KPI de la analítica web

Una vez se obtienen esos datos y para no caer en el error de realizar mediciones demasiado genéricas, es necesaria la utilización de métricas conocidas como indicadores clave de rendimiento (KPI) para cuantificar objetivos y que ayuden a definir y medir el progreso hacia aquello que el negocio necesita, tal y como se muestra en la Figura 22. Graham (2014) opina que tanto los objetivos como los KPI deben seguir el criterio SMART: ser específicos, medibles, alcanzables, relevantes y estar comprendidos en un periodo de tiempo concreto. Además, deben ser accionables, es decir, que se puedan realizar acciones para mejorar sus resultados. El número óptimo que estudiar a la vez es, según

esta misma autora, 5, aunque es aconsejable hasta 10. En la Figura 22 se puede observar esta estructura.

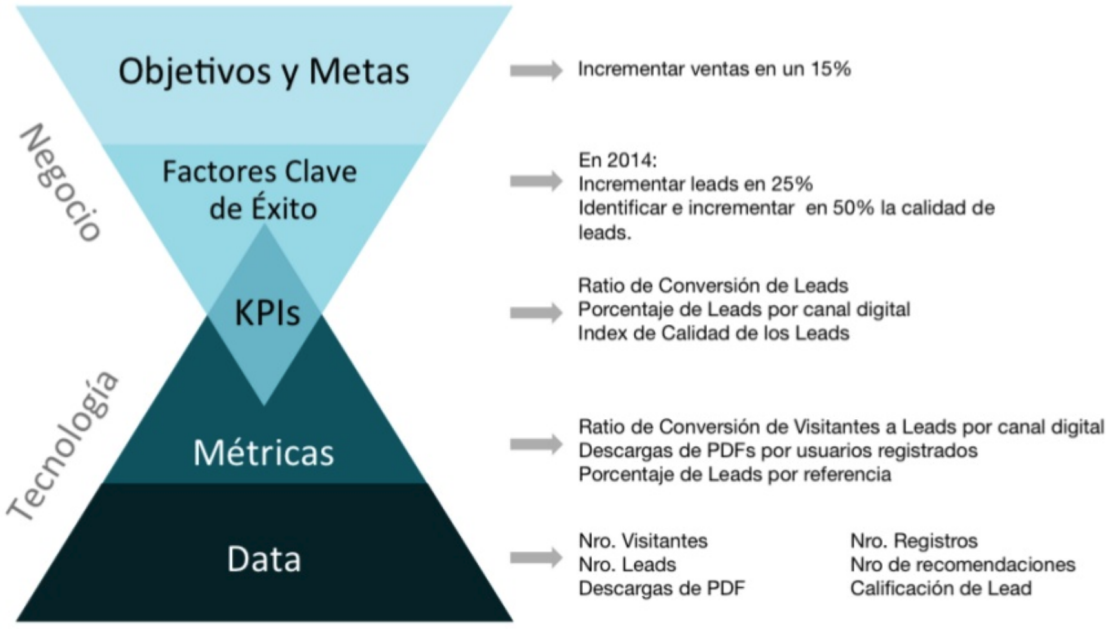


Figura 22. Relación entre los Objetivos de Negocio y los Datos Digitales (Graham, 2014)

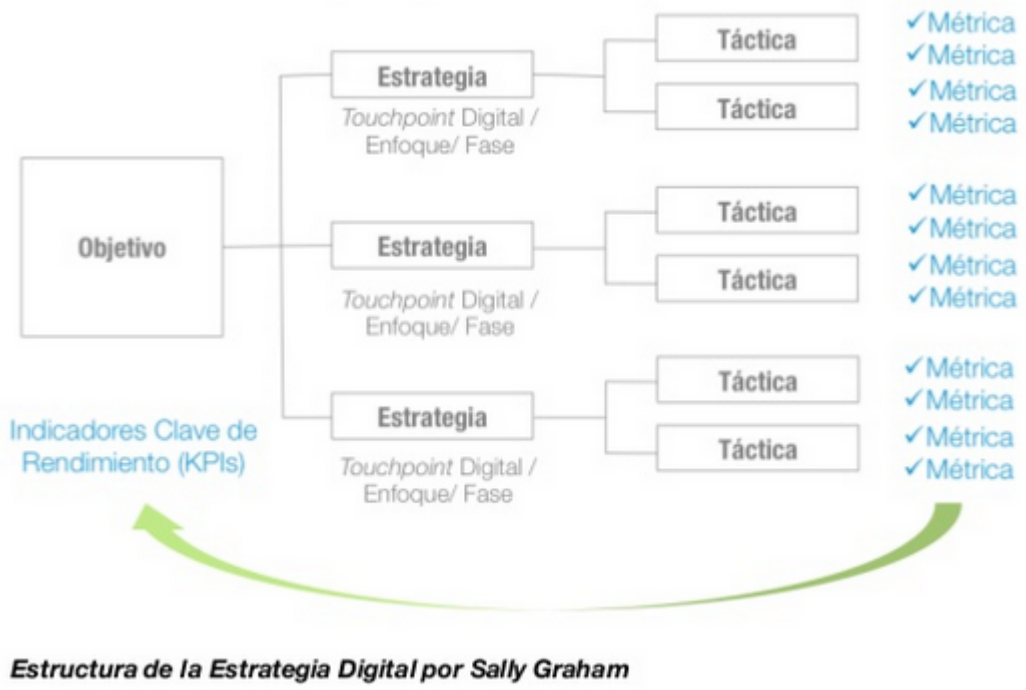


Figura 23. Estructura de la estrategia digital (Graham, 2014)

Balusamy et al. (2016, p. 253) explican en su estudio sobre la analítica web que el desarrollo web debe estar basado en objetivos y centrado en el usuario, de manera que los usuarios sean el parámetro para medir el éxito de la web y los objetivos de la web se basen en las métricas que aumenten el rendimiento de la web. Estas métricas deberían

buscar un incremento en la calidad, el rendimiento, la disponibilidad y el tiempo de vida de esta. Y la estrategia de contenidos es el factor clave para el desarrollo eficiente de la web, por lo que las métricas importantes son las relacionadas con el diseño, la interfaz de usuario, el comportamiento y la evaluación de contenidos. Todo ello en busca de unas características que designen calidad, como son la eficiencia, la funcionalidad, la usabilidad, la disponibilidad, la fiabilidad y la portabilidad.

Esa medición también puede tener en cuenta el carácter multidispositivo de la navegación online. Lamberti et al. (2017) comprobaron en su estudio que métricas como los mapas de calor y los clics se ven condicionados por el dispositivo con el que se navega, ya que por ejemplo un usuario podría tocar la pantalla para hacer clic o para hacer zoom, navegar por la página, etc. Por otro lado, el diseño web adaptable también permite controlar la visualización de los elementos según el dispositivo y el tamaño de la pantalla en que se muestre, por lo que todo ello permite realizar un análisis segmentado por estos parámetros para diferenciar el comportamiento y el rendimiento según el dispositivo utilizado. En el caso de los mapas de calor, consisten en estudiar la actividad del cursor e identificar las regiones de la página en las que se produce más interacción. Sin embargo, esta utilidad se ha visto dificultada por el diseño adaptable al tamaño del dispositivo en el que se navega. Algunos autores han propuesto estudiar esos mapas de calor considerando ya no solo los elementos en sí mismos, sino también el tiempo en que esa región de la página se está visualizando, en un enfoque desde el punto de vista del elemento y no de la ventana gráfica en sí.

Las métricas básicas en la analítica web por tanto son no solo las de los usuarios o visitantes únicos, sino también la duración de la visita y la tasa de rebote, según Alvarez Intriago et al. (2016). La tabla de estas métricas se puede ver en la Figura 24:

Figura 24. Métricas básicas y su descripción (Alvarez Intriago, et al., 2016).

La metodología a seguir para la analítica web desde un punto de vista teórico se basa en cinco fases, explicadas por Alvarez Intriago et al. (2016): la definición de objetivos y KPIs, la implementación de las herramientas adecuadas y la medición; el reporte gracias a una interfaz de usuario personalizada; el análisis basado en los objetivos iniciales; y la optimización mediante tareas o estrategias para mejorar la web.

El incremento de la complejidad de las webs ha sido paulatino, y así como evoluciona con el paso de los años, también deben evolucionar asimismo las herramientas asociadas a la analítica web. La Web 1.0 era solo informativa, la Web 2.0 se convirtió en colaborativa, la Web 3.0 en semántica y la Web 4.0 en ubicua (Montero, et al., 2010). Esta evolución se puede ver en la Figura 25:

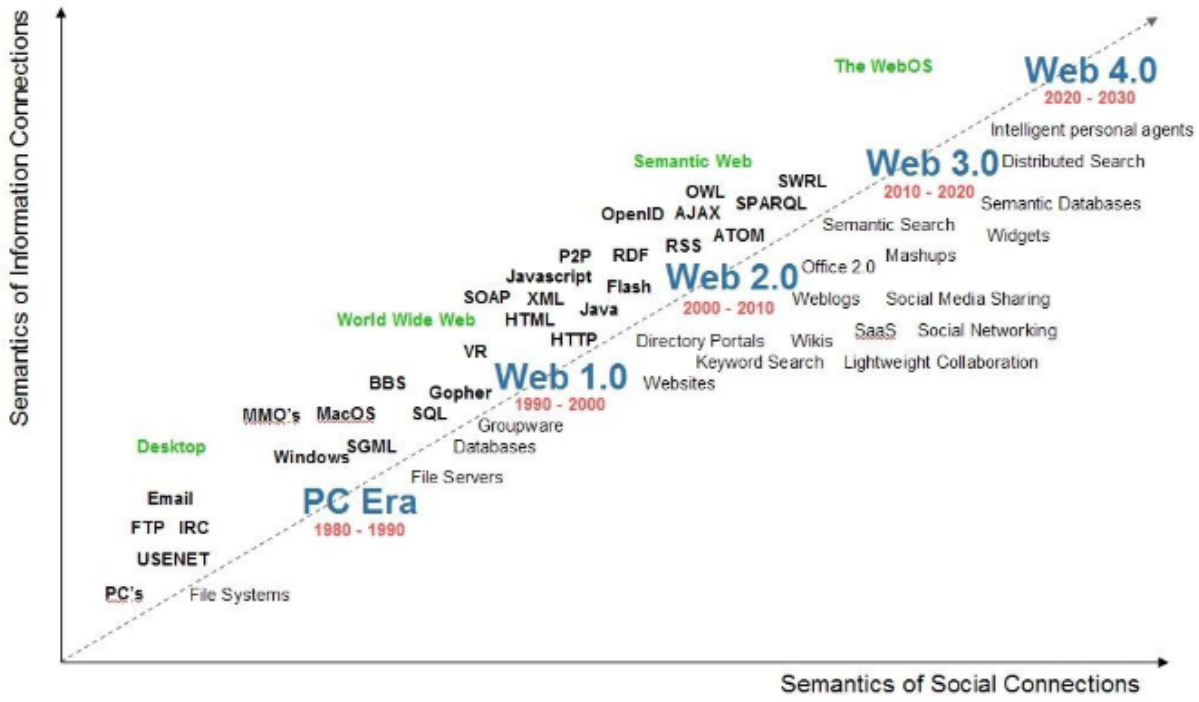


Figura 25. Evolución de la Web (Spivack, 2007)

Coleman (2016) indica que las categorías en las que se dividen los canales de adquisición de tráfico según herramientas como Google Analytics son:

- Orgánico: usuarios provenientes de buscadores, en los cuales suele predominar Google. Se trata de un tráfico con alta probabilidad de que se sienta habituado a los productos y contenidos específicos de la web.
- Directo: usuarios que conocen la URL de la web y han accedido a esta escribiéndola o a través de marcadores.
- Referido: Tráfico proveniente de enlaces de otras webs.
- Social: usuarios que provienen de redes sociales, así como plataformas de blog como Blogger[6] y WordPress[7] (aunque estos últimos deberían ser de tipo referido).
- Email: tráfico generado de boletines de email.

En el caso de las noticias, se trata de un tipo de contenido cuyo tráfico es en su mayoría orgánico o social según Coleman (2016), como se puede ver en el gráfico de la Figura 26:

[6] https://www.blogger.com/
[7] https://wordpress.com/

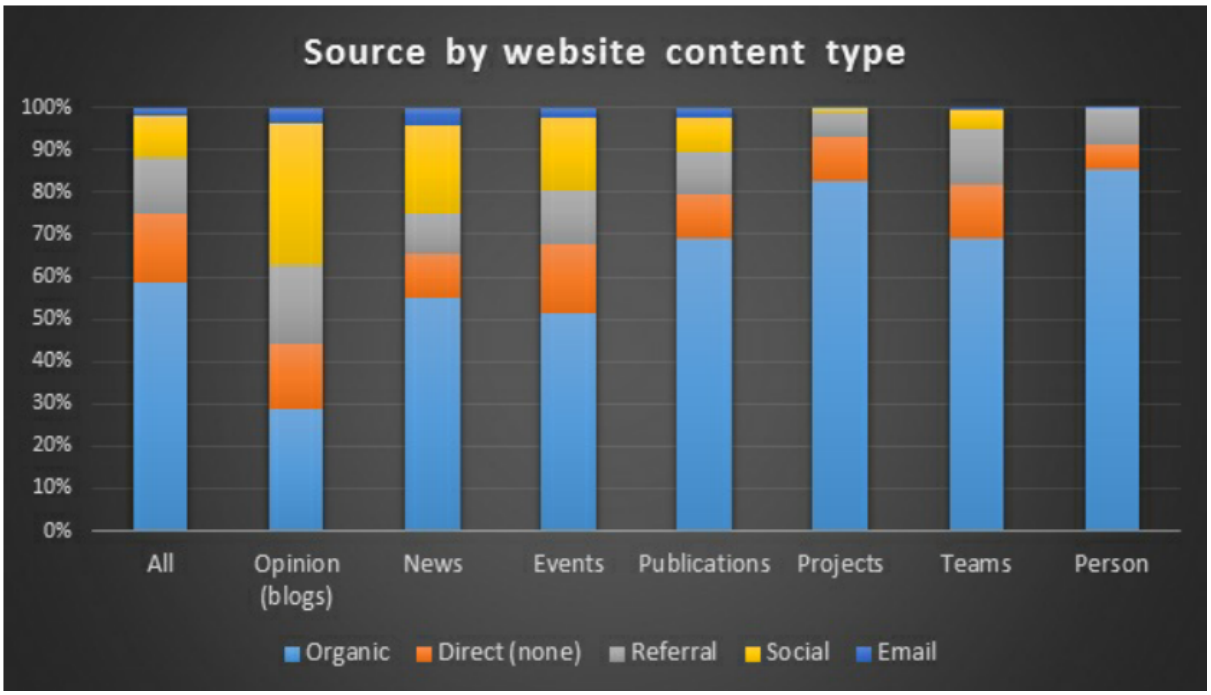


Figura 26. Canales de tráfico más activos en diferentes tipos de contenidos (Coleman, 2016)

López et al. (2017) aportan una lista de parámetros divididos por dimensiones, como puede verse en la Figura 27. Las pertenecientes a la analítica web (comportamiento, audiencia y adquisición) son las agrupaciones de indicadores de Google Analytics[8], una de las herramientas más utilizadas para ello. Las demás atienden a la calidad de la web según su SEO (posicionamiento web), contenido, usabilidad, factores técnicos y redes sociales.

Dimensión	Objetivo/definición	Parámetros analizados[a]
Analítica web		
Comportamiento	Saber cómo se comportan las personas usuarias de la web	Sesión, usuario, páginas vistas, páginas/sesión, tiempo en página, duración media de la sesión, tiempo medio de descarga de la página, porcentaje de rebote, tasa de abandono, tasa de conversión
Audiencia	Conocer comportamientos según ciertas características de las personas usuarias de una web	Idioma, ubicación, visitantes nuevos vs. recurrentes, frecuencia y visitas recientes, interacción, navegador, redes, dispositivos
Adquisición	Conocer comportamientos según los canales y redes sociales que utilizan las personas usuarias para llegar a la web	Canales, referencias, referencias sociales
Calidad de una web		
Autoridad SEO	Autoridad de una web según los factores externos que afectan a su posicionamiento	PageRank, ranking de Alexa, Moz Rank Backlinks, Autoridad de Dominio, Open Directory
SEO básico	Elementos técnicos de la propia web que afectan al posicionamiento de la web	Redirección www, título y meta description, meta keywords, robot.txt, sitemap, URL limpias
Contenido	Cantidad de contenido, frecuencia con la que se actualiza, optimización y estructura	Páginas indexadas, imágenes, enlaces on-page, etiquetas H, textos resaltados, blog
Usabilidad	Aspectos que tienen que ver con la facilidad de uso	URL y favicon, página de error 404, CSS para impresión, formulario de conversión, idioma, tiempo de descarga, optimización móvil
Aspectos técnicos	Aspectos que debe tratar un profesional de programación o informática	Protocolo seguro (HTTPS/SSL), etiquetas meta, ratio texto/código, validación W3 C, privacidad e-mail, Google analytics, optimización web, tecnologías web, localización del servidor, optimización wordpress
Redes sociales	Aspectos relacionados con las redes sociales ligadas a la web	Influencia social, página de Facebook, cuenta de Twitter

Figura 27. Tabla de parámetros de analítica web y calidad de una web (López, et al., 2017).

Hay una tendencia a priorizar métricas como la duración de la visita por encima de otras como las páginas vistas, según un estudio de la analítica en el periodismo digital de

[8] https://analytics.google.com/

Blanchett Neheli (2018), pero debido a que los anunciantes están más interesados en los clics que en la duración de la sesión, las páginas vistas siguen siendo una métrica utilizada por los medios de comunicación, ya que puede ayudar a crear y aumentar la audiencia mediante la optimización de los artículos y fomentar la elaboración de historias originales.

El valor de los KPI tiene, por otro lado, una relación directa con el valor del análisis que se hará con ellos. KPI como clics, páginas vista, número de visitas, vistas de vídeo, emails enviados o número de reportes de prensa no presentan la calidad suficiente, según Kaushik (2011), porque: son métricas tácticas de las que no se puede aprender nada concluyente ni impactante para el negocio; simplemente reportan datos pero no ayudan a razonar sobre ellos; requieren demasiada deducción, porque se infiere que más es mejor cuando no son indicadores de éxito directamente; y son datos de eventos inmediatos (una visita, una sesión, una página de salida…) que carecen de contexto de por sí.

Por otro lado, KPI como la lealtad, novedad, ratio de conversión, frecuencia de visita, amplificación, compartición de contenido, valor económico y ratio de finalización de tareas sí que son métricas interesantes para su análisis según Kaushik (2011) ya que:

- Fuerzan el análisis estratégico de las métricas que contienen datos del balance del negocio.
- Están inculcados de la voz directa del usuario, por lo que no es necesario que se infieran e interpreten los datos de forma sesgada.
- Analizan el comportamiento a través de sesiones y animan, de este modo, a centrar el negocio en valor a largo plazo.
- Debido a su naturaleza inherente en la mayoría de los casos muestran de manera muy clara si la optimización es buena o mala.

Por ello, Kaushik (2011) argumenta que resulta más conveniente medir la fidelidad y el tiempo desde la última visita de los usuarios que el número de las visitas; es preferible medir la conversación que se produce en las redes sociales (comentarios, respuestas recibidas y enviadas...) que los “me gusta” y los seguidores; es mejor analizar por qué los usuarios visitan la web y si consiguieron lo que se proponían (encuestas) que inferirlo del tiempo en página o el porcentaje de rebote; es más adecuado identificar el valor económico de la web según los esfuerzos y trabajos invertidos, el valor de las campañas, identificando dónde se está invirtiendo de menos y dónde de más… El valor económico en el marketing de contenidos se puede extraer también de los enlaces salientes. Es recomendable si es posible añadir etiquetas de seguimiento en los enlaces con parámetros de campaña,

códigos de afiliación, etc., para recoger los datos de las ventas que pueden haberse conseguido gracias a un enlace desde la web.

La selección de los KPI dependerá de múltiples factores, entre ellos las características y necesidades del negocio o medio de comunicación. Según el tamaño del negocio, Kaushik (2011) propone los siguientes KPI, cuyo resumen se puede ver en la Figura 28:

a) Webs de pequeños negocios

- Adquisición
    - Coste por adquisición (CPA): cifra más relevante según el gasto que suponga obtener una visita que el nº de clics, de visitas, de enlaces entrantes e incluso impresiones.
- Comportamiento
    - Porcentaje de rebote: ayuda a identificar público objetivo inadecuado, páginas de aterrizaje irrelevantes…
    - Porcentaje de abandono en el embudo de conversión: es más fácil convencer a los que dieron varios pasos en el embudo que a los que no dieron ninguno. En este caso podría ser importante hacer tests multivariantes y A/B, tipos de prueba que son explicados en el apartado 2.2.1.5.
- Resultados
    - Ratio de conversión en general: identificar las fuentes que generan más ingresos y potenciarlas. Tratar de identificar los usuarios que consiguen el doble de conversión: averiguar qué les interesa, cómo llegaron, etc…

b) Webs de medianos negocios

Los KPI de los pequeños negocios y, además, los siguientes:

- Adquisición
    - Proporción de clics (CTR): más específico que el CPA, investiga qué enlaces y acciones consiguen que los usuarios cliqueen en ellos.
- Comportamiento

    - Profundidad de página: a qué paso llegan del embudo de conversión o cuántas páginas visita el usuario.
    - Lealtad (% de visitas y nº de veces que vuelve): optimizar la web para que los usuarios la visiten más a menudo.

- Resultados
    - Ratio de conversión según objetivos: identificar los objetivos y el valor que aportan al proyecto.
    - Valor de objetivo por visita: identificar los objetivos que aportan más valor gracias al valor económico de cada visita. Determinar qué visitas son las más rentables.

c) Webs de grandes negocios

Todos los KPI de los medianos negocios y, además, los siguientes:

- Adquisición
    - Porcentaje de nuevas visitas: identificar las visitas de usuarios que nunca habían estado antes en la web y fidelizarlos. Se une al *re-targeting* y el análisis de comportamiento para mejorar la conversión de las visitas habituales.
- Comportamiento
    - Eventos / visita: medir los eventos que realiza cada visita para estudiar si han sido rentables según la estrategia a seguir y los objetivos de esta.
- Resultados
    - Días hasta la conversión (*e-commerce*) / *Time Lag* (webs de contenido): días hasta que se cumple la conversión o los objetivos.
    - Porcentaje de conversiones asistidas: si un canal externo está en el embudo de conversión de una visita… es que ha asistido a su conversión. Sirve para investigar qué canales ayudan a conseguir más conversiones.

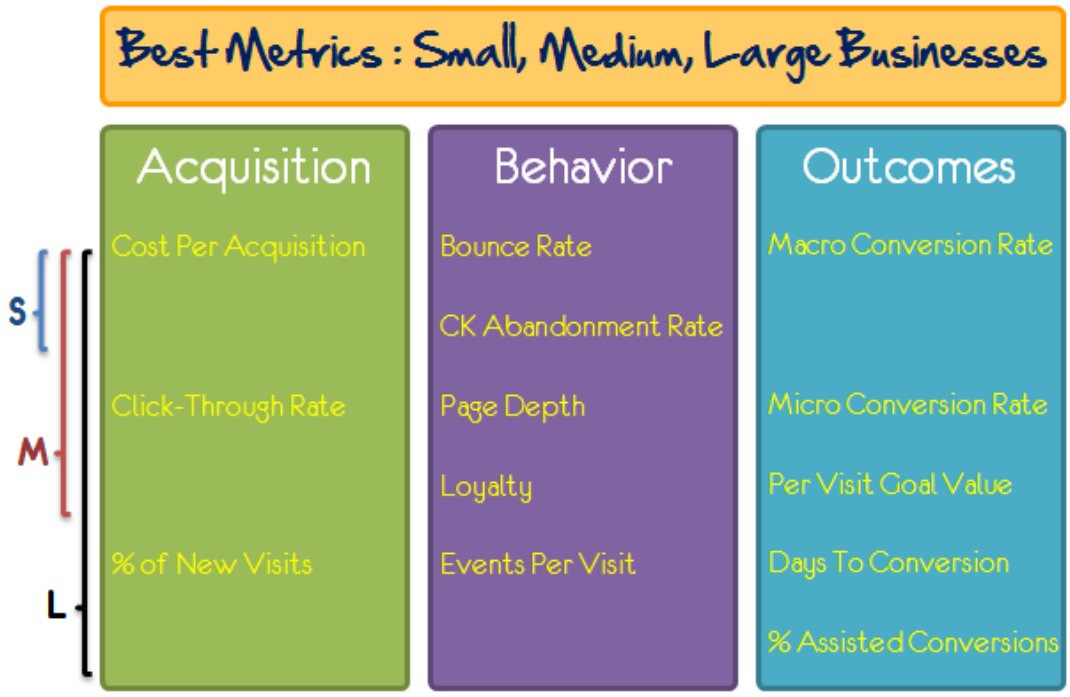


Figura 28. Lista de KPI según el tamaño del negocio al que pertenece la web (Kaushik, 2011)

Según Noguera (2009), la fidelidad de la audiencia se mide con páginas vistas y tiempo de navegación en el sitio. Otros medios como e-intelligent.es (2014) sugirieron las siguientes métricas: para medir el consumo de contenidos serían de utilidad el número de páginas vistas, el tiempo medio en la página, la tasa de rebote, el número de páginas vistas por cada visita y el número de usuarios únicos; para medir el alcance y la difusión se acudiría a las comparticiones en redes sociales, el número de enlaces recibidos y el número de conversaciones que genera el medio o el artículo; en cuanto a las conversiones se mediría el número de conversiones generadas gracias a la estrategia de contenidos, las suscripciones a boletines, las descargas de archivos importantes y las ventas generadas, en el caso de que la web dispuesta de ellas.

Saura et al. (2017) propusieron un listado de KPI segmentado en cuantitativos, cualitativos y de comportamiento:

- Cuantitativos: impresiones; número de visitas; número de usuarios únicos; y conversiones a modo de objetivos, como registros, descargas, formularios enviados y compras.
- Cualitativos: pruebas A/B, en forma de diferentes versiones de páginas y elementos; llamadas a la acción; experiencia de usuario; sistemas de valoración; encuestas y formularios; y el flujo de navegación de los usuarios.

- Comportamiento: ratio de conversión, ratio de cumplimiento de objetivos, tipos de usuarios, tipos de fuentes de tráfico, palabras clave y rango de estas.

Sin embargo, métricas básicas como usuarios o páginas vistas pueden no proveer del significado suficiente a la hora de medir el éxito de una web, en parte debido a que el usuario participa de una manera más activa en la experiencia de la web. Por ejemplo, para analizar el comportamiento de los usuarios, Phippen et al. (2004) aportan dos métricas: el factor de adherencia, basado en el tiempo de navegación en todas las páginas dividido por el número total de visitantes únicos; y el factor de relevancia, basado en el número medio de publicaciones consumida por el usuario dividido por el número disponible o esperado de publicaciones. De esta manera, la analítica web no solo se limita a la medición del tráfico sino también del comportamiento de este.

Google Analytics, la herramienta de analítica web de Google, unificó por su parte bajo el grupo “lealtad de la visita” las cuatro variables siguientes a modo de distribución de las métricas: lealtad, tiempo desde la última visita, duración de la visita y profundidad de la visita. Kaushik (2007) las describe de la manera siguiente:

a) Lealtad de la visita

Con qué frecuencia los usuarios visitan la web en un periodo de tiempo concreto. Sin embargo, es importante darle el contexto necesario según el número de publicaciones que se realicen para intentar determinar si ello afecta a su retorno. Por ello, la acción sería:

1. Identificar una meta para el número de visitas que se esperaría del tráfico en un periodo de tiempo.
2. Medir la realidad con el reporte de lealtad de Google Analytics.
3. Comparar la optimización a través del tiempo para comprobar cuál es el progreso.

b) Tiempo desde la última visita

Cuánto tiempo ha transcurrido desde que el usuario visitó por última vez la web.

La acción consistiría en definir qué tipo de web es y determinar si la mayor parte de la audiencia es nueva (0 días antes) de manera consistente y cuál podría ser la causa (contenido inadecuado, necesidad de vender mejor el valor de repetir la visita…). También sería necesario comprobar si el número de visitas repetitivas aumenta o disminuye y en qué franja de tiempo.

c) Duración de la visita

Calidad de la visita representada mediante la duración de la sesión en segundos. La acción sería:

1. Identificar la distribución de la duración de las visitas.
2. Planificar métodos creativos para obtener una mayor duración de la visita (especialmente más de 60 segundos).
3. Planificar un cobro variable de los anuncios (si los hay) tras 60 segundos.
4. Analizar si un tiempo demasiado largo es beneficioso o no (si es una web de soporte, un tiempo muy largo podría indicar que no se encuentra fácilmente la respuesta).

d) Profundidad de la visita

Cuál es la distribución del número de páginas en cada visita en un periodo de tiempo concreto. La acción sería muy similar a la de la duración de la visita, pero aplicada a la profundidad.

Es muy importante realizar también una segmentación de los datos: qué palabras clave atraen segmentos de tráfico más valiosos, qué tipos de contenido consumen las visitas que más tiempo permanecen en la web, etc. Hart, según cita Kaushik (2007), realizó un estudio para medir el valor de la web de su compañía (Trust) según la cantidad de contenido consumido y por qué fracción de las visitas de una web, como se ve en la Figura 29:

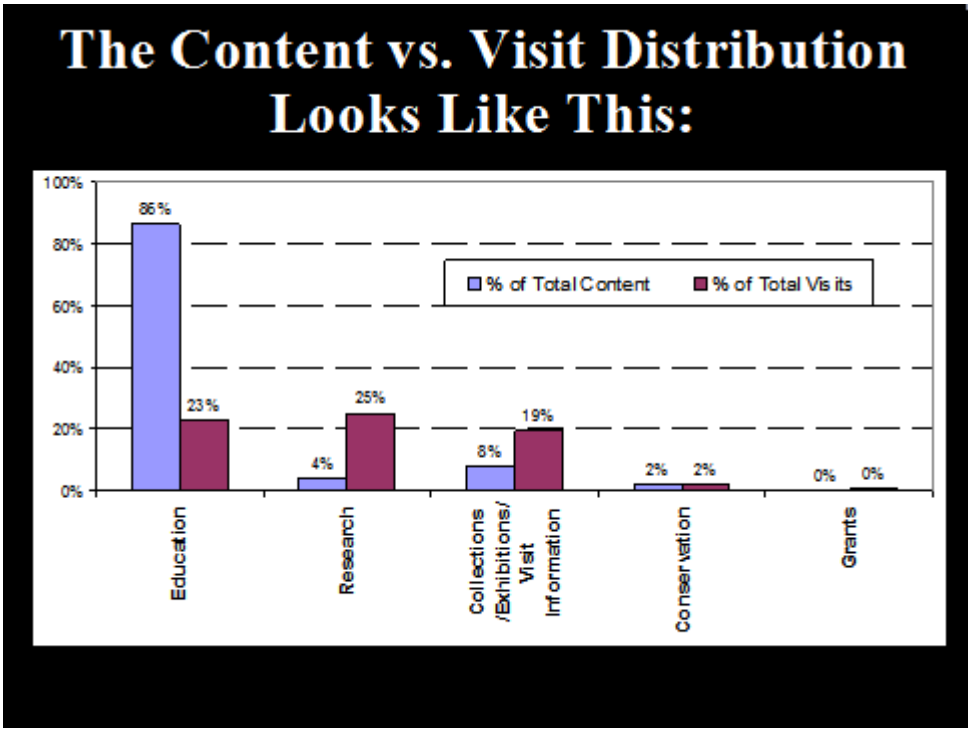


Figura 29. Estudio del contenido consumido en Trust según Hart (Kaushik, 2007)

En la Figura 29 se podría haber acudido a multitud de KPIs para intentar analizar el consumo de contenido, pero con dos variables muy utilizadas Kaushik (2007) pudo llegar a conclusiones muy interesantes: que el 86% del contenido es consumido por solamente el 23% de las visitas, que el 25% de las visitas consumía el 4% de los contenidos y que el 19% consumía otro 8% de los contenidos. De todo ello se podría extrapolar que, en el caso de la web de Trust, si se creara más contenido de tipo Investigación habría una mayor posibilidad de mejorar el número de visitas que aumentando el contenido de Educación.

Para medir el marketing de contenidos, Graham (2014) propone los siguientes KPI: porcentaje de tráfico orgánico (tráfico total / tráfico orgánico), número de visitantes únicos, número de visitantes recurrentes, porcentaje de usuarios que navegaron con un móvil, procedencia de los usuarios, porcentaje de rebote, duración de la visita, comentarios publicados, contenidos compartidos en redes sociales y visualizaciones completas de vídeos.

Por otro lado, Hochuli (2015) aporta las métricas relacionadas con el contenido que se ven en la Figura 30, siendo divididas según si la fuente es social o de búsqueda y si el objetivo es el tráfico, el *engagement* o la conversión:

	Goal: **Traffic**	Goal: **Engagement**	Goal: **Conversion**
Social content metrics	**Primary**: • Traffic via social channels	**Primary**: • Shares/retweets **Secondary**: • Comments	**Primary**: • Link click • Video play • Email sign-up/conversion • Transaction
Search content metrics	**Primary**: • Organic traffic **Secondary**: • SERP ranking • Geolocation traffic for local search algorithms	**Primary**: • Bounce rate • Time on site • Avg. pages per session **Secondary**: • Scroll tracking to separate readers vs. skimmers	**Primary**: • Link click (in body) • Download • Email sign-up • Visited a second blog post • Visited a product page • Transaction

Figura 30. Métricas del contenido social y de búsqueda (Hochuli, 2015).

Se clasifican según Hochuli (2015), por tanto, de la siguiente manera:

- Métricas sociales: el tráfico sería procedente de canales sociales; el *engagement* se mediría primero con retuits y después con comentarios; y la conversión con los clics en los enlaces, las reproducciones de vídeo, registros/conversiones y transacciones.

- Métricas de búsqueda: el tráfico sería el orgánico de los buscadores, junto con el ranking de las posiciones en éstos y la geolocalización del tráfico en el caso de los algoritmos de búsqueda locales; el *engagement* se mediría con el ratio de rebote, duración de la visita, número medio de páginas por sesión y seguimiento de la visualización del contenido; y la conversión con los clics en los enlaces del cuerpo, las descargas, los registros, visitas a otros artículos del blog o a un producto y transacciones.

Raso (2016), por su parte, explica que hay determinadas métricas que se confunden con el *engagement,* pero no son tales:

- Alcance vs *engagement*: el alcance es fácilmente manipulable debido a publicidad engañosa o confusa, ya que se puede conseguir más público publicitando algo que no se ofrece en la web, por ejemplo. Propone: conversiones, ya que lo que se necesita es que los usuarios hagan determinadas acciones cuando visitan la web.

Es mejor medir su comportamiento en lugar del tamaño de l audiencia a la que llega el contenido.

- Tiempo en página vs *engagement*: el tiempo en página no es fiable, ya que solo se mide en visitas que no tienen rebote, y a veces se puede dejar la página inactiva en una pestaña y no efectuarse una visualización activa... pero seguir conectados. Propone: "*scroll depth*", la distancia que se recorre con la barra de desplazamiento en la visita de un usuario, para medir si los usuarios están visualizando el final de los artículos.

- Compartir vs *engagement*: Hay poca correlación entre los artículos que leemos y los que compartimos (Haile, 2014). Por tanto, el número de veces que se comparte no es una variable fiable para medir el éxito del contenido. Propone: comentarios, ya que los usuarios no comentan a no ser que estén realmente interesados.

Otro análisis que podría dar conclusiones muy interesantes podría ser el de las búsquedas internas. Para estudiarlas, Kaushik (2007) propone los siguientes pasos:

1. Entender el uso de la búsqueda en la web: cuánta gente la usa y qué palabras clave son las más buscadas.

2. Averiguar dónde hacen los usuarios dichas búsquedas: analizar la página en la que se encuentran y lo que buscan puede sugerir cambios o tests multivariante para mejorar la optimización de la página. Además, si se añade en qué página acabaron tras la búsqueda, se puede entender más profundamente el comportamiento del visitante.

3. Analizar la calidad de la búsqueda interna de la web: es necesario medir el porcentaje de salidas del buscador, es decir, qué porcentaje de visitas buscaron y luego abandonaron la página de resultados sin elegir ningún resultado. Otro dato interesante es el número de páginas de resultados de búsqueda visitados por cada palabra clave, comparado con la media de la web. Por último, es útil estudiar el refinamiento de la búsqueda, es decir, cómo especifican la búsqueda los usuarios y cómo se puede evitar que necesiten hacerlo.

4. Segmentar a las visitas que utilizan el buscador: comparar las palabras clave que usaron las visitas en los buscadores para acceder a la web y qué palabras clave usaron, a su vez, en el buscador interno esas mismas visitas. O comparar el uso de la búsqueda interna, las salidas y el refinamiento con el tiempo de la visita después de haber realizado la búsqueda.

5. Medir la conversión de las visitas que usaron la búsqueda interna y compararlas con las que no lo usaron: ganancias, tamaño de pedido medio, ratio de conversión y valor por visita.

2.2.1.4. Metodologías de la analítica web

Es necesario utilizar una metodología formal para representar de la mejor manera posible la información de lo que se necesita analizar. Sin embargo, hay muchas maneras de abordar este desafío, llegando a ser necesario construir nuestra propia metodología para llegar a las conclusiones que deseemos. Graham (2014) sugiere como metodología a nivel general, fuera del ámbito digital y a partir de la cual construir una propia, la llamada RADAR:

- Reportar: Extraer datos de las fuentes seleccionadas.
- Analizar: Revisar los reportes e identificar patrones o tendencias basadas en una hipótesis.
- Decidir: Determinar si los datos confirman o no la hipótesis.
- Actuar: Crear e implementar un plan para afinar la hipótesis inicial.
- Repetir: Mejorar continuamente el proceso repitiendo la investigación de la hipótesis.

Un procedimiento parecido es el que propone Chaffey y Patron (2012) según las recomendaciones de Econsultancy-RedEye, haciendo especial énfasis en la importancia de dejar claros los objetivos de negocio y los KPI desde el inicio, para así crear un plan que priorice los diferentes tests que puedan aplicarse al proyecto.

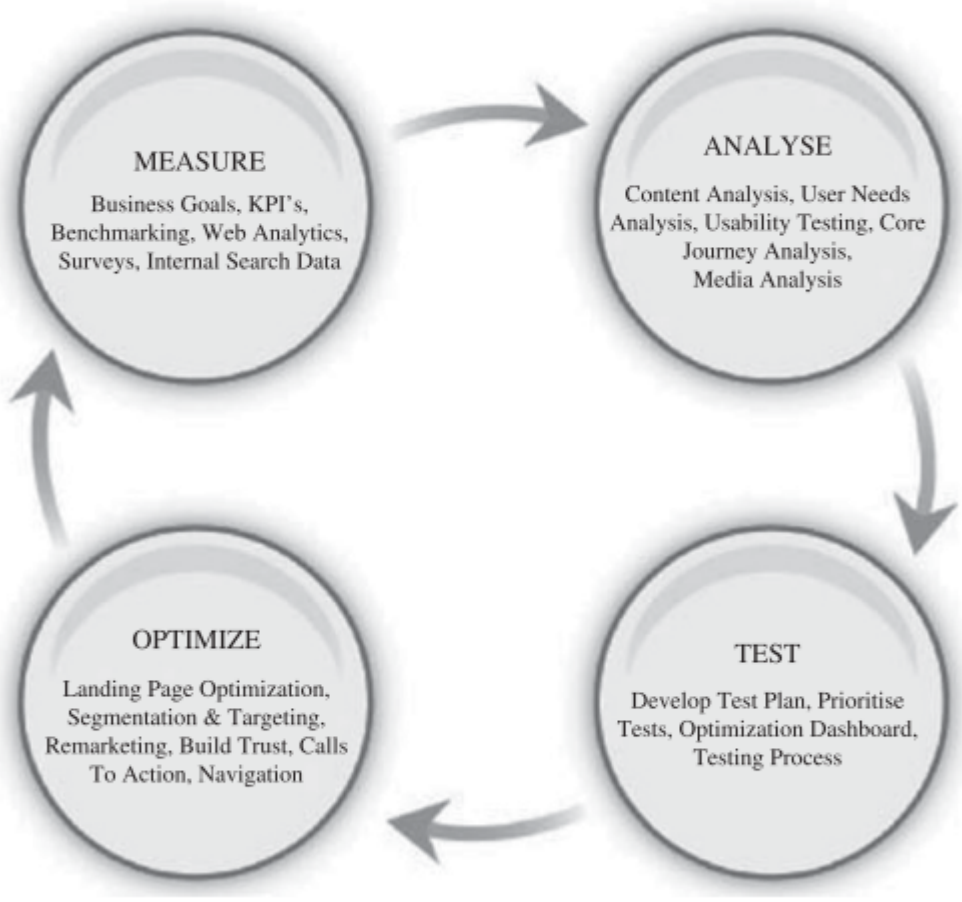


Figura 31. Proceso recomendado para la mejora continua gracias a la analítica digital (Chaffey & Patron, 2012)

Para crear un modelo de medición y marketing digital, es necesario una estructura y un conjunto de medidas objetivas con las cuales identificar el éxito o el fracaso. Las tres áreas clave a analizar en el marketing para Kaushik (2011) son: adquisición, en la cual se trata de anticipar el tráfico y medir y priorizar los esfuerzos dedicados a cada canal; comportamiento, en el cual se mide qué contenidos son de consumo prioritario y qué acciones deberían realizar los usuarios; y conversiones, en el sentido de acciones que puede realizar el usuario y que aportarían valor al negocio online. Para ello, Kaushik (2011) ha desarrollado un proceso de cinco pasos:

1. Identificar los objetivos del negocio por adelantado y elegir los parámetros más amplios para el trabajo que se está haciendo. Los objetivos deben ser: realizables, comprensibles, manejables y beneficiosos. Una tabla que podría servir como ejemplo de dicha clasificación sería la que se observa en la Figura 32:

Figura 32. Tabla de objetivos de negocio (Kaushik, 2011).

En la tabla de ejemplo se puede observar que los tres objetivos de negocio son: ganar visibilidad, generar conversiones y destacar eventos.

2. Identificar las metas definidas para cada objetivo del negocio. Las metas son estrategias específicas que se usarán para cumplir con los objetivos del negocio. Un ejemplo de dicha tabla sería el de la Figura 33:

Figura 33. Tabla de metas de la web (Kaushik, 2011).

En la tabla de ejemplo se han propuesto las siguientes metas de la web: para ganar visibilidad, reforzar la publicidad *offline* y *online*; para generar conversiones, conseguir conversiones vía email y contacto y proveer información y recursos de compra en la página principal; para destacar eventos, interactuar con la comunidad a través de los eventos locales.

3. Identificar los indicadores clave de rendimiento (KPI), es decir, las métricas que ayudan a entender en qué medida se están consiguiendo los objetivos. Un ejemplo de ello sería la tabla de la Figura 34:

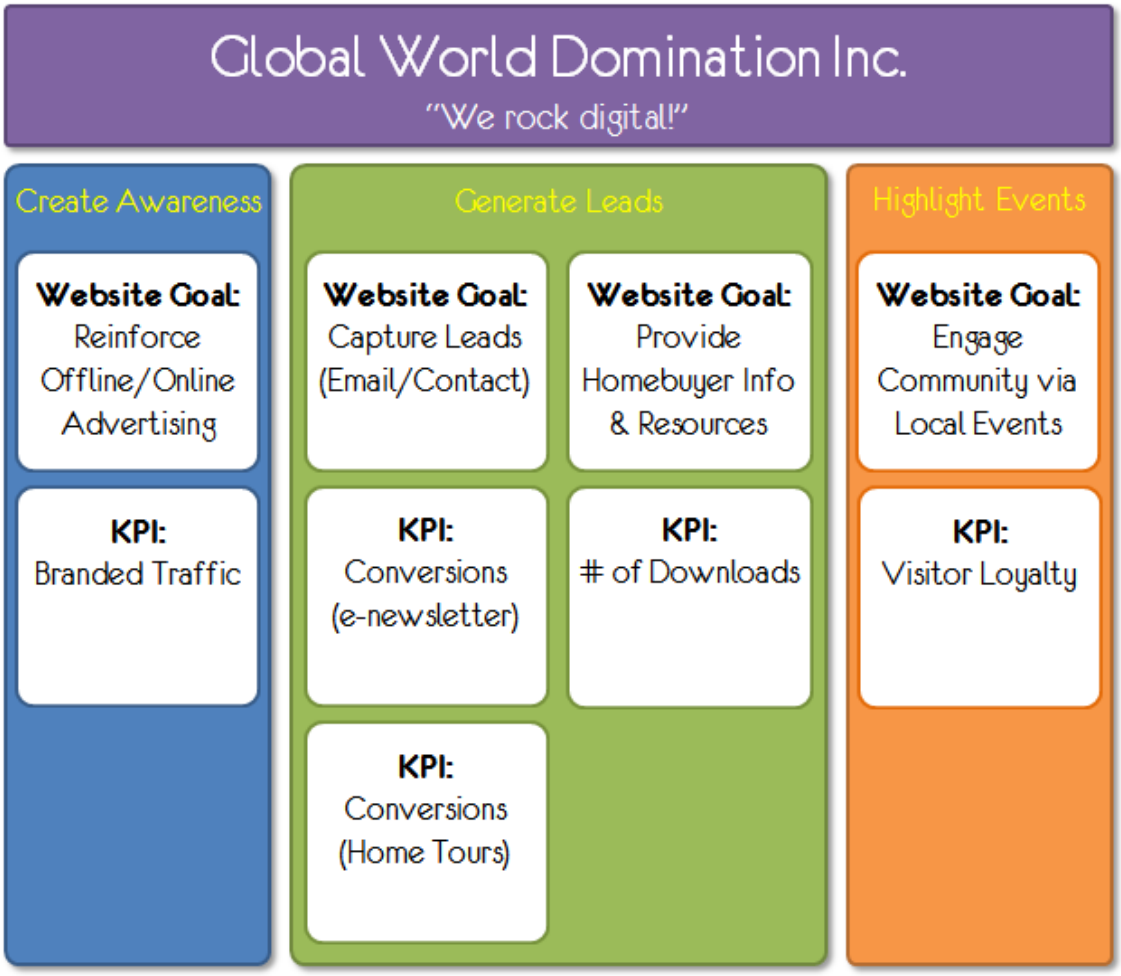


Figura 34. Tabla de KPI (Kaushik, 2011).

En la tabla de ejemplo se aportan los siguientes KPI: Tráfico de marca, conversiones a través de la e-newsletter, conversiones a través de la página principal, número de descargas y lealtad de la visita.

4. Elegir los parámetros de éxito por adelantado para identificar los objetivos de cada KPI. Es importante saber en qué medida se han conseguido las metas de cada KPI. Tanto es así que, si no se sabe qué cifra poner, es mejor poner una que sea muy alcanzable y ya se irá aumentando con el tiempo y la experiencia. Un ejemplo de dicha tabla sería el siguiente:

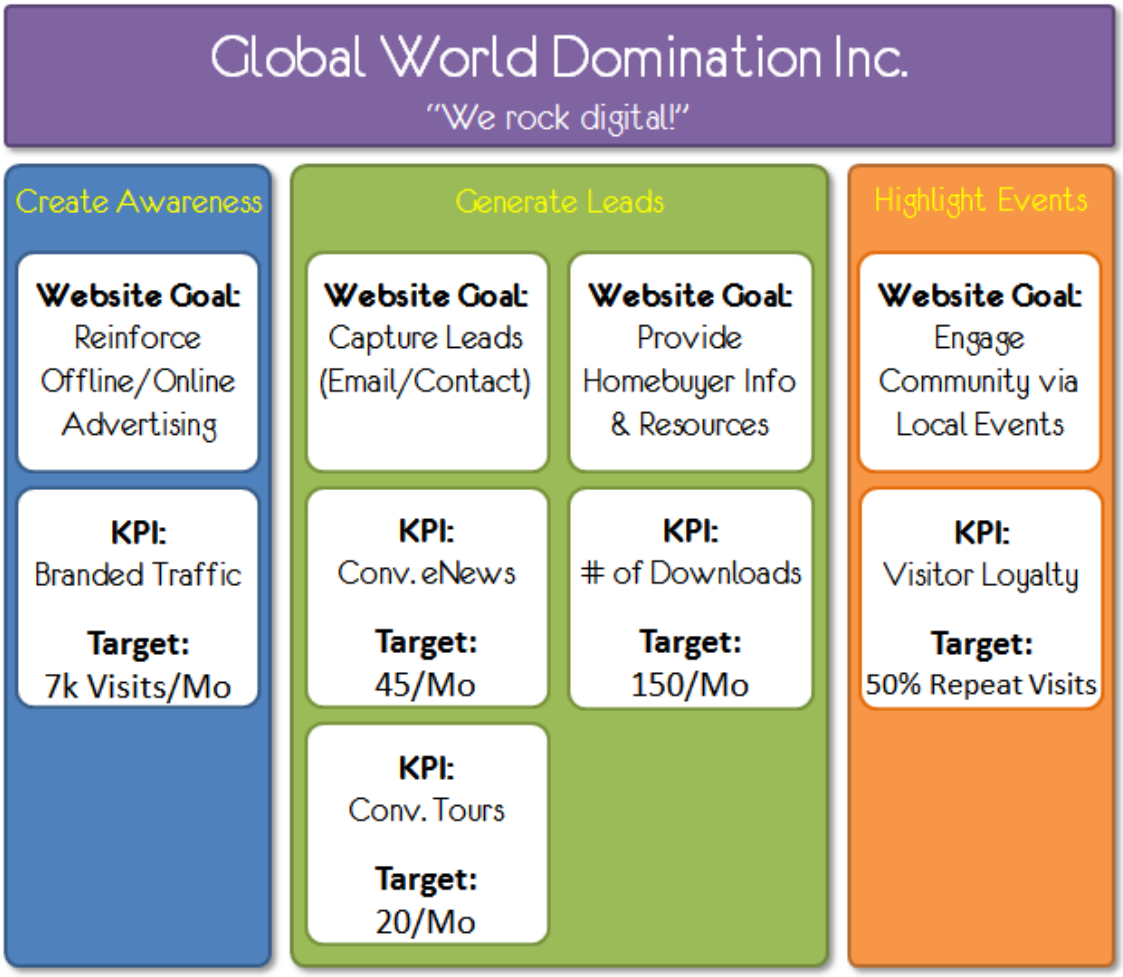


Figura 35. Tabla de metas de KPI (Kaushik, 2011).

En la tabla de ejemplo se ha elegido metas de KPI: 7000 visitas al mes, 45 conversiones al mes a través de la e-newsletter, 20 conversiones al mes a través de la página principal, 150 descargas al mes y 50% de visitas que vuelven de nuevo a la web.

5. Identificar los segmentos de usuarios, fuentes de tráfico, comportamiento y resultados que se analizarán para entender por qué se ha tenido éxito o se ha fracasado. Un ejemplo de ello sería el de la tabla de la Figura 36:

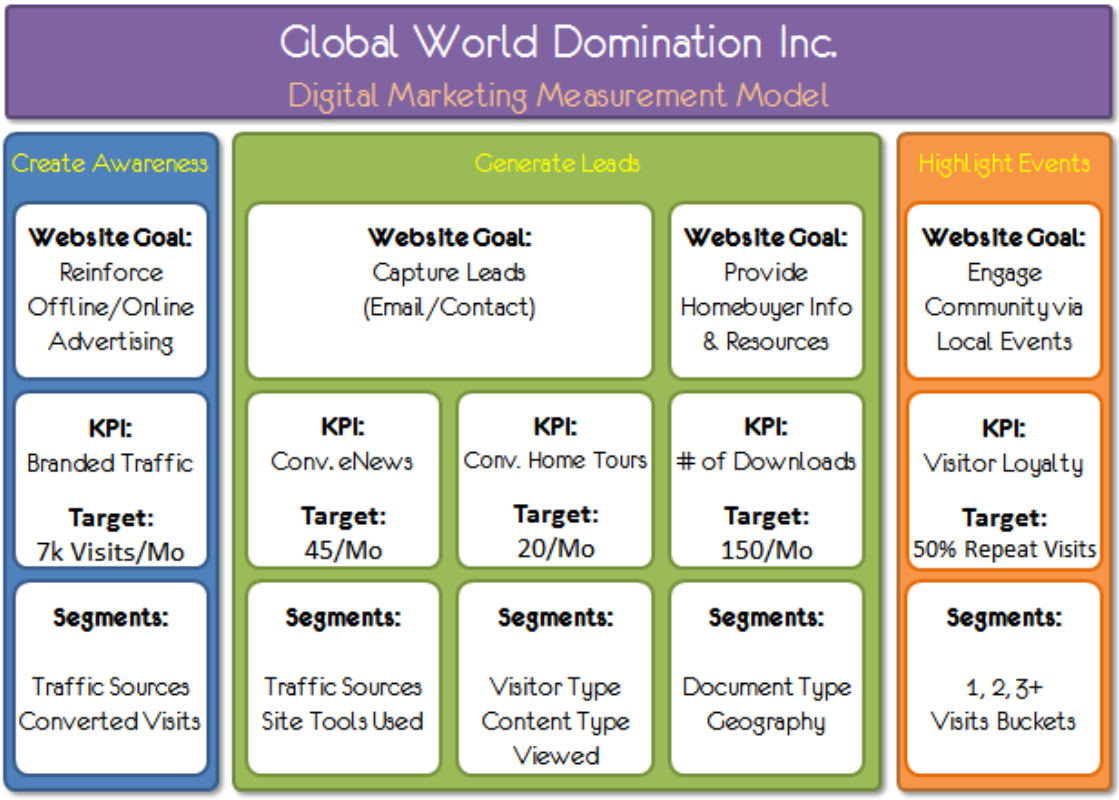


Figura 36. Tabla de segmentos a analizar (Kaushik, 2011).

> En la tabla de ejemplo se llega a la conclusión de que los segmentos a analizar son: fuentes de tráfico, visitas que convierten, herramientas de la web utilizadas, tipo de visitantes, tipo de contenido consumido, tipo de documento descargado, geografía de los usuarios que descargan contenidos y segmentación de usuarios según si han visitado una, dos o tres o más veces la web.

Segmentar los usuarios, último paso de la lista anterior, es un factor clave en el éxito de un proyecto para Kaushik (2010). La mayoría de las herramientas de estadística web incorporan segmentos por defecto que ayudan a interpretar mejor los datos: visitantes nuevos, tráfico proveniente de búsqueda de pago, tráfico directo… Sin embargo, los segmentos por defecto son para uso generalista y, como cada caso es único, resulta necesario crear segmentos personalizados según las prioridades, la estrategia a través de todos los canales y las herramientas utilizadas.

No hay que reportar nunca una métrica sin segmentar para realizar un análisis pormenorizado de dicha métrica, asevera Kaushik (2006). Las ventajas de segmentarla son que es imposible segmentar una métrica sin un cierto esfuerzo por comprender qué es lo que se está reportando y el valor que aporta al proyecto; que ayuda a pulir áreas más profundas de las que emergerán conocimientos claves para llevar a cabo acciones realistas

y significativas; y que facilita la comunicación de los resultados de los datos a personas que no son analistas. Por ejemplo, segmentar las visitas que al menos han estado 5 segundos visitando la web frente al número total, o que proceden de buscadores, o específicamente de Google o incluso todos los segmentos anteriores, como se ha realizado en la Figura 37:

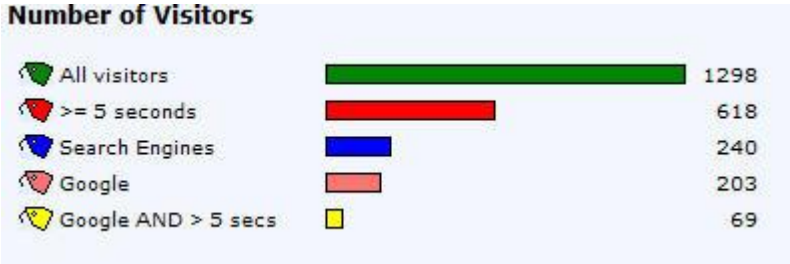


Figura 37. Ejemplo de segmentación de visitantes por la duración de la visita y el buscador utilizado (Kaushik, 2006)

Para segmentar, Kaushik (2010) especifica que es necesario definir los objetivos del negocio, las metas y las cifras que se quieren alcanzar. Una vez hecho esto, se crearán segmentos clasificados en tres categorías:

a) Adquisición

La adquisición es la actividad que se realiza para atraer gente a la web: campañas de PPC, de email, de afiliados, de anuncios en banners, de marketing en redes sociales, de SEO… Los segmentos más importantes de esta categoría serán aquellos en los que se está invirtiendo más dinero y/o tiempo. Así, su análisis servirá para optimizar en una mayor medida los esfuerzos del proyecto. El proceso sería el siguiente:

1. Identificar qué es lo más importante / prioritario.
2. Crear un segmento y los micro segmentos para ello.
3. Aplicarlo en los reportes relevantes para medir la optimización usando los KPIs.
4. Realizar las acciones para mejorar dichos datos.

Los reportes en los que aplicar el segmento dependerán de los KPI que se hayan elegido.

b) Comportamiento

El comportamiento es la actividad que realiza la gente en la web: qué es lo que hacen al llegar a ella, si añaden valor a la existencia online de la web, qué es lo que queremos que hagan y si lo llevan a cabo… En general, se pueden dividir en dos categorías: gente que

ve cierto número de páginas y gente que hace cierto número de cosas. Por ejemplo, podría ser interesante analizar las visitas que han accedido a más de tres páginas: de dónde provienen, si han participado en la web, qué tipo de contenido les interesa más… Otros ejemplos podrían ser: analizar las visitas que entran en la página principal y analizar qué hacen luego, o las que entran muy a menudo, o las que hacía 100 días que no entraban. Elegir los segmentos más interesantes solo se podrá hacer fruto de la experimentación, dada la complejidad y versatilidad del comportamiento de las visitas.

c) Resultados

Los resultados son las actividades del sitio que añaden valor. Posibles segmentos interesantes serían los de aquellas visitas que han reportado más dinero a la web (las "ballenas", llamados así por Voorhees et al. (2011) por primera vez), o los que han visto al menos cinco vídeos, o los que se han suscrito, o los que se han bajado un producto de prueba. Es decir, no solamente es importante segmentar por las macro conversiones, sino también por las micro conversiones.

Para estudiar y mejorar la eficiencia de una página web de manera individual, Kaushik (2007) señala que hay que seguir un programa de cinco pasos:

a) Medir más allá de la página principal: el SEO ha provocado que la mayoría de las visitas accedan a la web en una página profunda, por lo que cualquier página individual es como si fuera la página principal, ya que es la primera impresión de la visita.

b) Siguiendo el principio o distribución de Pareto, uno de los pilares de las leyes de potencias descritas por Newman (2004), el 80% del tráfico consume solo el 20% del contenido, por lo que hay que medir qué contenidos son los más consumidos para analizar las visitas que acceden a ellos y aplicar allí las mejoras. Una forma útil de segmentarlo es calculando el fuerte descenso que separa las páginas con más visitas del resto, como se observa de manera muy clara en la Figura 38:

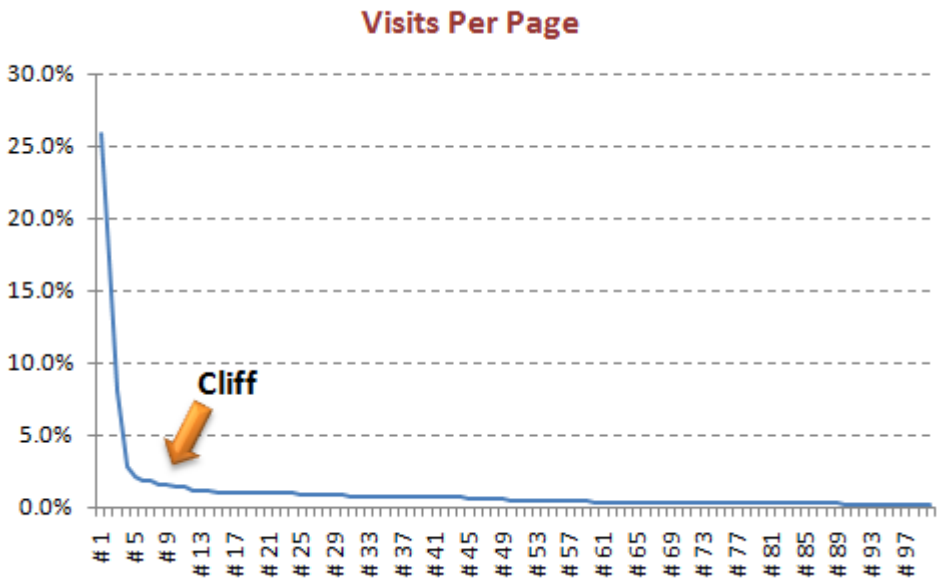


Figura 38. Ejemplo de comparación de páginas según su número de visitas (Kaushik, 2007)

c) Medir la ratio de rebote de las páginas más importantes. En contraposición de la ratio de salida (cuya entrada fue otra página), es crucial medir cuántas visitas entraron a esas páginas directamente y se fueron de la web sin visitar ninguna más. Posteriormente, hay que intentar averiguar la razón o las razones de dicho rebote: contenido, llamadas de acción, navegación, quiénes son esas visitas, cómo llegaron a dicha página, qué palabras clave usaron, si se trataba de una promoción expirada…

d) Medir la densidad de clic, en qué están clicando las visitas en las páginas más importantes.

e) Medir el contexto clave: páginas vistas, páginas vistas únicas, tiempo en la página, ratio de rebote, porcentaje de salida, palabras clave…

Para contrarrestar la saturación de datos, métricas y KPI, resulta conveniente según Kaushik (2010) crear reportes personalizados con las herramientas analíticas de que se dispongan. Algunos ejemplos podrían ser:

- Eficiencia de página: listado de páginas a partir de sus títulos, para facilitar su identificación posterior. Lo primero sería averiguar la frecuencia en la que dicha página es la primera que visitan los usuarios (entradas) y cuál es su efectividad (rebotes). Después, se analiza el consumo mediante los visitantes únicos (quién la ha visto), las páginas vistas (con qué frecuencia la han visto) y el tiempo en la página (cuánto contenido fue consumido de esta). Por último, se evalúa el valor

económico de la página (en qué medida aporta ingresos a la web) y los objetivos completados.

- Eficiencia de adquisición de visitas: listado de todas las fuentes de tráfico. Después se visualiza el número de visitas totales, los visitantes únicos y qué visitas son nuevas. También se analiza las visitas que han mostrado un comportamiento de valor. Por ejemplo: las visitas que están más de una cierta cantidad de tiempo o ven más de un cierto número de páginas. Para ello se identifica qué comportamiento significa que se ha captado la atención del visitante, se crea un objetivo para ello y se añade al reporte. Después, se analiza la ratio de conversión de los objetivos y el valor económico para detectar qué fuentes son más rentables. Y, por último, el coste de la visita (será 0 a no ser que provenga de Google AdWords y esté enlazado a Google Analytics).

2.2.1.5. Tests y experimentación

Los tests y la experimentación ayudan a analizar la intención o expectativa de un cliente / visitante y qué problemas encuentra. Kaushik (2006) explica en un artículo que los tres tipos de test más predominantes son:

a) Test A/B

Se trata de pruebas de más de una versión de la misma web. Cada versión normalmente será creada de manera única y separada. Por ejemplo, el objetivo podría ser probar tres versiones de la página principal, o de la página de producto, y ver qué versión obtiene mejores resultados. En la mayoría de los casos se hacen pruebas con un resultado (clics hacia la siguiente página, conversiones…). Lo óptimo sería probar las diferentes versiones al mismo tiempo, cargando de manera arbitraria una versión u otra. Si no es posible, se podría probar una versión por semana e intentar que los factores externos afecten en la menor medida posible.

Las ventajas de los test A/B serían que es quizá la manera más barata de hacer tests, al usar los recursos y las herramientas existentes. Además, es una buena manera de empezar a hacer tests antes de probar otros tipos de pruebas. Los inconvenientes serían que es difícil controlar todos los factores externos (campañas, tráfico de búsqueda, anuncios de prensa…) y no se puede confiar al 100% en los resultados; y que solo se puede hacer test con cosas simples y normalmente es difícil distinguir las correlaciones entre los elementos que se están probando.

Figura 39. Ejemplo de test A/B cambiando la imagen y la llamada de acción, según Quintero (2008)

b) Test multivariante

El test multivariante consiste en convertir ciertas zonas de la web en módulos que pueden cambiar dinámicamente su aspecto, su posición e incluso variar según el tráfico en sí, y medir qué versión de la página tuvo mejores resultados teniendo en cuenta también las correlaciones gracias a mecanismos matemáticos complejos.

Las ventajas son que gracias a herramientas como Google Website Optimizer[9] se puede realizar toda la funcionalidad de contenido, test de atributos, analítica y estadísticas de manera remota; y que puede resultar en una metodología de aprendizaje continuo. Los inconvenientes son que se necesita un conjunto limpio de ideas que provienen de posibles problemas de los visitantes u objetivos estratégicos de negocio, ya que de esta manera se evita optimizar con opciones que no se han estudiado lo suficiente y puedan ser perjudiciales; y que con el test multivariante solo se optimiza una página, así que no importa cuánto se haya optimizado, no podrá tener un papel externo en el resultado final, es decir, seguirá viéndose afectada por el resto de páginas visitadas por el usuario.

[9] https://optimize.withgoogle.com/

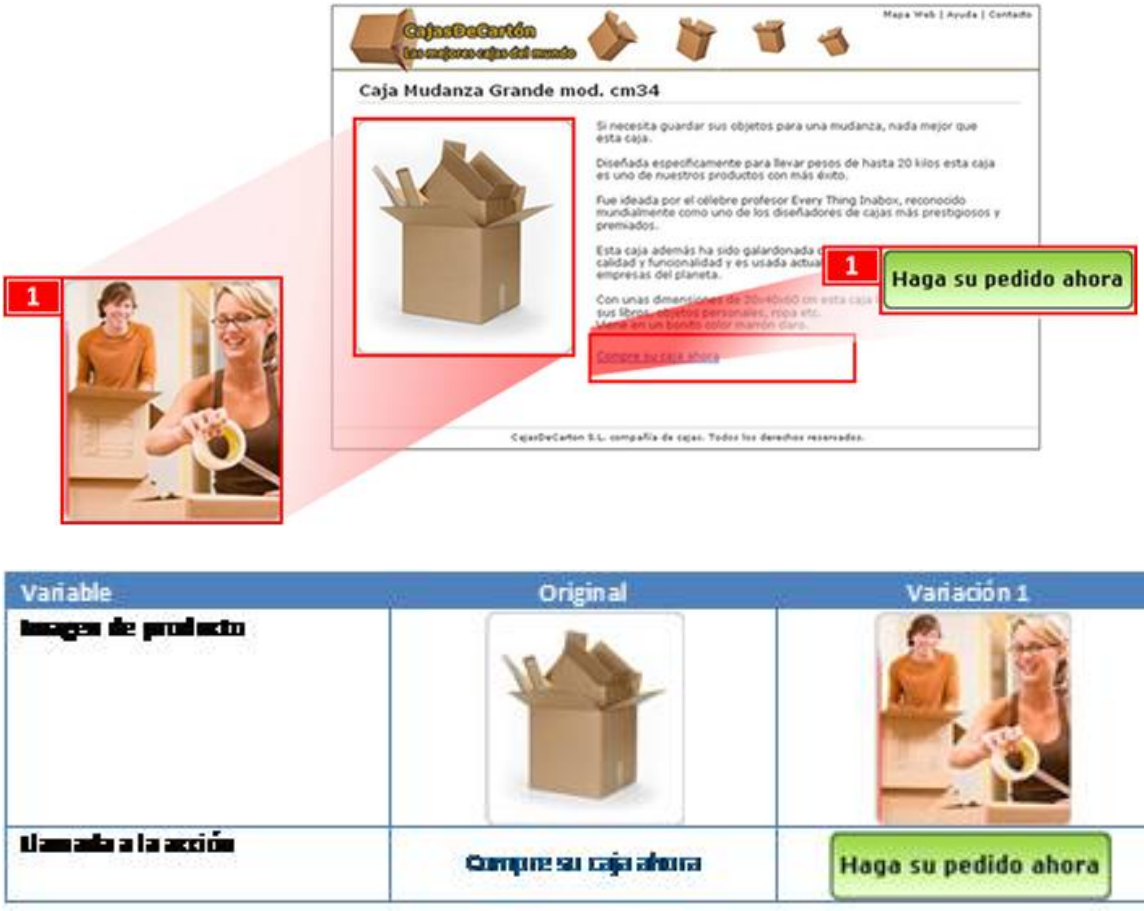


Figura 40. Ejemplo de test multivariante cambiando la imagen y la llamada de acción, según Quintero (2008)

c) Test de experiencia

Consiste en cambiar la experiencia de usuario de la totalidad de la web usando la capacidad de la plataforma de la web. En el test de experiencia no hace falta crear varias webs, sino que usando la misma plataforma de la web se pueden crear dos o tres experiencias persistentes y ver en cuál de ellas se obtienen mejores resultados. Por ejemplo: cambiar el fondo y el texto de color, quitar la navegación de la columna de la izquierda y poner fotos de la categoría del producto en vez de las fotos de cada producto en los recuadros correspondientes. Ya que las herramientas de estadística recogen los datos de todas las webs, el análisis se podrá hacer de manera simultánea.

Las ventajas de los test de experiencia son para Kaushik (2006) que se podrá hacer test con los usuarios en su ambiente natural y recoger los datos que reflejen de manera más cercana lo que piensan; que, si se integran métodos cualitativos, se puede leer literalmente los pensamientos de los usuarios sobre cada experiencia; y que se obtendrán resultados entre cinco y diez veces más poderosos que con cualquier otra metodología. Los inconvenientes son que se necesita una plataforma web que soporte test de experiencia;

que se necesita más tiempo que en las otras dos metodologías; y que se necesita de una implicación intelectual mucho mayor.

En una encuesta realizada por Econsultancy-RedEye en 2011 a 700 agencias de marketing digital, citada por Chaffey y Patron (2012), se comprobó que el test más utilizado era el A/B, mientras que otros como la segmentación, el test multivariante y las revisiones de usabilidad por parte de expertos eran los menos usados, tal y como se puede observar en la gráfica de la Figura 41. Los test de usabilidad tienden a ser más cualitativos, mientras que los test A/B, clic y reportes analíticos suelen ser cuantitativos. En la Figura 42 se puede ver la clasificación de los métodos de mejora de rendimiento según su tipología y enfoque:

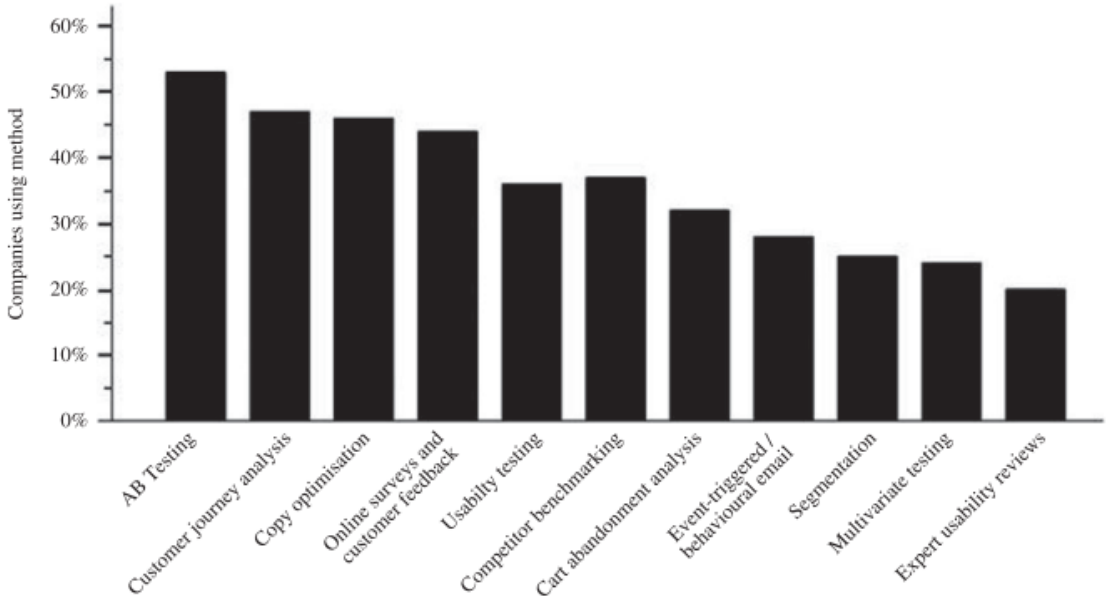


Figura 41. Métodos usados por las compañías para mejorar su ratio de conversión (Chaffey & Patron, 2012)

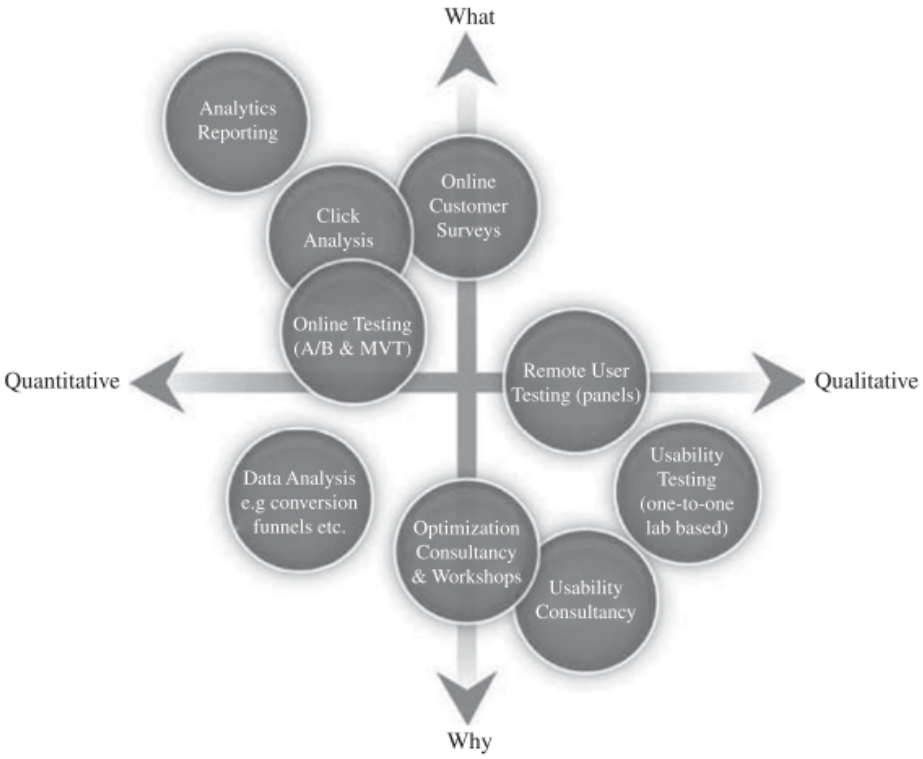


Figura 42. Clasificación de los métodos de mejora de rendimiento según su tipología y enfoque (Chaffey & Patron, 2012)

2.2.2. Cibermetría

Puesto que esta tesis propone una metodología cibermétrica, es conveniente añadir a este apartado un subcapítulo dedicado a la cibermetría, las áreas de trabajo y estudio, el rastreo de contenido y el análisis de la competencia.

2.2.2.1. Definición

La transición de los materiales impresos a los electrónicos y basados en la red provocó un incremento en el interés de los investigadores de la información de analizar cuantitativamente el ciberespacio, utilizando técnicas cuantitativas que le añadían una nueva dimensión al utilizar un acercamiento alejado del convencional que se llamó cibermetría. Se entiende como ciberinformación la información comunicada a través de medios electrónicos y que podría formar parte, según Shiri (1998), de las siguientes tipologías:

- Redes de información de todos los tipos; bases de datos online y metabases.
- Herramientas de internet como las webs, email, foros y grupos de noticias.
- Escuelas virtuales, universidades y organizaciones.
- Sistemas de tablones de anuncios.
- Conferencias, asociaciones y sociedades online.
- Libros, librerías, archivos y servicios de información electrónicos.
- Sistemas de información multimedia, hipermedia, polimedia y telemedia.

La cibermetría sería, por tanto, la medición, estudio y análisis cuantitativos de toda información del ciberespacio utilizando técnicas bibliométricas, cienciométricas y de la información (Shiri, 1998).

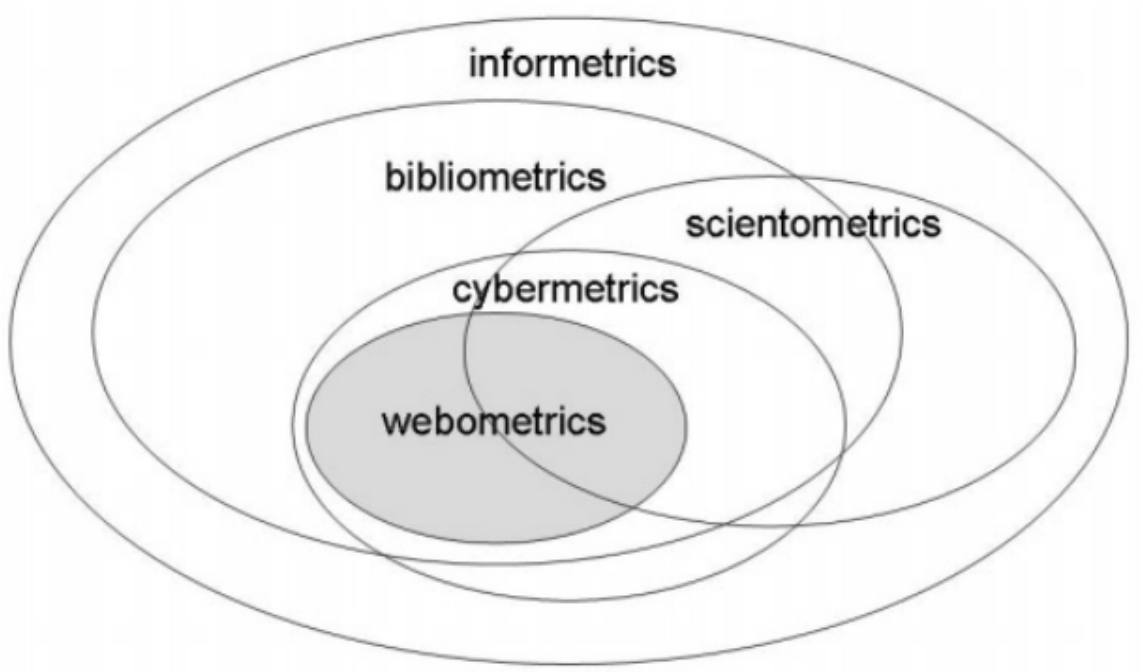


Figura 43. Las relaciones entre los campos de investigación de la informetría, bibliometría, cienciometría, cibermetría y webometría (Ingwersen & Björneborn, 2004, p. 339)

La bibliometría en sí ha sido utilizada en un gran número de contextos diferentes, según especifica McKiernan (2005), como los estudios científicos, la evaluación de la investigación, la gestión del conocimiento, la exploración del contexto, el análisis de tendencias y la optimización de recursos de bibliotecas e información. De esta manera, dentro de la bibliometría tradicional se pueden encontrar tres líneas de investigación principales:

- La informetría, que busca formalizar y modelar las normas de producción y uso de documentos así como sus implicaciones en el diseño y especificación de sistemas de información.
- La bibliometría, que estudia los procesos de comunicación en comunidades de conocimiento concretas, como los patrones de citas y la puntuación de documentos, y las implicaciones de éstos en la prestación de servicios de información.
- La cienciometría, que pretende entender la estructura y la dinámica de las comunidades del conocimiento y su productividad e impacto, con implicaciones en la política pública.

Mientras que la bibliometría había tenido mucho impacto antes de la aparición de internet, McKiernan (2005) explica que el surgimiento de internet y sus características inherentes hizo necesaria la aparición de la cibermetría como un acercamiento cuantitativo al nuevo

entorno hipermedia. Esto era debido a que Internet permitía por primera vez la aplicación de la investigación cuantitativa a gran escala.

En el entorno del periodismo, tuvo un gran impacto debido a la posibilidad de estudiar la frecuencia de citas de un artículo medido en un cierto período de tiempo, lo cual presentaba una gran variedad de utilidades según Rowlands (2000, p. 114): la clasificación, evaluación, categorización y comparación entre medios, factor de impacto para la guía editorial y de política de publicidad, etc. Todo ello a través de enlaces o hipervínculos como una fuente de información semántica adicional en la web. Se hace frente así a los inconvenientes del medio electrónico, entre los que se cuentan para Organista Sandoval & Cordero Arroyo (2001): la necesidad de un equipo para su lectura, la inexistencia de una regulación, la carencia de un código de ética establecido, la probable proliferación de información no regulada y la identificación de la información primaria frente a la secundaria.

El primer rastreo completo de internet se hizo entre julio y octubre de 1995 por Larson (1996), y devolvió 1,3 millones de documentos. En noviembre de ese mismo año hizo un segundo rastreo con 2,6 millones de documentos, duplicando así el tamaño en solo un mes. El crecimiento de la web ha provocado una necesidad de analizar los usuarios, contenidos y la estructura de internet en sí mismo. Dicho análisis comenzó siendo categorizado manualmente, y no se producía un análisis de citas, entendido como una herramienta para identificar conjuntos de artículos, autores y medios de campos particulares de estudio. Logró conjuntos de webs con similitudes de temática gracias a los enlaces entre estas, de una manera parecida a la que funcionan las citas en las fuentes académicas.

Inktomi Project, tal y como se describe en el artículo de Woodruff et al. (1996), examinó cerca de 2.6 millones de documentos HTML y estudió el uso de etiquetas, atributos, extensiones específicas de navegador, legibilidad, errores de sintaxis, etc. Se encontró inmediatamente con la dificultad de procesar una cantidad tan grande de datos, ya que en su momento ninguna de las herramientas existentes podía escalar adecuadamente a una muestra de ese tamaño.

Si la informetría es la investigación de información en un sentido amplio, la investigación de toda comunicación basada en la red aplicando métodos informétricos u otras medidas cuantitativas en la web se llamó webmetría o cibermetría, según especifican Almind e Ingwersen (1997). De esta manera, las webs se convierten en las entidades a estudiar gracias a los hipervínculos que actúan como citas. Así, se puede combinar el potencial de la indexación de citas y las bases de datos de textos completos, ya que la web permite

buscar una red de citas que además contiene los textos completos e incluso objetivos multimedia, si éstos son incorporados a las webs que son analizadas. Sin embargo, el gran tamaño y el dinamismo de internet dificulta mucho su indexación, por lo que Almind e Ingwersen (1997) propusieron una estrategia basada en crear una indexación distribuida de la web tal y como se observa en la Figura 44, pero se encontraron con el problema de que no se podía realizar un análisis de citas en la práctica, ya que no lograron encontrar una colección de webs que estuvieran lo suficientemente enlazadas para que fuera posible realizar la medición.

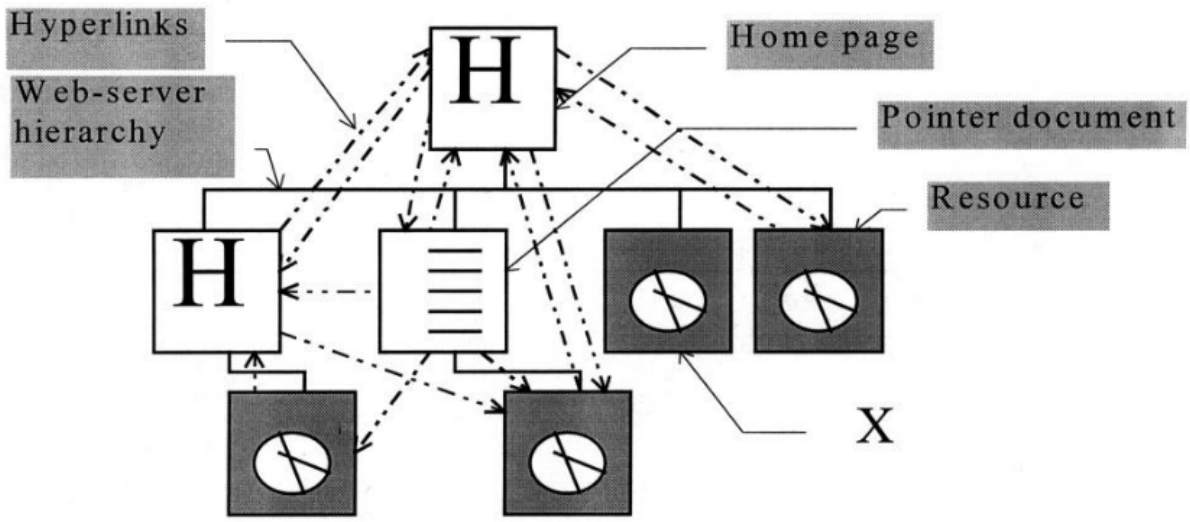


Figura 44. Modelo simplificado de relaciones entre sitios web (Almind & Ingwersen, 1997).

La cibermetría refleja, por tanto, las prioridades del usuario durante su navegación, la secuencia de pasos realizados en su sesión, las páginas destino visitadas y otras características pertenecientes a la experiencia de uso de dicho usuario. De esta manera, para Melnikov y Schönwälder (2010), la identificación de los usuarios basada en elementos cibermétricos aportaría ventajas para la administración del sistema, el mantenimiento de la red y la seguridad, como la verificación de la identidad o la habilitación del acceso para servicios específicos.

Aguillo (2000) por su parte evaluó el crecimiento explosivo de la información médica en internet gracias a herramientas en el lado del cliente para lidiar con grandes cantidades de datos e incrementar la usabilidad de los resultados obtenidos. Para ello, utilizó indicadores cibermétricos tales como el tamaño (número de páginas) y densidad (número de enlaces) de las webs, las citas a sí mismos, la visibilidad y la diversidad de estas. El objetivo era la clasificación de las webs y sugerir límites mínimos gracias a los cuales excluir webs que no tuvieran la cualificación suficiente para pertenecer a un directorio de contenido médico.

2.2.2.2. Áreas de trabajo e indicadores del análisis cibermétrico

Orduña-Malea y Aguillo (2015, p. 118) dividen la cibermetría en las siguientes áreas de trabajo, como también se observa en el esquema de la Figura 45:

- Cibermetría descriptiva: desarrollo teórico de la disciplina y definición y modelización de indicadores cibermétricos, así como las unidades de medida y su interpretación.
- Cibermetría instrumental: estudio del funcionamiento, cobertura y limitaciones de las fuentes de información cibermétricas (robots y motores de búsqueda), así como los métodos de atracción, análisis y visualización de información.
- Cibermetría aplicada: combinación de los indicadores cibermétricos en contextos específicos (entidades, temas, etc.), como pueden ser condiciones de contorno, académicas, sociales...

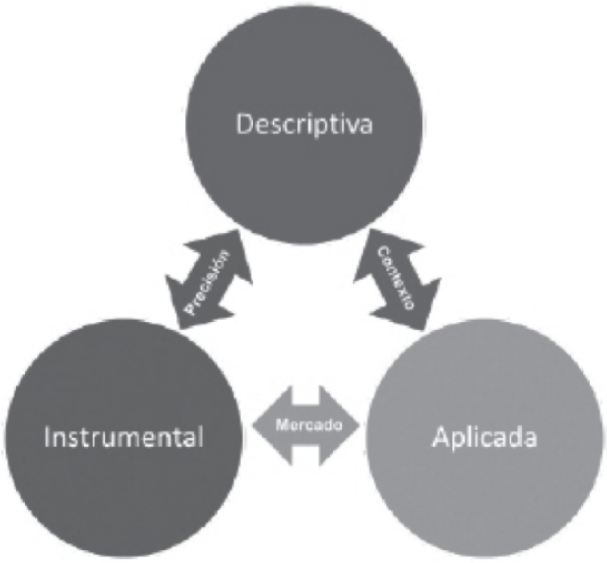


Figura 45. Interacciones de las áreas de trabajo de la cibermetría (Orduña-Malea & Aguillo, 2015).

Los indicadores cibermétricos a su vez podrían clasificarse, según Martínez Rodríguez (2006), en tres grupos:

- Medidas descriptivas: tamaño o número de objetos a estudiar, como la densidad de enlaces o el número de páginas. Útiles para medir la profundidad de Internet según parámetros como los países, organizaciones o grupos de usuarios.
- Medidas de visibilidad o impacto: patrones de enlaces entre páginas y webs distintas. Miden el número y la diversidad de éstos, el volumen con respecto al contenido, índices de peso relativo como el PageRank de Google...

- Medidas de popularidad: número y características de visitas de una web. Pese a lo complicado de obtener valores absolutos, los valores relativos pueden servir para análisis comparativos.

El proyecto europeo WISER, en el que se ha basado el estudio de Scharnhorst & Wouters (2006), demostró que no se pueden extraer los indicadores web desde un rastreo puramente automatizado, por lo que es necesaria una combinación de estudios cuantitativos y cualitativos para determinar la unidad de análisis, los límites de la medición debido al acercamiento técnico utilizado y el significado de aquello que se mida. Todo ello provoca una cierta tensión, ya que los estudios cuantitativos y cualitativos utilizan diferentes conceptos de muestreo y tamaño de muestra.

Siguiendo dicha lógica, Malinsky y Jelínek (2010) aplicaron el análisis de sentimientos a la cibermetría para reflejar mejor el contenido semántico de las páginas web, siendo el análisis de sentimientos un complemento cualitativo al acercamiento cuantitativo. Utilizaron análisis lingüísticos del texto y métodos matemáticos y estadísticos y alcanzaron una mejor comprensión de una página web y su significado semántico. El objetivo de este enfoque es el de reducir resultados irrelevantes en una búsqueda y facilitar el acceso de información al usuario. El análisis de sentimientos busca identificar la polaridad positiva y negativa de un texto y reconocer una impresión subjetiva u objetiva del mismo, siendo también un objetivo secundario el de determinar la actitud del autor al escribir o el efecto emocional intencionado que el autor pretende provocar en el lector.

La cibermetría identifica los elementos que constituyen el hiperespacio, es decir, la infraestructura (física, lógica y de comunicación), los usuarios y el contenido. Sin embargo, todo ello debe ser contextualizado, ya que Orduña-Malea y Alonso-Arroyo (2018, p. 9) estimaron que el 75% de las webs de Internet podrían estar inactivas, y el número de usuarios y de webs cambia en cada instante de tiempo, como se aprecia en la gráfica de la Figura 46.

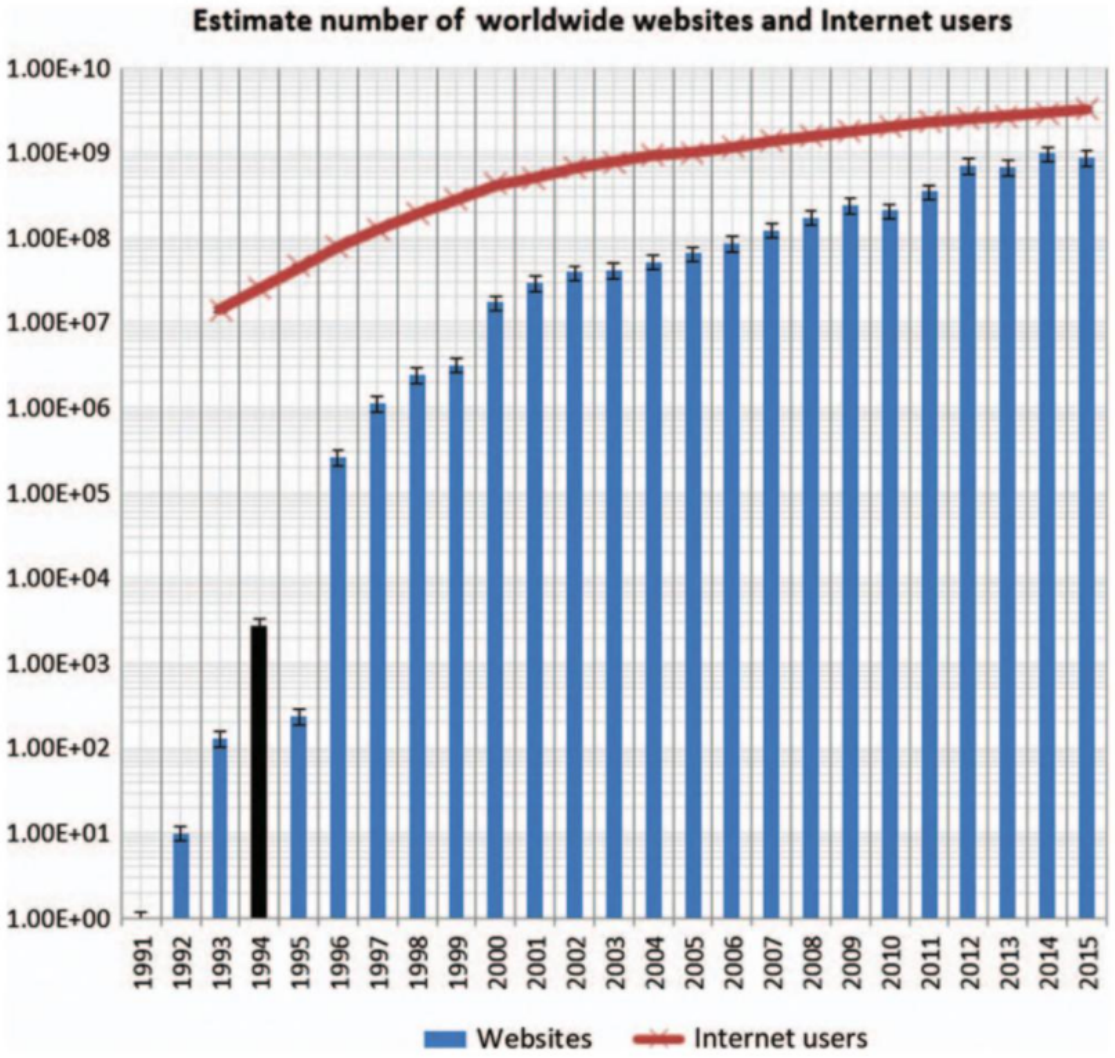


Figura 46. Número de webs y de usuarios estimados según el año (Orduña-Malea & Alonso-Arroyo, 2018, p. 10).

Según Danman Fugl (2001), hay dos factores para tener en cuenta antes de realizar ningún análisis cibermétrico, ya que son los que forman la base de este: los documentos a analizar y las herramientas que se aplicarán a las colecciones de datos. Las dos metodologías principales son la teoría de citas y la teoría de enlaces, ya que tanto la informetría como la cibermetría dependen de la legalidad y cuantificación de ambos. Así, para analizar webmétricamente a un macro-nivel se necesitaría una teoría normativa de enlaces globalmente homogénea, es decir, que incluyan los enlaces siguiendo unas mismas normas, y un control de procesos similar al proceso de revisión de pares. Siendo esto último utópico según el autor, sugiere una examinación cualitativa de los tipos de webs a incluir en el análisis para asegurar que cumplan las normas de enlaces a analizar.

Las principales áreas de investigación dentro de la cibermetría, según Ingwersen y Björneborn (2004, p. 339), son los problemas de cobertura, calidad y muestreo de

rastreadores de webs y de buscadores; el análisis de enlaces (incluyendo el análisis de impacto web); y los estudios del comportamiento de las interacciones web mediante logs. En el caso de los enlaces, no son necesariamente normativos, es decir, no tienen por qué proveer de crédito o reconocimiento, son más bien funcionales y navegacionales por naturaleza. En cuanto a los rastreadores de webs y buscadores, suelen realizar su procesamiento mediante enlaces o contenidos con objetivos como el estudio de espacios web concretos, indicadores web o interacciones web. Por otro lado, los mismos autores señalan una serie de problemáticas habituales cuando se realizan colecciones de datos y análisis de estudios de la web:

- Los métodos de iniciación de colecciones de datos web y su amplitud.
- Las estrategias de recuperación de datos y frecuencia de actualización de éstos por parte de los motores de búsqueda.
- Las variaciones de accesibilidad entre los motores de búsqueda.
- El algoritmo de clasificación de páginas de los motores de búsqueda.
- Los sesgos de género, nacionalidad, etc.
- Las limitaciones de profundidad del rastreo de webs.
- La limitación de número de páginas por web visitada.
- Las omisiones de webs: webs que están aisladas sin enlaces entrantes, o con formatos no comprendidos por el rastreador, o protegidas con contraseña u otros métodos...
- Páginas principales no incluidas en las estructuras de enlaces de una web.
- Nombres de dominios diferentes utilizados por las mismas entidades.
- Variación de la cualidad según los tipos de página o de datos de una web.
- Métodos de muestreo y significación.

2.2.2.3. Rastreo de contenidos y redes de enlaces

Los rastreadores, también llamados *crawlers*, *spiders*, *wanderers*, robots o *bots*, son aplicaciones informáticas que recuperan las páginas web extrayendo de estas sus redes de enlaces y, al mismo tiempo, las recorren mediante tres formas diferentes, según Sánchez-Pita & Alonso-Berrocal (2013): recorrido en anchura (*breadth-first*), recorrido en

profundidad (*depth-first*) y el mejor posible (*best-first*).

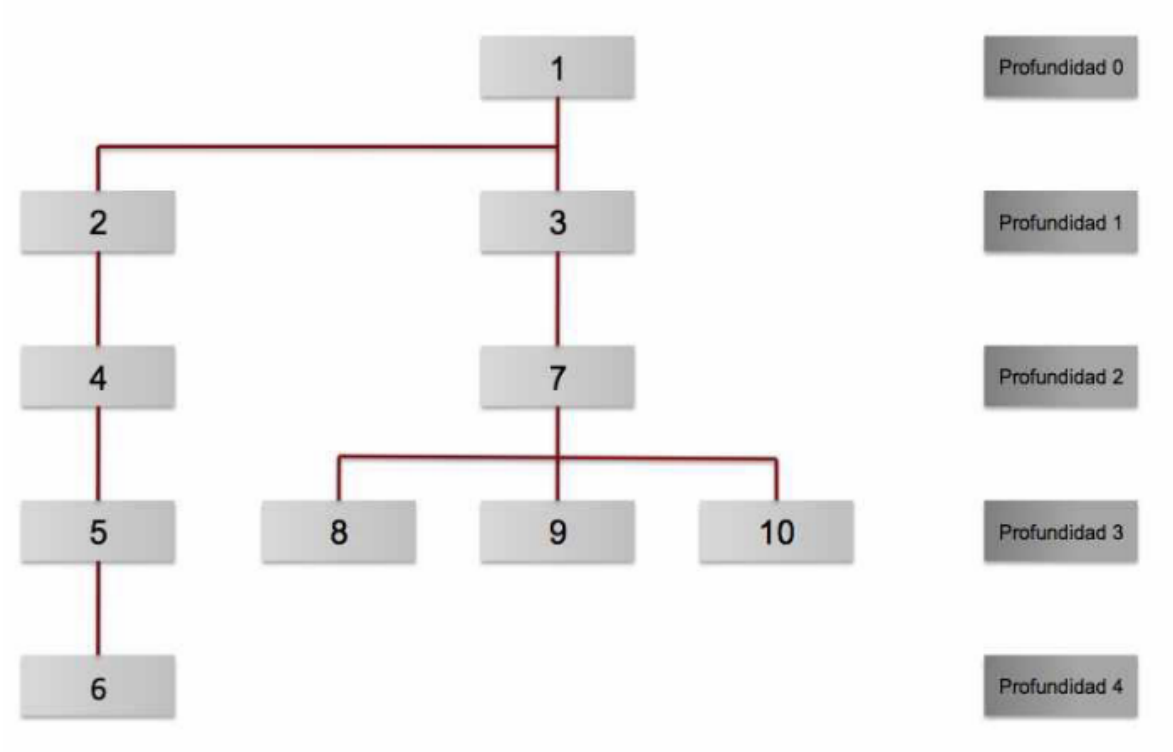


Figura 47. Esquema de profundidad del recorrido de un rastreador (Sánchez-Pita & Alonso-Berrocal, 2013)

Ortega Priego (2004), en cambio, apostó por el uso de ficheros log como manera de aplicar minería web a la revista Cybermetrics en 2004. Se encontró con el problema de las IPs dinámicas y el creciente uso de proxis, que dificulta la identificación del usuario y en su estudio provocó un 24,9% de margen de error. Por otro lado, señaló que también era complicado cuantificar el consumo de información en Internet debido a la dificultad de conocer la intención de información que llevaba al usuario a visitar la web, ya sea mediante un enlace entrante o un buscador. Habría que añadir el hecho de que en ocasiones se visitan archivos de tipo pdf (*Portable Document Format*) o ps (*PostScript*), por lo que algunas publicaciones los incluían en sus mediciones. Y si, como es el caso de la revista Cybermetrics, no se producía una publicación regular y con un número suficiente de artículos, no permite la detección de conductas o patrones que se puedan extrapolar a otras webs. Pese a todo ello, sí que estudió de manera óptima el consumo web de la revista a través logs, gracias al cual pudo constatar qué secciones se leían más, la procedencia de los usuarios y en qué campos era más relevante la revista.

En un estudio posterior, Ortega Priego y Aguillo (2009) declararon la necesidad de limpiar los datos eliminando los accesos que pudieran dificultar el análisis. Así, si el interés está situado en las páginas web, es necesario eliminar los accesos a objetos y elementos

gráficos (ficheros .gif, .jpg, .bmp...) y scripts (.js, .cgi, .pl...). Además, para ayudar a evitar el acceso de robots, se hace uso del archivo “robots.txt” situado en el directorio principal de la web, en el cual se pueden incluir las normas que los robots deberán aplicar: qué páginas podrá visitar, con qué frecuencia y otros parámetros.

Para relacionar las visitas entre sí y las páginas que han sido visitadas, Ortega Priego y Aguillo (2009) definen dos conceptos:

- Soporte (sop): proporción de sesiones que han visitado tanto la página A como la B, de manera que si por cada 10 visitas, 4 han accedido a la página A y B, el soporte de dicha relación es sop = 4/10 = 0,4.

$$sop\left(A \longrightarrow B\right) = sop\left(A \cup B\right)$$

Figura 48. Fórmula del soporte de la relación entre una página A y una página B (Ortega Priego & Aguillo, 2009)

- Confianza (conf): proporción de sesiones que han visitado tanto la página A como la B (soporte de A y B) entre las que solo acceden a la página A (soporte de A). De esta manera se puede calcular la probabilidad de que, si se visita una página, también se visite la otra.

$$conf\left(A \longrightarrow B\right) = \frac{sop\left(A \cap B\right)}{sop\left(A\right)}$$

Figura 49. Fórmula de la confianza de la relación entre una página A y una página B (Ortega Priego & Aguillo, 2009).

El concepto de web es el de un grupo de páginas que tienen una relación semántica concreta y están relacionadas por una estructura. Puede coincidir con el dominio físico o no, debido a lo que algunos autores como Arroyo Vázquez y Pareja Pérez (2003) lo definen como “sede web” para alejarse de otros que consideran más confusos como lugar web, página web o página principal. La web o “sede web” sería una unidad documental reconocible e independiente de otras, ya sea por su temática, por su autoría o por su representatividad institucional. Consiste en una página web o conjunto de ellas que están enlazadas jerárquicamente a una página principal. Por ejemplo, la web de Apple sería el conjunto de contenidos publicados bajo el dominio “www.apple.com”, siendo distinguible de otras webs ya que están alojadas en otros dominios o subdominios. Según los mismos autores, las webs se pueden clasificar en institucionales, temáticas y personales, siendo

codificadas de manera institucional, geográfica a varios niveles y por materias o temáticas. En esta tesis se usará indistintamente web o sitio web en referencia al mismo concepto de "sede web" planteado antes, ya que su uso no es habitual en el momento de la publicación de este estudio.

El rastreo puede ser de contenido y de enlaces, según Thelwall y Vaughan (2004). El primero se centra en conseguir un listado de páginas distintas y rechazar el contenido duplicado, mientras que el rastreo de enlaces ignora dicho problema y busca un análisis topológico de la web. Sin embargo, se ha apreciado una relación de similitud entre el contenido de webs diferentes si estas están vinculadas entre sí, lo cual significa que el rastreo de enlaces podría responder a ambas cuestiones. Por otro lado, los logs o historiales de los servidores web también aportan información sobre cómo los usuarios interaccionan con las webs, principalmente gracias al tiempo de espera o *timeout*, que consiste en que una sesión se considera terminada cuando se ha producido un determinado número de minutos de inactividad. La elección de ese número de minutos supone un límite para este acercamiento, por lo que es necesario estandarizarlo mediante el método estándar, el método de longitud de referencia y el de máxima referencia anticipada.

Las tres direcciones principales a la hora de estudiar el conocimiento en internet son a través de los contenidos de las páginas web, las estructuras de enlaces y la información del comportamiento de los usuarios. Björneborn e Ingwersen (2001) estudiaron las estructuras de enlaces a través de la teoría de grafos, siendo esta una representación matemática de una red consistente en vértices o nodos conectados por aristas. De esta manera se puede estudiar los aspectos estructurales de internet gracias a rastreos que siguen los enlaces de las webs. Estas pueden tener solo enlaces salientes, solo enlaces entrantes, ambos o ninguno. Las webs se incluyen así en clústeres o grupos de webs de una temática relacionada, en las que también pueden aparecer enlaces transversales hacia otras temáticas creando de esta manera vinculaciones entre dos clústeres heterogéneos de webs.

Para Björneborn e Ingwersen (2001) hay una falta de fiabilidad en los datos secundarios de los buscadores, provocada por la incerteza de la cobertura, la frecuencia de actualización, las reglas de indexación, el rendimiento del procesamiento informático, los algoritmos de clasificación, etc. La diversidad de las personas creadoras de contenidos web y enlaces también afecta a la calidad y fiabilidad de estos documentos online, así como la falta de metadatos adjuntos a todo ello. Es por ello por lo que estos autores sugieren recoger la información directamente de las webs gracias a los caminos aleatorios

de la teoría de grafos. Para localizar los enlaces transversales, sería necesario realizar un gran número de caminos aleatorios a través de los enlaces y aplicar criterios de heterogeneidad entre los dominios de las webs gracias a la coocurrencia de palabras similares.

Alonso Berrocal et al. (2001), en cambio, proponen el concepto de las leyes de exponenciación, mediante las cuales evalúan internet pese a sus características de crecimiento aleatorio y desregulado, aprovechando así la ventaja de que su procesamiento sea más fácil y rápido. Las leyes de exponenciación se basan en la premisa de que Internet se rige por unas pautas de funcionamiento genéricas que pueden servir para analizar diferentes webs. En su estudio sobre la cibermetría de Internet, presentan las siguientes leyes: Exponente de Orden R, Exponente de Grado de Apertura O, Exponente de representación de saltos H y Exponente de valores propios “?”.

Con todo lo anterior, el estudio de la web se puede realizar mediante análisis cuantitativo, medidas topológicas vinculadas al análisis de citas y el factor de impacto, y las leyes de exponenciación para analizar el crecimiento del contenido online. El mayor desafío según Berrocal et al. (2002) es que la información no está estructurada, por lo que es necesario aplicar formalismos para representar los documentos gracias a los términos como elementos básicos de su representación. No obstante, la falta de una normalización también en este aspecto ha dificultado la clasificación de las diferentes variantes. También es crítico elegir un punto de partida lo más cercano posible a los nodos más relevantes de la red. Un método para ello es el uso de servicios metabuscadores que recogen los resultados, los organizan y los devuelven, siendo éstos los candidatos a ser puntos de entrada. Es recomendable la utilización de diferentes agentes que permitan el análisis desde puntos de entrada diferentes, ya que así se puede hacer un procesamiento paralelo y hacer más improbable la aparición de cuellos de botella, servidores lentos, etc. También es posible la priorización de los puntos de entrada según las especificaciones del usuario y el contenido de estos.

Una vez obtenidos esos puntos de entrada, se procedería a la extracción de sus enlaces, almacenaje en una lista y procesamiento de éstos para recuperar las páginas a las que apuntan y aplicar sucesivamente el mismo proceso a estas. El orden de los elementos obtenidos será, indican Berrocal et al. (2002), según la relevancia de los enlaces (atendiendo a las especificaciones del usuario) y la posibilidad de acceder a redes más amplias de páginas web. Es por ello por lo que la importancia de una página se podría determinar según el número de *backlinks* que tenga, es decir, el número de páginas que le enlacen. De ahí surge el algoritmo llamado PageRank, que determina la importancia de

una página de manera directamente proporcional a su número de *backlinks* y el peso o importancia de éstos. El inconveniente de este procedimiento es que es muy costoso en tiempo de proceso. Todo lo anterior supone una aproximación a la estimación de la proximidad de un nodo a las necesidades informativas de un usuario.

El PageRank es una técnica de cálculo del posicionamiento que simula un usuario que navega aleatoriamente en Internet, navegando a una página aleatoria con probabilidad que, o que sigue un enlace aleatorio de la página actual con probabilidad 1 - q. Dicho proceso se modela, según Alonso Berrocal et al. (2008), como una cadena de Markov con posibilidad de calcular la probabilidad estacionaria de estar en cada página. La cadena de Markov es un concepto estadístico en el que se estudia la fuerte dependencia entre un evento y otro anterior dentro de un sistema aleatorio. La relevancia resultante de una página sería la dada por todas las páginas que la enlazan:

$$PR(a) = q + (1-q)\sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

Figura 50. Fórmula del PageRank (Alonso Berrocal, et al., 2008).

En la anterior fórmula, un usuario navega a una página aleatoria con probabilidad q, o sigue un enlace aleatorio desde la página actual con probabilidad 1 – q.

Las fuentes de autoridad serían las webs con una ratio alto de enlaces entrantes. En cuanto a los tipos de enlace recíprocos, Danman Fugl (2001) los clasifica en tres: reciprocidad verdadera o enlaces entre tres o cuatro páginas de dos webs diferentes, como se puede ver en la Figura 51, la Figura 52 y Figura 53 a continuación:

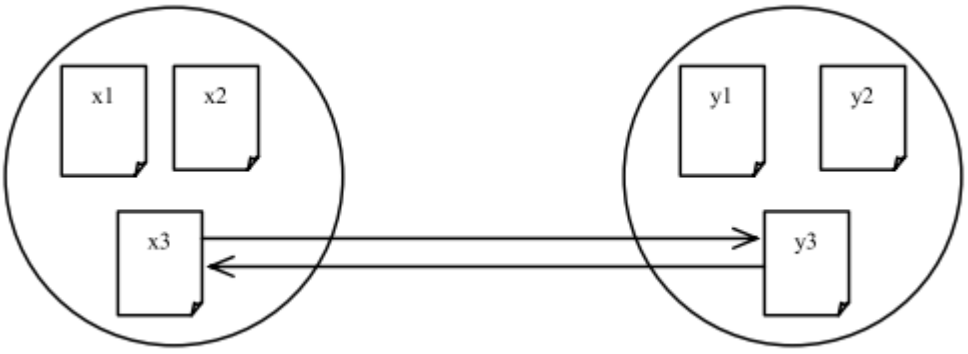


Figura 51. Reciprocidad verdadera entre dos páginas de dos webs (Danman Fugl, 2001).

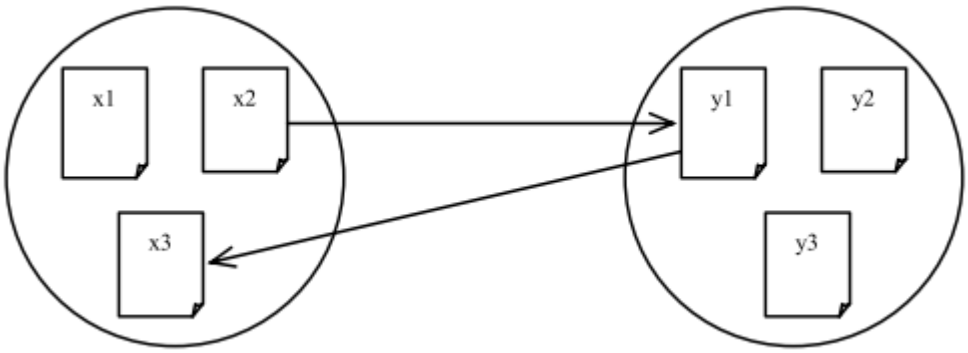


Figura 52. Reciprocidad entre tres páginas y dos webs (Danman Fugl, 2001).

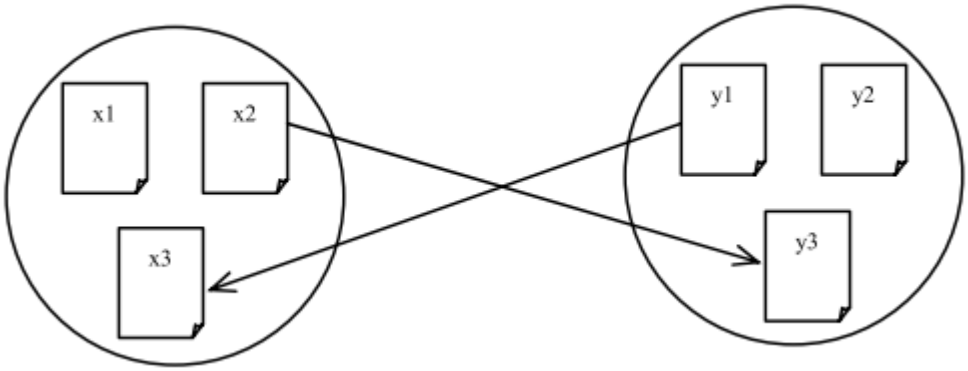


Figura 53. Reciprocidad entre cuatro páginas y dos webs (Danman Fugl, 2001).

Las tres medidas más comunes de la actividad en una web son, para Jana y Chatterjee (2004), las peticiones de archivos por parte de un navegador, las páginas vistas y las sesiones realizadas por los usuarios, todo ello en un periodo determinado de tiempo. Estas medidas dan lugar a la siguiente serie de métricas:

- Peticiones: de toda la web, una media por día y las realizadas a la página principal.
- Páginas vistas: páginas vistas totales, media por día, media por visitante único y visualizaciones de documentos.
- Visitas: visitas totales, media por día, duración media de la visita, duración mediana de la visita, visitas internacionales, visitas de origen desconocido, visitas desde el país a estudiar, visitas provenientes de buscadores y visitas de robots.
- Visitantes: visitantes únicos, visitantes que solo realizaron una visita y visitantes que realizaron más de una visita.

Con ánimo de clasificar los tipos de enlaces de una manera más amplia, Ingwersen y Björneborn (2004, p. 339) aportan el siguiente ejemplo de la Figura 54 en uno de sus

estudios:

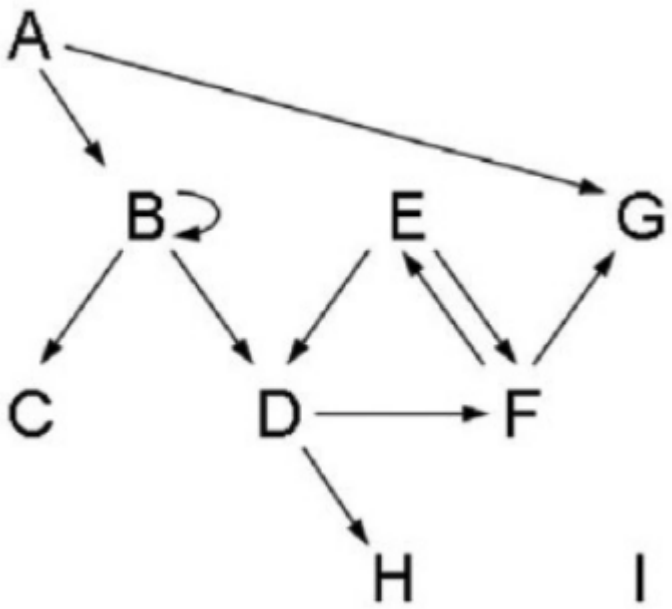


Figura 54. Grafo para terminología básica de enlaces webométricos en el que cada letra representa un nodo web diferente (Ingwersen & Björneborn, 2004, p. 344).

Siguiendo el anterior esquema:

- B tiene un enlace entrante desde A (inlink). A le enlaza, por tanto, está dentro del vecindario entrante de B.
- B tiene un enlace saliente hacia C (outlink), por lo que C está en vecindario saliente de B.
- B tiene un enlace a sí mismo o auto enlace (*selflink*).
- A no tiene ningún enlace entrante, por lo que no está enlazada.
- C no tiene enlaces salientes, por lo que no está enlazando a nada.
- I no tiene enlaces entrantes ni salientes, así que está aislada.
- E y F tienen enlaces recíprocos.
- D, E y F tienen enlaces entrantes y salientes conectándose entre ellos, así que están inter-enlazados tríadicamente.
- A tiene un enlace saliente transversal hacia G, que funciona a modo de atajo.
- H es alcanzable desde A por un camino directo.

- C y D están coenlazados por B, así que C y D tienen enlaces coentrantes (*co-inlinks*).
- B y E están coenlazando a D, por lo que B y E tienen enlace cosalientes (*co-outlinks*).
- Los enlaces coentrantes y cosalientes son ambos coenlaces (*co-links*).

Ingwersen (1998) definió el Factor de Impacto Web (FIW) como la suma del número de enlaces externos y auto enlaces que apuntan a un país o sitio web en concreto, dividido por el número de páginas en ese país o sitio web en un instante de tiempo determinado. A partir de esta definición se obtienen tres cálculos:

- FIW total: número de enlaces recibidos por un sitio web dividido entre el número de páginas de este.
- FIW externo: número de enlaces externos recibidos por un sitio web dividido entre el número de páginas de este, considerado el más apropiado para medir la relevancia de una web.
- FIW interno: número de enlaces internos recibidos por un sitio web dividido entre el número de páginas de este.

Sin embargo, el mismo Ingwersen (2009) incidió en que el FIW no tiene nada que ver con el reconocimiento y menos aún con la calidad, ya que la mayoría de los enlaces son navegacionales. Debido a la falta de normalización, no pueden ser consideradas citas, ya que no existen convenciones en su construcción. Es por ello por lo que resulta más conveniente medir el reconocimiento o impacto de un sitio web teniendo en cuenta el número de enlaces entrantes de esta (sin tener en cuenta sus enlaces internos).

Sánchez-Pita y Alonso-Berrocal (2013) miden la densidad de los enlaces como la proporción de enlaces entre las relaciones posibles de una red, pudiendo variar entre 0 y 1, siendo los próximos al 1 los mejores resultados:

$$D = \frac{r}{N(N-1)} \qquad \text{N: nº de nodos} \qquad \text{r: nº de enlaces}$$

Figura 55. Fórmula para calcular la densidad de enlaces de una red (Sánchez-Pita & Alonso-Berrocal, 2013).

Por otro lado, los mismos autores calculan el diámetro como la distancia geodésica más larga del grafo obtenido del análisis de una red.

Las webs, además, pueden subdividirse en dominios derivados (subdominios) y directorios dentro de la misma jerarquía del dominio principal, a modo de subsitios, sub-subsitios y sucesivamente, siguiendo la tesis de Björneborn & Ingwersen (2004). En la Figura 56, cada círculo sería una web, un círculo doble sería una web hija y un círculo triple sería una web hija de una web hija:

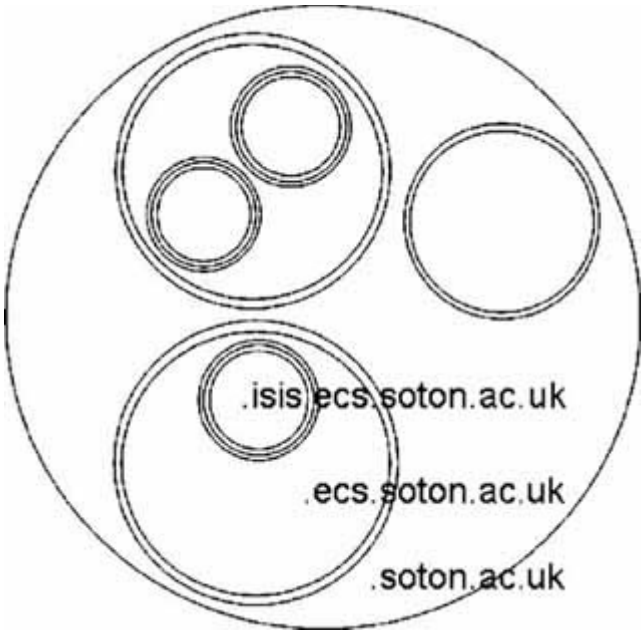


Figura 56. Diagrama simplificado de nodos web de un sitio web contenedor de subsitios web y sub-subsitios web (Björneborn & Ingwersen, 2004).

Pero Orduña-Malea y Alonso-Arroyo (2018, p. 38) especifican que la dispersión no solo se produce a nivel interno sino externo, consistente en una distribución horizontal del contenido en varias webs que pertenecen a la misma entidad. Webs corporativas, delegaciones, webs de productos, webs de servicios o fundaciones son algunos ejemplos, como se pueden ver en la Figura 57.

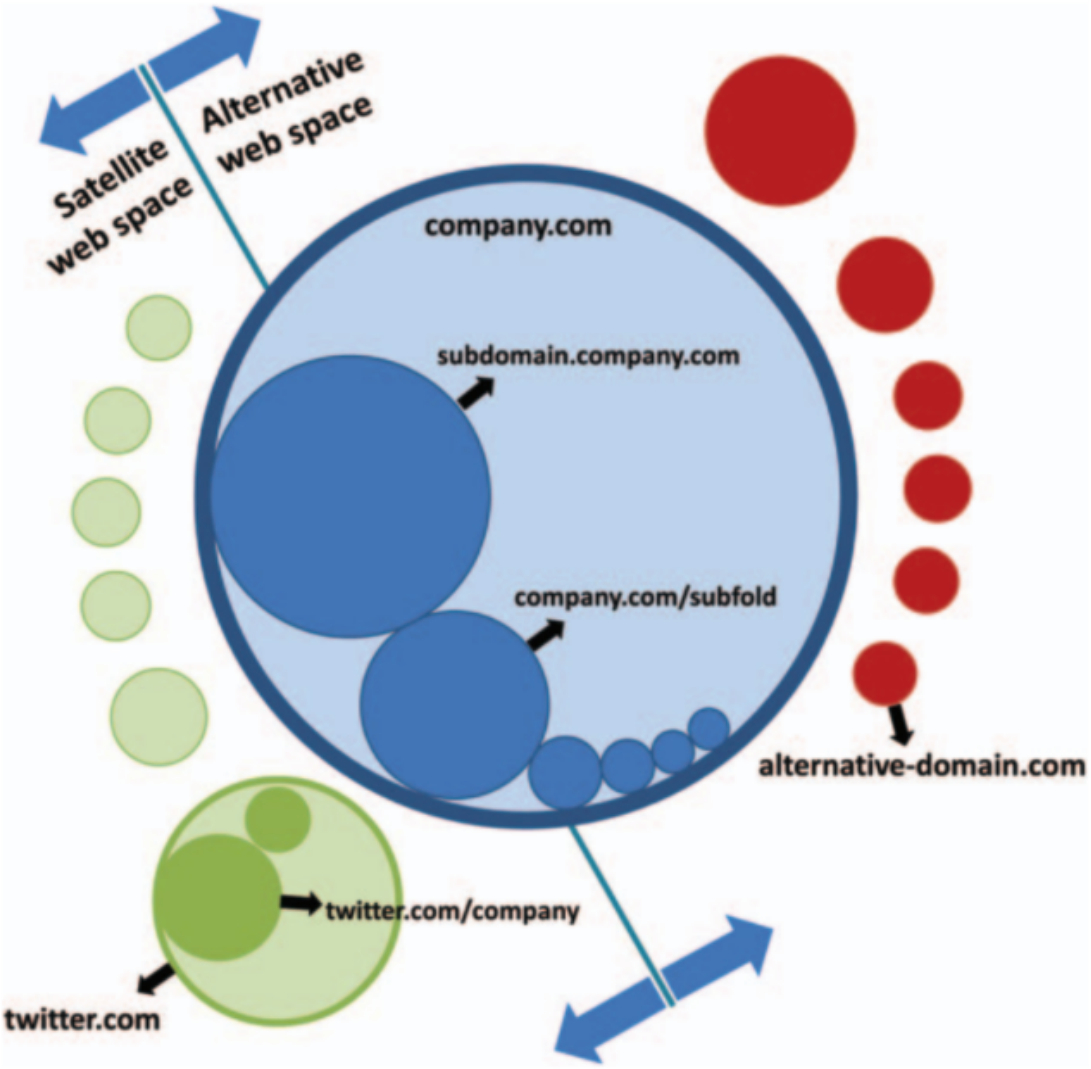


Figura 57. Dispersión web de tipo interna y externa (Orduña-Malea & Alonso-Arroyo, 2018, p. 39).

Debido a la falta de normalización en la forma en que se enlazan las diferentes webs, su similitud con las citas debe ser tomada con cautela y, por eso, Smith (2006) estudió Google Scholar como una herramienta que tiene la ventaja de seleccionar material orientado a la investigación y el potencial de ser un índice de citas para la bibliometría. Sin embargo, también señaló que Google Scholar no era transparente en sus algoritmos y alcance ni tenía funciones orientadas a la búsqueda cibermétrica. No obstante, se trata de una herramienta que correlaciona con los métodos convencionales de la investigación y es muy útil como una forma simple de obtener contenidos web basados en la investigación.

Orduña-Malea (2019), por su parte, analizó si Google Trends[10] pudiera ser válido para análisis cibermétricos, llegando a la conclusión de que no se puede utilizar para esa

[10] https://trends.google.es/trends/?geo=ES

finalidad. Esto es debido a que la asignación de temas y entidades a los términos de búsqueda no es consistente; el explorador no dispone de los comandos de búsqueda suficientes para expresiones de búsqueda complejas; el filtro de categorías y subcategorías no es completo y está sesgado; y los indicadores de interés relativo presentan limitaciones como que no se conocen los volúmenes de búsquedas por término. Esto último puede provocar malas interpretaciones, pues no se puede comprobar la precisión ni reproducir los indicadores.

2.2.2.4. Áreas de estudio

Con ánimo de medir la anchura y amplitud de Internet, han surgido las siguientes áreas de estudio: (Alonso Berrocal, et al., 2008)

- Número de sitios web y páginas principales en el mundo y su distribución por países.
- Clasificación de las páginas web por tipos de documentos.
- Número de páginas web por dominios.
- Clasificación de páginas web por el idioma de sus documentos y el modo en que éstos se representan.
- Estadísticas de uso y usuarios en un período de tiempo determinado.
- Número de citas recibidas por cada página web.
- Ordenación de webs y páginas más citadas según el tipo de documento.
- Tipos de colecciones electrónicas disponibles en cada sitio web.
- Factor de Impacto Web (FIW) y productividad de los autores.
- Análisis del contenido de las páginas web.
- Identificar la variedad de publicaciones por tipo, idioma y distribución geográfica.

Y para estos estudios se han utilizado medidas como las siguientes, según los mismos autores:

- Tamaño medio de los documentos.
- Protocolos utilizados por las URLs de los documentos HTML.

- Tipos de ficheros.
- Recuento de dominios científicos.
- Tipología documental de las páginas web.
- Recursos textuales o audiovisuales de cada página web.
- Número promedio de enlaces por página.
- Densidad media de enlaces.
- Tamaño informático.
- Densidad hipertextual.
- Densidad multimedia.
- Profundidad.

Con la intención de ejemplificar estas áreas y medidas de estudio, a continuación se mencionarán varios análisis cibermétricos de diferentes tipos o aplicados en diferentes ámbitos:

Rahimala y Marthandan (2010) realizaron un análisis cibermétrico de las 50 mejores compañías de pequeño tamaño mediante la exploración de los enlaces para determinar la asociación entre estos datos cibermétricos y los medidores de rendimiento y ranking de tráfico. Sus resultados sugirieron que los datos cibermétricos podrían ser indicadores para la evaluación del rendimiento de esas compañías, así como la visibilidad de sus webs para una audiencia determinada.

Orduña-Malea (2013) también comenta el potencial de la cibermetría para estudiar la división digital entre Estados Unidos y Europa, la detección de la influencia de patrones lingüísticos y culturales en las relaciones institucionales y la verificación del crecimiento de la visibilidad web académica en Asia. En cuanto al ámbito de las universidades, en 2013 una encuesta asegura que el 88% de los estudiantes preferían usar Google que el buscador interno de estas. Las características inherentes de la cibermetría le permiten utilizar indicadores que ayudan a estudiar qué quieren los usuarios, cómo acceden y qué piensan de los servicios ofrecidos, entre otros, si bien es cierto que también puede ayudarse con elementos como la altmetría y el análisis de sentimientos de los comentarios en las redes sociales.

Groselj (2014) aplicó la cibermetría a la información digital de salud a través del análisis de búsquedas, rastreo de webs y limpieza de datos, tras la cual categorizó las webs resultantes según el patrocinio, el tipo de plataforma y la estructura de enlaces.

En el ámbito del periodismo digital, el FIW, definido por Ingwersen (1998) y ya comentado en el subcapítulo 2.2.2.3, es una de las métricas más empleadas según Maina (2012). Se considera que una web es muy visitada si contiene información de interés para una audiencia objetivo, ofrece servicios de alta calidad y es mencionada por webs de mayor influencia. El Factor de Impacto Periodístico (FIP) mide las citas en un intervalo determinado de tiempo (2 o 5 años), y otros indicadores que también podrían influenciar en la visibilidad y el grado de cibermetría son: el diseño, la velocidad y facilidad de encontrar información de la web, la estructura de esta y los enlaces. Pradhan (2019) hizo a su vez un análisis cibermétrico de webs de periódicos indios, utilizando como métricas los datos de Alexa, el tráfico, las páginas vistas, la velocidad de descarga, el porcentaje de rebote, la duración de la visita, la ratio de búsqueda y los porcentajes de visitantes de India y fuera de esta.

Herrero Gutiérrez et al. (2012) efectuaron un análisis cibermétrico de cinco revistas de comunicación de reciente aparición, utilizando como criterios el género, la nacionalidad, el grado académico y la institución de adscripción de los autores de cada publicación. Este análisis permitió estudiar hipótesis como que el número de artículos es bajo en revistas emergentes y la existencia de una igualdad de género.

En cuanto al análisis cibermétrico de la prensa digital española, Trillo Domínguez (2008) destacó la débil situación de estos ciberdiarios debido a la existencia de problemas y fallos que complican el acceso a la información. También apreció una aparición desigual y relativa del desarrollo hipertextual y multimedia con una media de 32,53 sobre 100. Sí que comprobó un uso creciente de enlaces internos y externos (siendo mayoría los primeros), el predominio de las páginas HTML y la presencia de enlaces erróneos tanto internos como externos.

2.2.2.5. Análisis de la competencia

La cibermetría también puede ser utilizada en el análisis de competencia, es decir, el análisis de las webs que compiten con la entidad que se desea estudiar. Una estrategia que se ha utilizado mucho en el pasado para realizar análisis de marketing de la competencia es, explica Lavinsky (2013), el denominado como SWOT (Fuerzas, Debilidades, Oportunidades, Amenazas), presente en la Figura 58. El principal problema de este tipo de análisis es que puede conducir a la parálisis debido a que no se pueden

evitar completamente ni las debilidades ni las amenazas.

Figura 58. Análisis SWOT según el origen y si ayuda o no a obtener el objetivo (Lavinsky, 2013).

Es por ello por lo que Lavinsky (2013) opina que es más recomendable realizar un análisis de marketing de la competencia más amplio gracias a un sistema de inteligencia competitiva basado en: la revisión y documentación de lo que ha funcionado en el pasado y la optimización y repetición de estas acciones; sistematizar aquello que funciona y crear procedimientos de negocio basándose en ello; e identificar los canales, demografías y tendencias que puedan sugerir acciones para encontrar oportunidades de acciones que lleven al éxito.

Nesterenko (2017), por su parte, incluye los siguientes apartados en un análisis de marketing de la competencia, recomendando además las herramientas Compete[11], Marketing Grader[12], Moz[13], Raven[14], Simply Measured[15], Market Samurai[16] y BuzzStream[17]:

1. Estrategia de negocio: comparar los segmentos de audiencia, el porcentaje de mercado, qué temáticas, tendencias y conceptos tienen como objetivo, qué

[11] https://www.compete.com/
[12] http://marketing.grader.com/
[13] http://moz.com/
[14] http://raventools.com/tasks/competitor-analysis/
[15] http://simplymeasured.com/free-social-media-tools
[16] http://www.marketsamurai.com/
[17] http://www.buzzstream.com/

canales de distribución usan y los precios, descuentos y especiales anuales.

2. Comparaciones de la web: número de páginas, navegabilidad, integración social y tiempo de carga.

3. Público objetivo: descubrir qué perspectiva han adoptado los competidores y comparar si es la misma que la nuestra.

4. Social media: plataformas, fecha, hora y frecuencia de publicación, crecimiento de seguidores, tipos de contenido publicados, control de la voz activa sobre la temática o el producto y tiempo de respuesta.

5. Marketing de contenidos: tipos de contenido publicados, frecuencia de publicación, calidad (medición de la edición, la investigación y la originalidad), relevancia y audiencia.

6. Email marketing: frecuencia de emails, contenido, optimización para móvil y si los emails entran en spam directamente.

7. SEO: ranking de palabras clave, tráfico de visitas, autoridad de dominio, mozrank...

Una manera de mejorar la ratio de éxito del marketing de contenidos es realizando un análisis de contenidos de la competencia antes de empezar a crear contenidos, propone Gratt (2012). Así, se puede aprender a trabajar con la audiencia antes de poner el lápiz en el papel. Este autor introduce un proceso que consta de los siguientes pasos:

1. Definir los objetivos, para así poder analizar la competencia en base a cómo los consiguen ellos.

2. Reunir una lista de los competidores, sin confundir los competidores en SEO y redes sociales con los competidores en negocio. Los competidores en SEO y redes sociales son cualquier sitio que consiga atención en las SERPs (*Search Engine Results Page*, las páginas de resultados de un buscador) y feeds sociales.

3. Reunir una lista de URLs de contenidos de los competidores para poder analizar su efectividad a la hora de obtener enlaces y comparticiones. Posibles herramientas: Xenu[18], Screaming Frog[19], *sitemaps* en formato XML de los competidores…

[18] http://home.snafu.de/tilman/xenulink.html
[19] https://www.screamingfrog.co.uk/

4. Analizar la popularidad en enlaces y redes sociales de esas URLs con herramientas como Open Site Explorer[20], Majestic SEO[21], SharedCount[22], Social Crawlytics[23], AudienceWise[24]…

5. Escalar los datos a comparaciones multisitio. Cuando se obtienen los datos de popularidad de competidores individuales, es necesario hacer una comparativa entre diferentes webs para analizar tendencias (tipos de contenidos que tienen una gran popularidad a través de una variedad de webs de los competidores, representando nichos de demanda de ese tipo de contenido). Se crea una "*Relative Popularity Metric*" para cada web. Se elige el contenido más compartido de cada canal y se escala cada artículo como porcentaje de compartición con respecto a ese contenido. Sin embargo, para ello es necesario contar con una actividad mínima, ya que, si no, no se podrá separar la señal del ruido.

6. Combinar los datos para sacar conclusiones. Estudiar qué tipos de contenido son más populares, qué tono es el más popular, qué número de palabras, qué tipo de titulares, qué temáticas, qué mecanismos utilizan para provocar credibilidad, emoción, persuasión… y quién comparte esos contenidos (herramienta: Google+ Ripple Tool de AJ Kohn[25]).

7. Ser diferente, pero utilizando elementos de lo que es más popular divididos en tres grupos: motivación para compartir/enlazar, forma y temática. De esa manera se pueden mezclar los elementos de los grupos para su posterior combinación (Gratt, 2012).

Rayson (2014) introduce cinco pasos fundamentales para estudiar el éxito de contenido de la competencia, con el objetivo de conseguir inteligencia de contenidos y percepción de su estrategia:

1. Crear alertas de contenido de los competidores. Consiste en crear alertas vía email de las menciones a los competidores en otras webs. También alertas que se produzcan cada vez que se publican nuevos contenidos en las webs de los competidores, así como alertas que avisen de webs que han enlazado a los competidores. Esto último sirve para descubrir nuevas oportunidades de conseguir

[20] https://moz.com/link-explorer
[21] https://es.majestic.com/
[22] http://sharedcount.com/
[23] http://socialcrawlytics.com/
[24] http://audiencewise.com/
[25] http://www.blindfiveyearold.com/ripples-bookmarklet

enlaces hacia el proyecto.

2. Resumen del rendimiento de contenidos de los competidores, en el que se puedan ver los últimos artículos publicados y cuánta actividad han provocado en las redes sociales. De esta manera se puede averiguar qué tipo de artículos han sido los más exitosos, su cantidad de palabras, el día en que se publicaron…

3. Comparación entre los contenidos propios y los de los competidores. Consiste en un reporte de comparación a nivel de dominio de las medidas de compartición en redes sociales, el ritmo de publicación, el día en que se publica, la comparativa entre cantidad de palabras…

4. Revisar los contenidos más compartidos de los competidores. Sería necesario elaborar un reporte en el que ver una lista de los artículos más compartidos, lo que permite revisar sus titulares, en qué redes sociales han sido más vitales…

5. Revisar quién comparte sus contenidos. Estudiar quiénes son sus seguidores más fieles, cuántos seguidores tienen, su autoridad de dominio, su ratio de respuesta, su número medio de retuits… incluso se propone crear una lista de usuarios que comparten los contenidos de los competidores para intentar crear una relación con ellos e identificar qué tipo de contenido comparten.

La comparación de los contenidos con los de los competidores es una parte muy importante a la hora de desarrollar técnicas de marketing entrante o *inbound marketing*, ya que según Wainwright (2012) puede ayudar a optimizar la estrategia de contenidos de manera que se pueda competir e incluso superar a los competidores de un medio de comunicación. Este autor aporta la siguiente guía de pasos para dicho análisis:

1. Encontrar los contenidos de la página web.

2. Realizar una auditoría de contenidos: cantidad de cada tipo de contenido publicado, frecuencia de publicación y distribución de las temáticas. Con todo ello, el propósito es descubrir los puntos fuertes y débiles de los competidores.

3. Evaluar la calidad del contenido, basándose en la rigurosidad, ortografía, profundidad, tono, número de palabras, estructura, autores y la accesibilidad en cuanto a qué es necesario hacer para poder consumir los contenidos. También se puede analizar con la interacción de los lectores con el contenido: cómo interactúan en las redes sociales, qué temáticas tienen más éxito, el grado de fidelidad de los lectores... También es recomendable evaluar la estructura del blog: localización y

existencia de botones de seguir y compartir en redes sociales, categorización de los contenidos, biografías, diseño...

4. Establecer sus objetivos SEO: evaluar cómo se usan las palabras clave en los contenidos: título de página, arquitectura de URL, título, encabezados, densidad de palabra clave, textos alternativos de imágenes y uso de enlaces internos.

5. Investigar la integración de social media con la estrategia de contenidos: en qué redes sociales están presentes, qué cantidad de fans/seguidores, frecuencia de publicaciones, tipos de contenidos publicados y frecuencia de interacción de los fans con esos contenidos.

6. Aplicar la inteligencia competitiva adquirida, con el objetivo de sobrepasar a los competidores en ciertos aspectos si los recursos lo permiten, pero viendo todo el conjunto como un todo.

2.2.2.6. Altmetría

La cibermetría también se ha expandido desde el análisis de webs generales o académicas a las investigaciones de webs sociales, a través del rastreo o peticiones de datos a través de rutas permitidas (API). De esta manera se ha podido investigar patrones de amistad y uso del lenguaje, así como análisis de sentimiento hacia grandes eventos como ha sido en el caso de Twitter, o factores asociados a los comentarios de los vídeos de Youtube[26]. De esta manera, la cibermetría ha ido evolucionando desde un planteamiento más teórico a uno más aplicado, pudiendo acceder a ciencias sociales más amplias (Thelwall, 2012). Adoptando este enfoque complementario, ya que la cibermetría estudia lo que ocurre más allá de la web de la compañía, es posible analizar el impacto y la visibilidad de esta en el ciberespacio y, por ello, reflejando el estatus en la realidad (Orduña-Malea & Alonso-Arroyo, 2018, p. 12).

Nace así la altmetría, es decir, la ciencia que proviene de la cibermetría y que cuantifica la información existente en las redes sociales. La cibermetría no dispone de los parámetros suficientes para la contabilización de la información de las redes sociales, ya que tanto esta como la bibliometría se basan en citas, según el estudio de Priem et al. (2010) sobre las citas de los investigadores en Twitter. La altmetría, en cambio, puede medir las conversaciones y los marcadores a corto plazo, de manera que se puedan estudiar en menor tiempo. La altmetría refleja el impacto del artículo en sí mismo, expandiendo el

[26] https://www.youtube.com/

estudio del impacto en sí hacia aquello que está produciendo el impacto. Además, estudia el impacto fuera de la academia, es decir, el impacto de trabajos influyentes, pero no citados, así como el impacto de fuentes que no han sido revisadas por pares. Los sistemas altmétricos podrían, así, ser más robustos, diversos y óptimos a la hora de aprovechar el poder estadístico del *big data* algorítmicamente, por ejemplo, en la detección y corrección de actividad fraudulenta. Su velocidad, además, presenta la oportunidad de crear recomendaciones en tiempo real y sistemas de filtrado colaborativos.

Las redes sociales permitieron a los científicos manifestarse de tres maneras, descritas por Manzano Zambruno et al. (2019, pp. 208-209):

- Compartiendo sus investigaciones, ya sea en fases iniciales o finales.
- Compartiendo recursos útiles como pueden ser enlaces, referencias bibliográficas, etc.
- Compartiendo los resultados de sus análisis gracias a la naturaleza abierta de las redes sociales o incluso los blogs, revistas de acceso libre, etc.

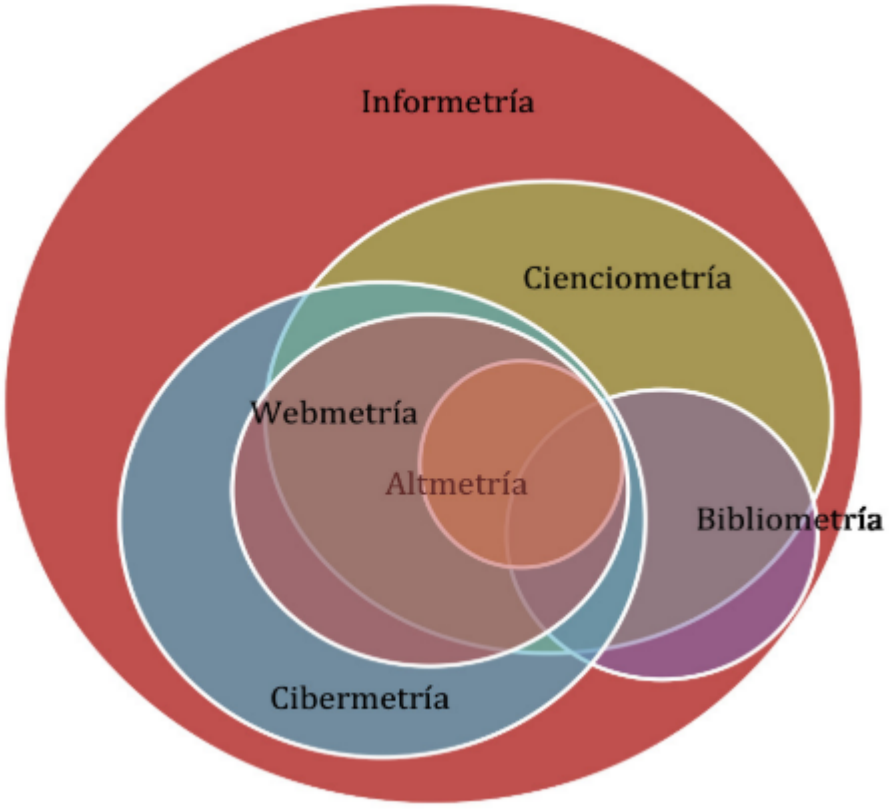


Figura 59. Las relaciones entre los campos de investigación de la informetría, bibliometría, cienciometría, cibermetría, webmetría y, también, altmetría (Manzano Zambruno, et al., 2019, p. 208).

Costas (2017) ha observado que el uso de estas plataformas para la difusión y promoción de las investigaciones ha ido en aumento con el transcurso del tiempo, ya no solo por parte

de los investigadores, sino también por las universidades, bibliotecas académicas, sociedades científicas y editoriales.

Con todo ello, para Bornmann (2014), una de las ventajas de la altmetría es su capacidad de medición de una mayor amplitud de impacto de la academia al cubrir entornos más allá de la ciencia en sí. Esto lleva a fuentes de datos más diversas y una evaluación de productos más allá de las publicaciones científicas, lo que proporciona una recolección de datos de impacto de estas en tan solo días o semanas tras su publicación. Además, al expandir la escucha a un público no especializado, se puede medir también el impacto a nivel social, así como en otros ámbitos de estudio. Sin embargo, según Bornmann (2014), también presenta algunas desventajas, como son:

- La comercialización de las redes sociales.
- La calidad de los datos en cuanto a quiénes las usan.
- Las diferentes versiones de las publicaciones.
- Las intenciones de estas citas sociales (desde menciones hasta debates extensos).
- Los estándares de medición y de menciones.
- La normalización de las menciones.
- La replicación de los datos, ya que los proveedores de éstos cambian o se vuelven obsoletos rápidamente.
- La falta de evidencia debido a los pocos estudios empíricos sofisticados en este campo.
- La posible manipulación de los datos gracias a cuentas falsas o la publicación automatizada de tuits.

Es por ello por lo que otros autores como Wouters et al. (2019) señalan que las métricas de redes sociales solo deberían ser usadas en conjunción con una revisión personalizada e informada y unos indicadores normalizados y avanzados, buscando transparencia y accesibilidad abierta, siempre teniendo en cuenta los efectos sistémicos de cada red social.

El análisis de menciones asume, pues, el papel del análisis de citas, y una técnica englobada en la altmetría. Consiste en la extracción de términos o frases y la evaluación de su presencia de manera cuantitativa (Aguillo, 2012). Estas menciones o citas también podrían ser en modo de enlaces, aunque todas ellas aprovechando el ámbito conversacional y la rapidez de acción y divulgación en tiempo real de redes como Twitter. Priem y Costello (2010) estudiaron este fenómeno en el ámbito de la academia y comprobaron que la mitad de las citas académicas de Twitter de su muestra era producida hacia enlaces de intermediarios que, a su vez, enlazan o incluyen a las fuentes originales.

Sin embargo, pese a sus características diferentes, representan y transmiten un impacto académico. Haustein (2019) estudió asimismo la comunicación de la academia en Twitter y comprobó que la mayoría de los tuits que hablaban de artículos científicos eran en su mayoría de miembros de la academia y no del público general, por lo que reflejan la comunicación académica más que el impacto social de los mismos. La mayoría de los tuits que enlazaban a artículos científicos se publicaban en las fechas posteriores a su publicación, por lo que la conversación podría medirse en horas más que en días, provocando un engagement muy bajo de usuarios enlazándolos en Twitter. Haustein (2019) argumenta que el límite de caracteres podría influir en los debates sobre los mismos, ya que no fomentan un debate intenso y pormenorizado sobre un tema de carácter científico.

2.2.3. Analítica en Twitter

Esta tesis se sirve de la red social Twitter para medir algunos parámetros de éxito, ya procedan de la propia cuenta del medio o del análisis de tendencias, por lo que se verá a continuación el estado de la investigación relacionada con Twitter, desde el análisis de la actividad, el análisis de la autoridad, los hashtags y el análisis de sentimientos.

2.2.3.1. Minería de datos en Twitter

En cuanto a la minería de datos de Twitter, se han producido dos ramas principales: la minería de grafos basada en el análisis de enlaces entre los mensajes, y la minería de textos basada en el análisis del texto en sí.

El análisis de redes en el entorno de una red social, en la que los nodos son individuos u organizaciones conectadas entre ellas a través de relaciones dentro de la propia red social, es una técnica englobada dentro de la altmetría y que Harinarayana (2015) llamó netmetría (*nettometrics*). Esta técnica se ha convertido en una metodología clave en el ámbito de la investigación sociológica y también es utilizada en otros muchos ámbitos, como la antropología, biología, estudios de la comunicación, etc. La minería de grafos se ha utilizado, según explican Bifet y Frank (2010), a la hora de: medir la influencia y la popularidad de los usuarios a través de los seguidores, los retuits y las menciones, ya que los enlaces directos crean un flujo de información y, asimismo, un flujo de influencia; el descubrimiento y formación de comunidades; y la difusión de información social a través de estrategias de muestreo de datos. La minería de textos, en cambio, se han volcado en: el análisis de sentimientos; la categorización de tuits; la agrupación de tuits; y el análisis y detección de tendencias.

Las utilidades de los datos de Twitter son muy numerosas, desde el seguimiento de la actividad política, la detección de terremotos y brotes de gripe y la detección de falsas alarmas al no suscitar tuits sobre ellas hasta cuestiones sociológicas de muchos tipos, ya que analizar los tuits es mucho menos costoso en tiempo y dinero que realizar encuestas, asegura Savage (2011). Filippo Menczer, director asociado del Center for Complex Networks and Systems Research de la Universidad de Indiana, comentó que “las redes sociales nos dan una oportunidad que no teníamos hasta ahora de seguir qué dicen todos, sobre todo. Es increíble.” Noah Smith, profesor adjunto de informática en la Universidad Carnegie Mellon explicó que “cuantos más datos tenemos, más cerca está la verdadera representación de lo que una población subyacente es. [...] Conforme Twitter y otras redes sociales crezcan, seremos más capaces de realizar preguntas más concretas. Si imaginas

a la sociedad como un gran organismo, esta es simplemente otra herramienta para mirar dentro de él."

La analítica de redes sociales es un proceso de tres pasos, descritos por Fan y Gordon (2014):

- Capturar en cuanto a que se obtienen datos relevantes de las redes sociales gracias a la monitorización o escucha de sus fuentes. Se puede hacer mediante feeds de noticias, APIs y rastreos. Tras ello es necesario modelar los datos, capturar los enlaces, capturar los hashtags o etiquetas, extraer las características y otras semánticas y sintácticas operaciones.

- Entender consiste en realizar una selección y archivo de la información más relevante, así como eliminar los datos de baja calidad que producen ruido en el análisis. Se trata, pues, de generar métricas útiles para la toma de decisiones, interpretando los datos y proveyendo de información sobre el sentimiento y el comportamiento de los usuarios. Es el apartado central de la analítica en redes sociales.

- Presentar consta de mostrar los resultados y su interpretación de una manera comprensible. Resumir, evaluar y aportar al usuario la información en un formato que sea fácil de entender, como es el caso de los paneles visuales, algunos de ellos con la capacidad de vistas personalizadas.

2.2.3.2. Análisis de la actividad

A la hora de analizar la actividad en Twitter, es importante estudiar por qué y cómo utilizan los usuarios esta herramienta. La intención de uso de éstos podría categorizarse en: conversación diaria, conversaciones, intercambio de información y estructura de comunidad. Bajo esta intención, los usuarios adquieren roles de fuente de información, amigo o buscador de información, siendo activo en el transcurso de una semana si ha publicado al menos un tuit a lo largo de esta y siendo conservado si retuitea al menos una vez en las siguientes X semanas. En un estudio realizado por Java et al. (2009) se comprobó que aproximadamente la mitad de los usuarios eran activos, y la mitad de éstos habían retuiteado en la siguiente semana. En la Figura 60 se puede observar la actividad y retención de los usuarios analizados en este estudio:

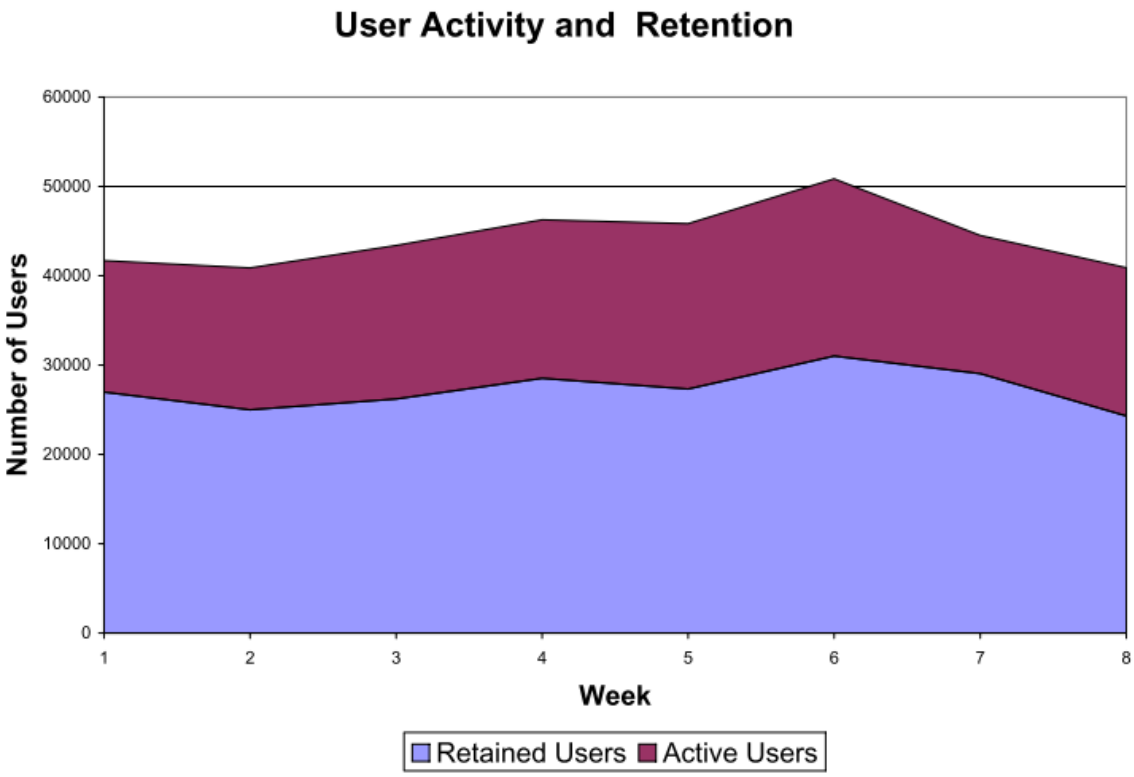


Figura 60. Usuarios activos y conservados (Java, et al., 2009).

Mathews et al. (2017) modelaron el tiempo que se tarda en retuitear utilizando la distribución de una ley de poder con corte exponencial, que demostró ser mejor que una ley de poder estándar. Así, analizaron el tiempo que tarda un usuario en retuitear desde el momento en el que abre la red social a partir de los 100 usuarios con más seguidores de Twitter.

A la hora de calcular la intención del usuario, Java et al. (2009) eligieron un enfoque que combina el algoritmo HITS y la detección de comunidad. Para ello, adaptaron el algoritmo HITS para detectar concentradores y autoridades en Twitter, siendo el valor de autoridad la suma de los valores escalados de los concentradores (*hubs*) de sus seguidores y el valor de concentrador es la suma de los valores escalados de autoridad de los que sigue:

$$Authority(p) = \sum_{v \in S, v \to p} Hub(v)$$

$$Hub(p) = \sum_{u \in S, p \to u} Authority(u)$$

Figura 61. Fórmulas de autoridad y concentrador (hub) (Java, et al., 2009).

En la anterior fórmula, Hub(p) es el valor de *hub* de una página p y Authority(p) es el valor de autoridad de una página p.

Por otro lado, la detección de comunidades se hizo a través de las relaciones de amistad considerando únicamente los enlaces bidireccionales entre dos usuarios. De esta manera, una comunidad sería una red de nodos más densamente conectados entre ellos que hacia otros fuera del grupo. Estas comunidades pueden solaparse y un usuario podría adquirir roles diferentes en cada comunidad. Para detectar el solapamiento de comunidades y la densidad de estas, se recurrió al Clique Percolation Method (CPM), gracias al cual se observaba aquellos miembros de la comunidad que están enlazados a muchos otros miembros de esta, pero no necesariamente a todos. Las comunidades k-clique se identifican por las uniones de todas las k-cliques que puedan alcanzarse entre ellas a través de k-cliques adyacentes, esto eso, que compartan k-1 nodos.

El análisis de la actividad es necesario ya que incluso en los primeros años de Twitter ya se apreció un gran número de sus cuentas como inactivas o con un bajo índice de participación en la conversación, según Anger y Kittl (2011). Cuando un usuario es influyente, lo es porque muchos están de acuerdo con su contenido; porque tiene un estatus que es bien acogido, respetado y/o de celebridad; o porque está relacionado con una creencia o comportamiento aceptado en público y en privado, es decir, a nivel de contenido y conversación. Lo contrario de dicha influencia se notaría en una reacción basada en ignorar su contenido o en provocar rechazo. En un estudio de Kwak et al. (2010) se observó que un retuit alcanza a un promedio de 1000 usuarios, sin importar el número de seguidores de la cuenta que publicó el tuit original. Es por ello por lo que para estudiar la influencia sería conveniente averiguar qué usuarios retuitearán los mensajes. Luo et al. (2013) propusieron una metodología que usaba el histórico de retuits, el estatus de seguidores, el tiempo de actividad e intereses de éstos. Demostraron así que los parámetros que podrían ser efectivos para ello serían el histórico de retuits del autor y sus

seguidores, el estatus de éstos y la similitud entre el contenido del tuit en concreto y los últimos tuits publicados por los seguidores

A la hora de analizar mediante grafos los retuits, Kumat et al. (2014, p. 35) utilizan un enfoque en el que se trata de obtener los usuarios que formen la centralidad mediante grado, es decir el número de usuarios que han retuiteado un tuit, así como qué usuarios son más influyentes (centralidad de vector propio) y cuáles presentan un camino más corto hacia la información (centralidad del camino más corto). En la primera visión, se trata de averiguar qué usuarios son más retuiteados, por lo que, en el siguiente ejemplo de la Figura 62, sería Alice:

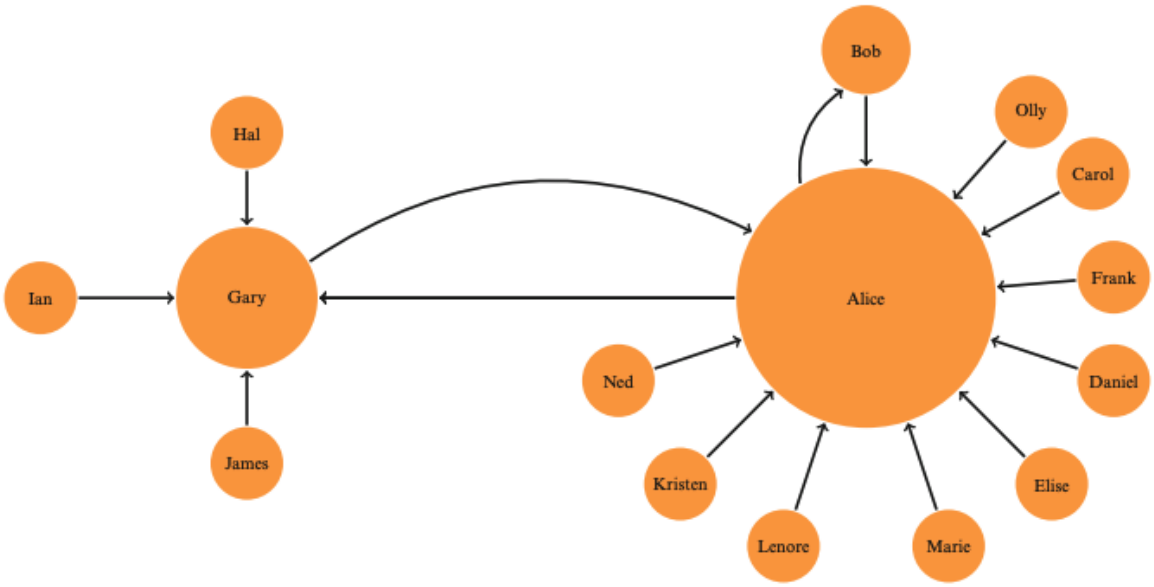


Figura 62. Alice es el usuario con más retuits, por lo que es la que tiene una centralidad de grado mayor (Kumar, et al., 2014, p. 35).

Sin embargo, Bob y Gary son los nodos de los cuales obtiene Alice la información, por lo que Bob y Gary son los nodos más influyentes, como se ve en la Figura 63:

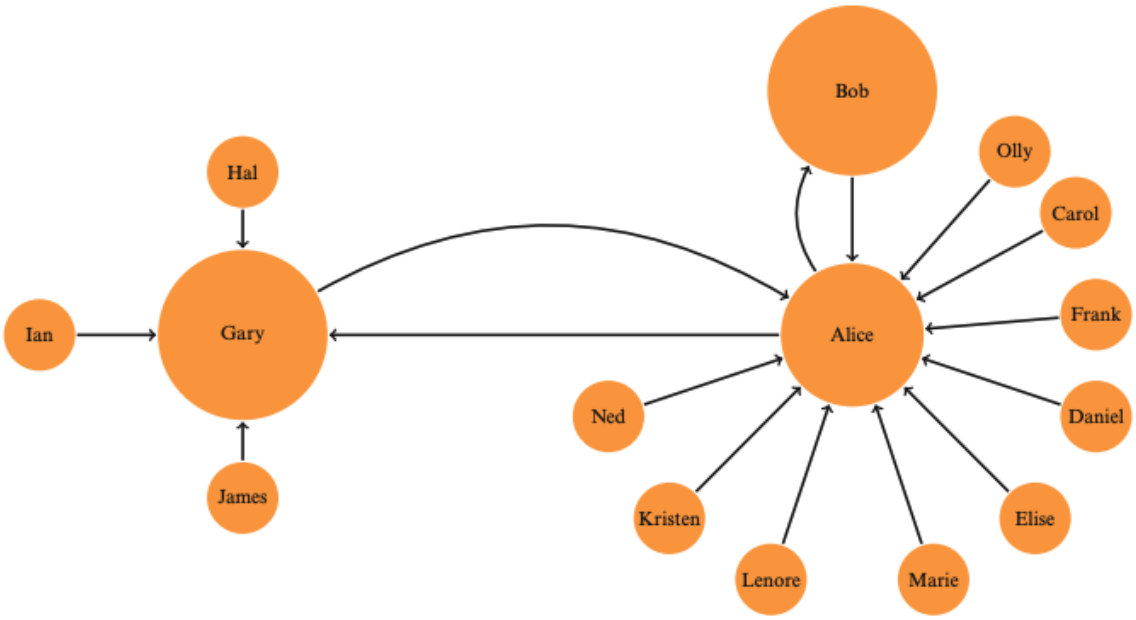


Figura 63. Bob y Gary son los nodos que aportan la información más retuiteada, por lo que son los de mayor centralidad de vector propio (Kumar, et al., 2014).

En cambio, Alice es la que está haciendo llegar la información a más usuarios en el camino más corto y por tanto directo, destacándose en la Figura 64:

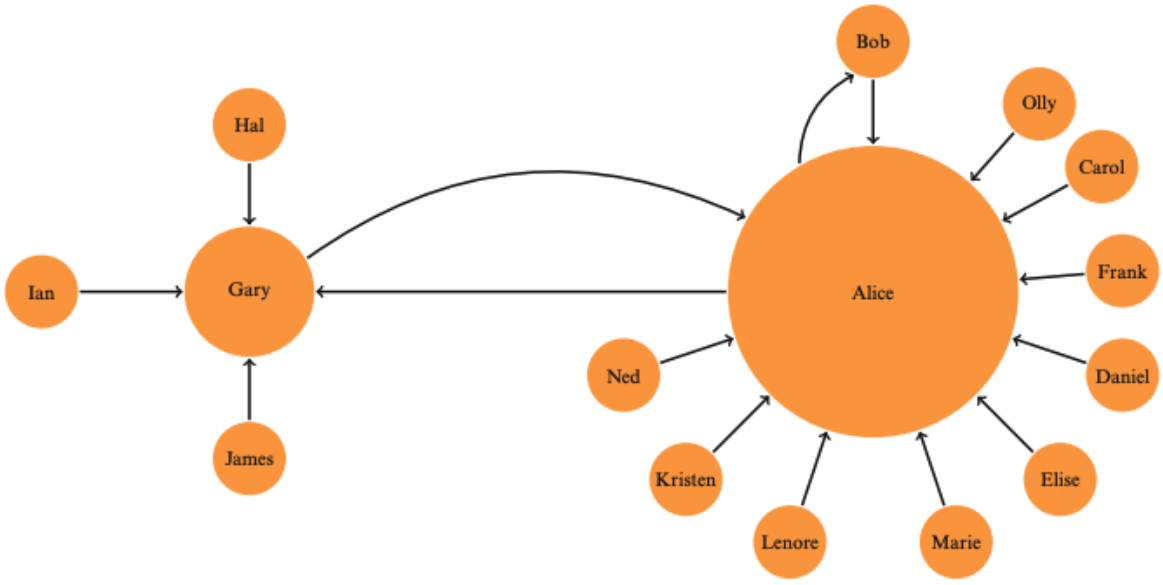


Figura 64. Alice es la que aporta la información a más usuarios en un camino más corto, por lo que tiene la mayor centralidad del camino más corto (Kumar, et al., 2014, p. 35).

2.2.3.3. Análisis de la autoridad

Hay algunas métricas que, aunque por definición deberían estar relacionadas con la autoridad, sin embargo, en la práctica pueden ser engañosas. Si la influencia es el poder o capacidad de causar un efecto indirecta o intangiblemente, en el ámbito de Twitter se da

con la conexión entre los usuarios, y esta puede ser producida por motivos muy variados, desde una amistad íntima hasta intereses comunes. Los usuarios populares con un gran número de seguidores no tienen por qué suscitar muchos retuits y menciones. Además, pueden ser influyentes en varios temas a la vez, aunque no de manera espontánea, sino gracias a la especialización sobre ciertas temáticas en sus tuits (aunque pudiendo ser varias y no solo una) y con contenido creativo y de calidad que pueda considerarse de valor para los demás usuarios.

El número de seguidores representa la popularidad de un usuario y, según un estudio de Cha et al. (2010), no está necesariamente relacionado con los retuits y menciones. Esto se debe a que los retuits representan el valor del contenido de los tuits y las menciones el valor del nombre del usuario. Otros autores como Suh et al. (2010) difieren, asegurando que tanto el número de seguidores como de usuarios a los que se sigue y la edad de la cuenta afectan al número de retuits, mientras que la métrica que no le afecta es el número de tuits publicados anteriormente. Otras métricas relacionadas con el número de retuits son la inclusión de una URL en el tuit, así como el uso de hashtags.

Si bien es cierto que las tendencias suelen ser creadas por usuarios que ya han sido influyentes en el pasado y con un gran número de seguidores, es complicado predecir si una URL o un usuario en concreto crearán tendencia, por lo que Bakshy et al. (2011) llegaron a la conclusión de que la única manera de captar el boca a oreja sería a través de grandes cantidades de usuarios potencialmente influyentes con la intención de conseguir una media de sus efectos.

Yamaguchi et al. (2010) proponen otro enfoque a la hora de calcular la autoridad de un usuario, ya no a través de sus relaciones con otros usuarios, sino del análisis de enlaces de su contenido. Si Twitter es considerado como un tipo de fuente de información, un usuario de autoridad enviará frecuentemente información útil que se compartirá rápida y ampliamente. Para ello, aportan un grafo usuario-tuit que valora las diferentes conexiones entre uno y otro con diferentes pesos, de manera que se pueda calcular el peso de todas las interacciones entre los usuarios en una temática concreta.

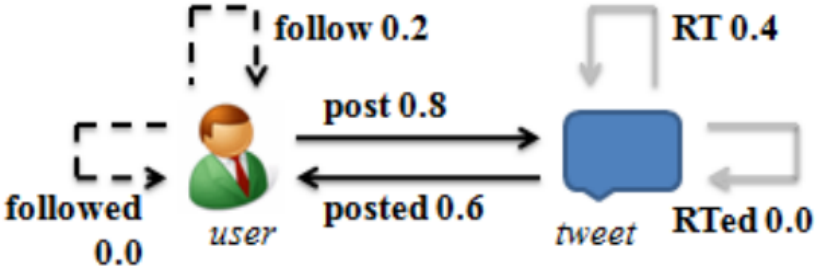


Figura 65. Grafo Usuario-Tuit (Yamaguchi, et al., 2010).

También puede surgir la necesidad de elegir a qué usuarios seguir para estar informados de las temáticas en las que se desee participar y las comunidades de las que se desee formar parte. Para ello, Hannon et al. (2010) describen como necesario atender a métricas como la descripción del usuario, su actividad reciente en Twitter y su grafo social, por lo que han centrado sus esfuerzos en estudiar la recomendación automatizada ya no de potenciales seguidores, sino de usuarios a los que seguir, como es el caso de Twittomender. Este tipo de nociones han ayudado a la implementación de algoritmos, arquitecturas y marcos de evaluación para herramientas como Who to Follow, que ofrecía un servicio de recomendaciones a gran escala (Gupta, et al., 2013).

Otro enfoque de aplicación de machine learning o aprendizaje automático es el de la clasificación de usuarios según atributos como la orientación política, la etnicidad o la afinidad por una compañía en particular. Para ello, Pennacchiotti y Popescu (2011) utilizaron el perfil del usuario, el comportamiento de su publicación, el contenido lingüístico de los mensajes y las características de su red social. Comprobaron que las características del contenido pueden tener un gran valor si se toman en un modelo a gran escala sobre una temática determinada, por lo que podrían ayudar en la clasificación de usuarios incluso en otras temáticas que las anteriormente mencionadas como ejemplos.

Siempre hay que tener en cuenta para estos tipos de análisis los límites de la propia API de Twitter, que especialmente desde su versión 1.1 aumentó todavía más los límites de ejecución de sus funciones. Es uno de los problemas con los que se encontraron Zheng et al. (2013) a la hora de desarrollar un rastreador para obtener perfiles, seguidores y seguidos de usuarios de Twitter en 2013.

Por otro lado, no hay que olvidar que es posible obtener una falsa sensación de influencia a través de estrategias simples pero efectivas, consistentes en *bots* que producen interacciones automáticas o seguir usuarios desconocidos para llamar su atención y que les sigan a cambio, por lo que los sistemas de puntuación de influencia deberían tener en

cuenta la actividad automática (Messias, et al., 2013).

2.2.3.4. Procesamiento de hashtags y spam

La naturaleza conversacional también afecta a la hora de etiquetar los tuits con los llamados hashtags (#), ya que en el caso de Twitter los usuarios hacen un uso de ellos a modo de filtro, promoción y categorización de contenido. Según un estudio de Huang et al. (2010), el 0,001% de los hashtags pasa por un proceso de adopción en el que una gran cantidad de usuarios deciden utilizarlo, mientras que su abandono se produciría cuando una gran cantidad de usuarios deja de utilizarlo y el uso del hashtag se vuelve infrecuente.

Huang et al. (2010) procesaron los hashtags de la siguiente manera: para elegir solo los que se habían creado en el período estudiado, eliminaron aquellos que aparecieron más de 0,001% del tiempo en los primeros 10 días; para normalizar la frecuencia de cada hashtag por día, lo dividieron entre el número total de hashtags de ese día; para eliminar el spam, efectuaron una eliminación manual detectando la compartición repetida de una misma URL.

Para calcular la desviación estándar (una medida de dispersión que se suele utilizar en la estadística descriptiva) de las marcas de tiempo de cada hashtag, midiendo así la propagación de la actividad de este y por tanto el tiempo que estuvo en uso, utilizaron la siguiente fórmula:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Figura 66. Fórmula de la desviación estándar de un hashtag (Huang, et al., 2010).

También calcularon el sesgo de las marcas de tiempo de los hashtags, es decir, la rapidez con que crece y decrece su uso, gracias a la siguiente fórmula:

$$g_1 = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{3/2}}}$$

Figura 67. Fórmula del sesgo de un hashtag (Huang, et al., 2010).

Y, a la hora de calcular el cuarto momento, curtosis (medida estadística del grado de

concentración de los valores de una variable alrededor de un valor central) para comprobar cuánto se mantuvo el hashtag en el pico de popularidad, utilizaron la siguiente fórmula:

$$g_2 = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2}} - 3$$

Figura 68. Fórmula de la curtosis de un hashtag (Huang, et al., 2010).

Gracias a la aplicación de estas fórmulas, Huang et al. (2010) observaron una relación entre las métricas estadísticas y la adopción de los hashtags, por lo que el estudio de sus patrones de uso podría ayudar a su clasificación automática.

Las temáticas o eventos a los que hacen referencia los hashtags, además, influyen en la forma en que los usuarios reaccionan ante éstos, según Bruns y Stieglitz (2012). Por ejemplo, una respuesta estándar ante las noticias de última hora es la de encontrar, compartir y volver a compartir información relevante, por lo que los tuits con URLs y retuits aumentan en número. En cambio, para eventos en tiempo real, los tuits suelen ser mayoritariamente comentarios de fuente propia y que no interactúan tanto con los tuits de otros usuarios ni incluyen enlaces con más información. Estudiar los papeles cambiantes de los participantes en las publicaciones de un hashtag podría servir para identificar fases en la discusión de este o formar grupos de participantes en torno a él. (Bruns, 2012)

Los temas con mayor tendencia (*trending topics* o TT) pueden ser vistos como oportunidades para generar tráfico e ingresos, por lo que muchos usuarios aprovechan este canal para crear contenido spam con los términos y hashtags más utilizados, a menudo utilizando acortadores de URLs. Todo ello puede devaluar el contenido y por tanto la imagen de Twitter como plataforma de contenido en tiempo real. Para detectarlos, Benevenuto et al. (2010) propusieron un enfoque de cuatro pasos para detectar usuarios que publican spam:

1. Realizaron un rastreo casi completo de Twitter, con 54 millones de usuarios, 1,9 billones de enlaces y 1,8 billones de tuits.

2. Crearon una colección de etiquetas para los usuarios manualmente clasificados como *spammers* o *no-spammers* según si sus tuits incluían palabras clave relacionadas con los temas de mayor tendencia y el punto de vista de hasta tres voluntarios: dos de ellos los clasificaban y, en caso de empate, un tercero rompía el empate con su voto.

3. Estudiaron las características del contenido de los tuits y el comportamiento de los usuarios de la colección anterior para entender la manera en que se discrimina y distingue el spam. Para las características del contenido de los tuits comprobaron el máximo, mínimo, media y mediana de atributos como: el número de hashtags por número de palabras de cada tuit, el número de URLs por palabra, el número de palabras por tuit, el número de caracteres por tuit, el número de URLs por tuit, el número de hashtags por tuit, el número de caracteres numéricos por tuit, el número de usuarios mencionados por tuit, el número de veces que un tuit es retuiteado, etc. Para los atributos del usuario, capturaron las métricas: número de seguidores, número de seguidos, fracción de seguidores por seguidos, número de tuits, edad de la cuenta, número de menciones del usuario, número de respuestas recibidas por el usuario, número de respuestas enviadas por el usuario, número de siguiendo de los seguidores del usuario, número de tuits recibidos de los usuarios que sigue, la aparición de palabras spam en el nombre de usuario y el mínimo, máximo, media y mediana del tiempo entre tuits, el número de tuits publicados por día y por semana.

4. Aplicaron un método de aprendizaje automático supervisado y comprobaron su fiabilidad, la cual resultó ser alta.

Este aprendizaje automático supervisado, teniendo en cuenta el contenido y el comportamiento de los usuarios, también se ha aplicado en otras ocasiones en la academia, como es el caso del algoritmo propuesto por Zheng et al. (2015) basándose en SVM para clasificar los *spammers*. El problema de este tipo de enfoques es que dependen de una selección manual basada en un análisis estadístico, y si se tiene en cuenta que Twitter funciona en un entorno en tiempo real, el procedimiento podría no ser suficientemente adaptativo ni rápido.

Según Wang (2010), las cuentas de usuarios que pueden ser consideradas spam suelen presentar algunas de las siguientes características:

- Seguir a una gran cantidad de usuarios para ganar su atención. La misma política de spam y abuso de Twitter advierte de que si tienes una pequeña cantidad de seguidores y sin embargo una gran cantidad de seguidos, puedes ser considerado spam. Por ello, parámetros como el número de amigos, el número de seguidores y la reputación de un usuario pueden ayudar en este cometido.

- Publicar contenido duplicado de otras cuentas, ya que no es una actitud común entre usuarios legítimos.

- Publicar tuits con enlaces maliciosos, normalmente mediante acortadores de URL.
- Uso abusivo de menciones y respuestas para ganar la atención de los demás usuarios.
- Publicar tuits que contienen TT (*trending topics* o tendencias) sin estar relacionados con ellos solo por ganar visibilidad de esta manera.

Hay diversas maneras de producir spam en Twitter, entre las que se cuentan los acortadores de URLs que en realidad dirigen a publicidad, vendedores de cuentas, programas de afiliados de spam... Este spam se distribuye a través de los tuits de los usuarios, los hashtags y palabras clave más usados, las búsquedas de temáticas populares y los mensajes directos de aquellos usuarios que siguen con sus cuentas. No obstante, Twitter presta atención a la hora de suspender dichas cuentas, ya que el 77% de las que detectaron Thomas, Grier y Paxson (2011) en un estudio fueron suspendidas en solo un día.

2.2.3.5. Social Media Analytics

En un estudio de Teevan et al. (2011) en el que se comparaba la búsqueda web con la búsqueda en Twitter, se comprobó que la intención de los usuarios a la hora de utilizar la herramienta de búsqueda dentro de Twitter es, principalmente, la de buscar información relevante temporalmente como noticias de última hora, contenido en tiempo real y tendencias populares; así como información relacionada con gente en concreto, como contenido dirigido al usuario que busca, información de otros usuarios de interés u opiniones en general. Las consultas de búsqueda en Twitter son más cortas, más dependientes de lo que es popular y evolucionan con menos probabilidad con el transcurso de la sesión del usuario que las consultas de búsqueda realizadas en una web. Esto según los autores es en parte porque la intención de los usuarios de Twitter a la hora de buscar es la de revisar los resultados que están relacionados con la consulta, mientras que en una web suelen buscar una mayor información sobre la consulta en sí. Todo ello está relacionado con el propio carácter del contenido en Twitter, más conversacional y en tiempo real que el proveniente de una web en el que son más habituales los hechos y el contenido navegacional.

Según Kaushik (2011) una de las claves de la analítica de redes sociales consiste en medir si se está participando en los canales de una manera óptima. Es decir, si se es capaz de captar la atención, de hacer disfrutar, de provocar ganas de compartir, de iniciar conversaciones, de realizar ciertas acciones… Para ello, se basa en cuatro métricas de

social media:

- Ratio de conversación, que equivale al número de comentarios o respuestas por artículo. Para ello, es necesario estudiar quién es la audiencia, cuáles son los atributos de la marca, en qué se es bueno, qué valor se puede ofrecer a los seguidores y cómo es el ecosistema en el que se participa.

- Ratio de amplificación. En Twitter es igual al número de retuits por tuit. En otras redes sociales, la amplificación se entiende como el número de comparticiones por entrada en Facebook, y el número de clics de botón de compartir por entrada o vídeo en un blog o en Youtube. Para ello, es necesario medir qué contenidos provocan amplificación (acceder al 2º o 3er nivel de la red: los seguidores de los seguidores) y crear más contenidos de ese tipo que provocará: que lo compartan más y, sobre todo, que lo valoren más, ya que lo consideran tan bueno que lo comparten con su propia audiencia. De esa manera, muchos seguidores de los seguidores se podrían convertir en seguidores de 1er nivel, si valoran los contenidos publicados.

- Ratio de elogio. En Twitter es el número de clics en favoritos por tuit. En otras redes sociales, el elogio se mide mediante el número de “me gusta” por entrada en Facebook o el número de +1 y “me gusta” por entrada o vídeo en un blog o en Youtube. De esta manera no solo se valora la visita sino también si esta ha considerado el contenido de la calidad suficiente para valorarlo positivamente.

- Valor económico, cuya definición es la suma a corto y largo plazo de las ganancias y costes, tal y como se observa en la Figura 69.

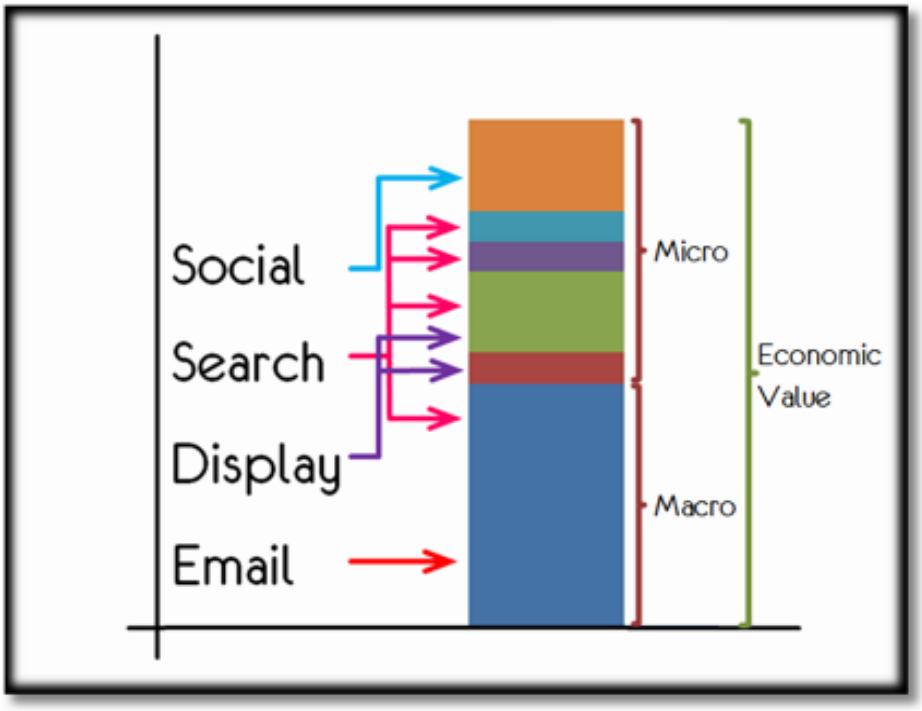


Figura 69. Valor económico en los diferentes canales (Kaushik, 2011).

En el caso de las redes sociales, se trataría de identificar y medir la flecha azul y del bloque naranja del gráfico anterior. Y, después, segmentar en las diferentes redes sociales en las que se está participando. Un ejemplo de la medición de todas estas métricas podría ser la siguiente gráfica de la Figura 70, creada a partir de las explicaciones anteriores por Erik Ohlen, citado por Kaushik (2011):

	Twitter	Facebook	Google+	Blog	Youtube
# Posts:	8	9	4	3	4
# Comments:	13	22	5	22	5
# Re-tweets/Shares:	14	19	7	3	3
# Favorite clicks/Likes/+1:	3	81	19	16	28
1. Conversation Rate # Comments per post:	1.63	2.44	1.25	7.33	1.25
2. Amplification Rate # Re-tweets/Shares per post:	1.75	2.11	1.75	1.00	0.75
3. Applause Rate # Favorite clicks/Likes/+1 per post:	0.38	9.00	4.75	5.33	7.00
Imported data from analytics tool 1 time/day					
	Twitter	Facebook	Google+	Blog	Youtube
4. Economic Value: Value per visitor:	0.13 $	3.72 $	1.22 $	1.43 $	0.74 $

Figura 70. Gráfica de mediciones de redes sociales según Ohlen (Kaushik, 2011).

En su momento, el servicio Klout proveía de una métrica Klout que se calculaba con la fórmula de la Figura 71:

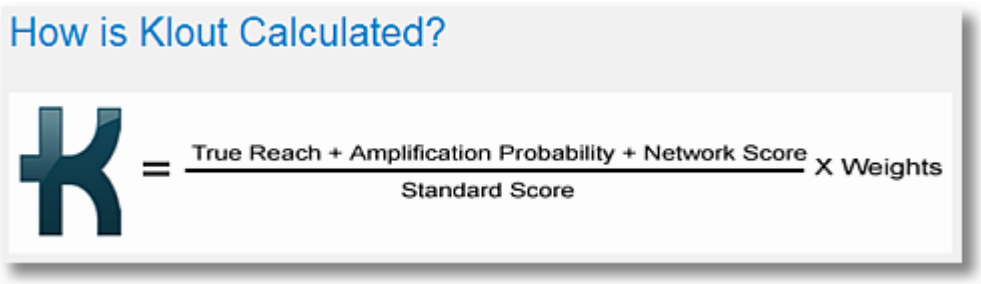


Figura 71. Fórmula de la métrica Klout (Kaushik, 2011).

Sin embargo, según Kaushik (2011), las fórmulas compuestas pueden ser subjetivas, no aplicables a muchos casos y esconder de manera eficiente información que es necesario comprender para las acciones a realizar. Es por ello por lo que destaca otras métricas también evaluadas por Klout, categorizadas en cuatro grupos: alcance, demanda, compromiso y rapidez, tal y como se resume en la Figura 72:

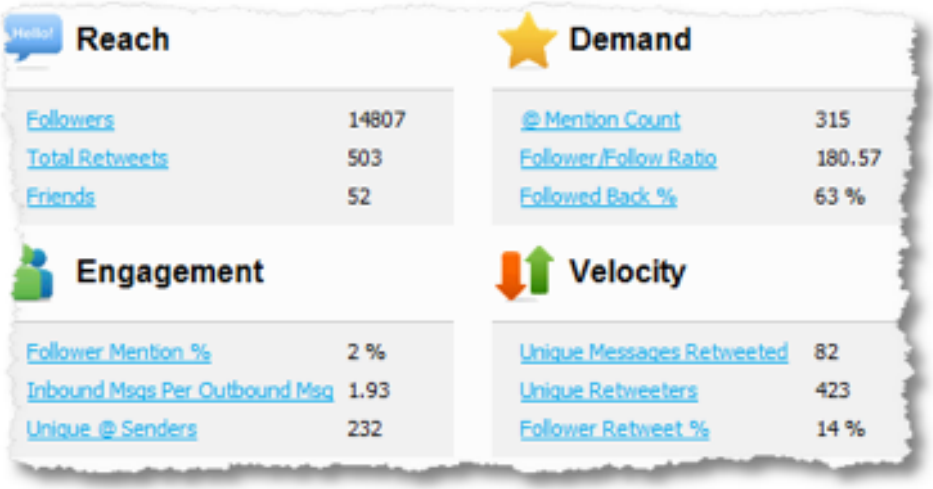


Figura 72. Métricas de redes sociales según Klout (Kaushik, 2011).

El alcance es la capacidad de atraer audiencia a los contenidos, calculada gracias a métricas como el número de seguidores, el número total de retuits y el número de amigos, según indica Kaushik (2009). La demanda consistiría en el número de menciones, la ratio de seguidores y usuarios a los que se sigue y el porcentaje de relaciones de seguimiento bidireccionales. El *engagement* o grado de compromiso consiste en comprobar en qué grado se contesta a los tuits y se participa en las conversaciones. Se puede medir, especifica Kaushik (2009), con el porcentaje de menciones de los seguidores, el número de mensajes por cada mensaje saliente y el número de usuarios únicos que realizan menciones. En cuanto a la rapidez, es necesario estudiar las probabilidades de que el

contenido publicado sea retuiteado y por quiénes, gracias a métricas como el número de mensajes únicos retuiteados, el número de usuarios únicos que han retuiteado mensajes, el porcentaje de seguidores que retuitean contenidos e incluso el número de retuits por cada mil seguidores.

El *engagement* también puede ser calculado, según otros medios como Social BlaBla (2020), como la cantidad de respuestas y retuits de cada tuit en un período de tiempo, dividida entre el número total de seguidores activos en el instante de publicación del tuit. Dicho número, multiplicado por 100, ofrece el porcentaje de *engagement* o vinculación de los seguidores de una cuenta.

Rodríguez Martínez (2013) aporta los siguientes cinco indicadores para medir las redes sociales: la visibilidad gracias al tráfico que recibe la web, el número de visitas, de seguidores, de suscriptores, de clics, de enlaces entrantes; la interacción mediante los tuits, retuits y “me gusta” recibidos por los usuarios; la influencia según quiénes y cuánto han compartido los contenidos, número de retuits y número de listas; la fidelización y el *engagement* gracias a la calidad de los tuits recibidos independientemente de si son positivos o negativos así como los enlaces compartidos, todo ello mediante un análisis manual; y la popularidad medida con el número de seguidores. Pallares (2014) señala los siguientes indicadores: tráfico que deriva a la web, número de seguidores, número de retuits, número de favoritos, número de menciones, Klout, promedio de seguidos y seguidores, usuarios que dejan de seguir, comunidades que generen *engagement* en forma de conexiones emocionales, y solicitudes de seguimiento en el caso de ser un perfil privado.

Alvarez Intriago et al. (2016) eligieron los indicadores de la Figura 73 aplicados a las redes sociales de Facebook y Twitter para analizar la estrategia de marketing digital de una escuela de postgrado de Ecuador:

Tabla # 1

Indicadores en las redes sociales Facebook y Twitter

Análisis		Indicadores	
		Facebook	Twitter
Interacciones		Comentarios	Menciones
		Compartir	Retweet
		Me gusta	Clics en enlaces
Audiencia		Fans	Seguidores
		Publicaciones	Tweets
Compromiso	Conversación	Comentarios	Menciones
		Fans	Seguidores
	Amplificación	Compartir	Retweet
		Fans	Seguidores
	Acción		Clic enlace
			Seguidores
	Aceptación	Me gusta	
		Fans	
	Interacciones	Comentarios+Compartir+Me gusta	Menciones+Retweet
		Fans	Seguidores
Interés	Conversación	Comentarios	Menciones
		Fans	Tweet
	Amplificación	Compartir	Retweet
		Publicaciones	Tweet
	Acción	Clic enlace	Clic enlace
		Publicaciones	Tweet
	Aceptación	Me gusta	
		Publicaciones	
	Interacciones	Comentarios+Compartir+Me gusta	Menciones+Retweet
		Publicaciones	Tweet

Figura 73. Tabla de indicadores (Alvarez Intriago, et al., 2016).

López et al. (2017) elaboraron, por su parte, una lista de indicadores de uso de Twitter y Facebook que se puede ver en la Figura 74, respondiendo a cuatro estrategias: audiencia o comunidad, compromiso, viralidad y conversión.

Estrategia	Objetivo	Indicadores Twitter[a]	Indicadores Facebook[a]
Audiencia o comunidad	Saber si aumentan nuestros seguidores	→ Seguidores *(followers)* → Nuevos seguidores → Han dejado de seguir *(unfollowers)* → Alcance potencial	→ Fans *(Me gusta)* → Nuevos fans → Alcance total → Alcance de la publicación
Compromiso	Conocer las acciones y reacciones (grado de compromiso) que generan nuestros contenidos en la comunidad	→ Favoritos → Respuestas → Mensajes directos → Clics a enlaces de nuestros contenidos → Listas	→ Visitas a la página fan → Publicaciones que "gustan" → Comentarios y sentimientos (positivo/negativo/neutro) → Clics a enlaces de nuestros contenidos → Mensajes privados → Impresiones
Viralidad	Saber si nuestros contenidos se comparten en la red	→ Tweets → Retweets → Impresiones → Menciones → Influencia (*Klout* o *Kred*)	→ Tasa de interacción
Conversión	Conseguir los objetivos fijados	→ Número de descargas → Número de nuevos registros en la web → Número de compras → Usuarios/as que dejan sus datos *(Lead)*	→ Número de descargas → Número de nuevos registros en la web → Número de compras → Usuarios/as que dejan sus datos *(Lead)*

Figura 74. Tabla de indicadores de uso según su estrategia en Twitter y Facebook (López, et al., 2017).

A la hora de proponer un universo de acciones desde el punto de vista de una cuenta de Twitter, Navas (2018) tuvo en cuenta tanto los tuits nativos como las menciones. Así, los tuits nativos se dividen en los nuevos tuits nativos, las respuestas nativas y los retuits nativos de la cuenta, y las menciones se dividen en menciones originales, respuestas a menciones y retuits a menciones. Esta clasificación se puede ver mejor en la Figura 75, mostrada a continuación:

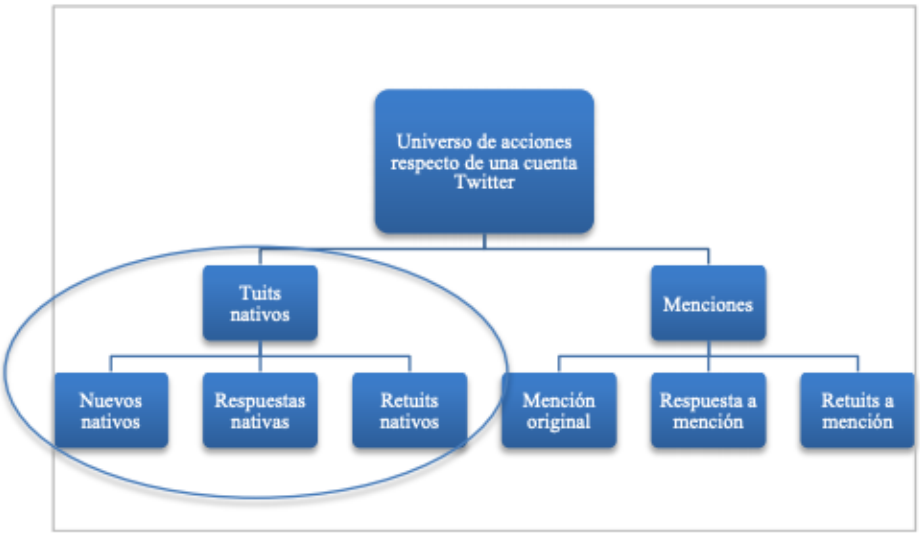


Figura 75. Universo de acciones respecto de una cuenta de Twitter, según Navas (2018)

Sarzosa Rivera y Medina Chicaiza (2017) hicieron una revisión teórica de los indicadores más utilizados en Facebook, Twitter e Instagram, y propusieron para medirlas: Piwik (ahora conocido como Matomo)[27], Google Analytics[28], Tweet Binder[29], Followerwonk[30],

[27] https://matomo.org/
[28] https://analytics.google.com/
[29] https://www.tweetbinder.com/
[30] https://followerwonk.com/

Audiense[31], Improvely[32], Clicktale (ahora conocido como Contentsquare)[33] y 4Q iPerceptions[34].

2.2.3.6. Análisis de sentimientos

Crawford (2009) menciona el término de "tener una voz" a la hora de estudiar la experiencia de la presencia online, ya sea en blogs, wikis, redes sociales o fotos, haciendo uso de otros términos relacionados como el de "escuchar" en las redes sociales, que clasificó luego en una escucha de fondo, recíproca o delegada.

Esa escucha da la posibilidad de detectar rápida y efectivamente la perspectiva de los clientes hacia lo que es importante para el éxito en el mercado (Sarlan, et al., 2014). O incluso modular la comunicación de manera que suscite sentimientos y sea más proclive a ser compartida. Por otro lado, en opinión de Stieglitz y Dang-Xuan (2013), el tono del autor actúa como un dato adicional al del contenido y el uso intencional de sus publicaciones en una plataforma de microblogging como es Twitter, en la que su flexibilidad y efímera naturaleza juegan a su favor.

Bollen et al. (2009) hicieron un análisis de sentimientos de todos los tuits publicados entre el 1 de agosto y el 20 de diciembre de 2008 y clasificaron el tono en una de las seis dimensiones siguientes: tensión, depresión, enfado, vigor, fatiga y confusión. De esta manera comprobaron que un análisis de sentimientos de gran escala podría ayudar a modelar tendencias emocionales colectivas y tratar de predecirlas según indicadores sociales y económicos. Se trató de uno de los primeros estudios a gran escala de este tipo en Twitter, y constó de los siguientes procesos:

1. Compilación manual de eventos en el período analizado.
2. Compilación de todos los tuits publicados en Twitter en el período analizado.
3. Procesamiento de los tuits: separación de términos individuales por palabras; eliminación de caracteres no alfanuméricos; conversión de los caracteres resultantes a minúscula; eliminación de las apariciones de 214 palabras conectoras y verbos muy comunes; y búsqueda según el algoritmo de Porter de los términos finales, asegurando gracias a este que la forma de las palabras (género, número o persona) no penalice a su frecuencia.

[31] https://es.audiense.com/
[32] https://www.improvely.com/
[33] https://contentsquare.com/
[34] https://4q.iperceptions.com/products/astute-social

4. Selección de tuits que tuvieran una expresión explícita de sentimiento individual, es decir, aquellos que incluyeran expresiones regulares como “sentir”, “soy”, “siendo”, “ser”.

5. Filtro de spam, eliminando tuits con expresiones regulares como “http:” o “www”.

Go, Bhayani y Huang (2009) analizaron los tuits dividiendo el sentimiento según si era positivo o negativo, con la limitación de que decidieron no tener en cuenta los emoticonos en su análisis debido a que los consideraban unas etiquetas ruidosas que podrían provocar un análisis erróneo del tuit. En este caso, el post procesamiento consistió en los siguientes pasos:

1. Eliminación de emoticonos.

2. Eliminación de tuits con tono positivo y negativo a la vez.

3. Eliminación de retuits, puesto que son copias de tuits originales con un posible comentario adicional y aumentaban la complejidad del proceso.

4. Eliminación de tuits con “:P” debido a que en 2009 Twitter API devolvía estos también al buscar para “:(“, pero “:P” no indica un sentimiento negativo.

5. Eliminación de tuits repetidos (Go, et al., 2009).

Sin embargo, los emoticonos están asociados a estados emocionales, hasta el punto en que podrían ser considerados etiquetas de estos en los tuits (Bifet & Frank, 2010), por lo que otros científicos sí que deciden usarlos en sus estudios. Un ejemplo de ello fue el análisis de Pak y Paroubek (2010), en el que asumieron que un emoticono representa una emoción vinculada a todo el mensaje y, por tanto, todo el contenido de dicho tuit está relacionado a esa emoción. Kouloumpis et al. (2011) valoraron más elementos como los emoticonos o abreviaciones de emociones que el número de tipologías de palabras como parte del texto. Otro ejemplo es el de Davidov et al. (2010), que utilizaron 50 hashtags y 15 emoticonos como etiquetas de sentimiento en su estudio como manera de evitar la introducción de un gran conjunto inicial de datos o palabras que manualmente clasifiquen el sentimiento de los tuits. También eliminaron de su conjunto de tuits aquellos que presentaran más de un hashtag o emoticono para reducir la ambigüedad.

Si bien la mayoría de los estudios profundizan en el estudio de los sentimientos independientemente de una temática objetivo, Yu et al. (2011) exploraron la posibilidad de analizar los sentimientos en Twitter siguiendo una estrategia dependiente de un objetivo dado, es decir, teniendo en consideración características de la temática objetivo y los tuits

relacionados de la misma. Saif et al. (2012) analizaron el impacto de cada concepto en la fiabilidad de su análisis, ya que a veces las características semánticas la mejoran, pero en otros casos la reducen. En un estudio posterior, Saif et al. (2015) propusieron una representación de palabras semántica y de sentimientos llamada SentiCircle, que trataba de identificar sentimientos tanto a un nivel de entidad como de tuit usando métodos diferentes y alcanzando mejores resultados que otros métodos basados en el léxico como SentiStrenght.

Otros estudios han enfocado el análisis de sentimientos ya no midiendo la polaridad de cada tuit relevante en una temática, sino desde el punto de vista del hashtag en sí. Wang et al. (2011) propusieron un modelo basado en tres tipos de información: la polaridad de los sentimientos de los tuits que participan en cada hashtag, la relación de coocurrencia entre los hashtags a través de un modelo de grafos y el significado literal de esos hashtags según información semisupervisada. Se puede ver un ejemplo de este modelo en la Figura 76:

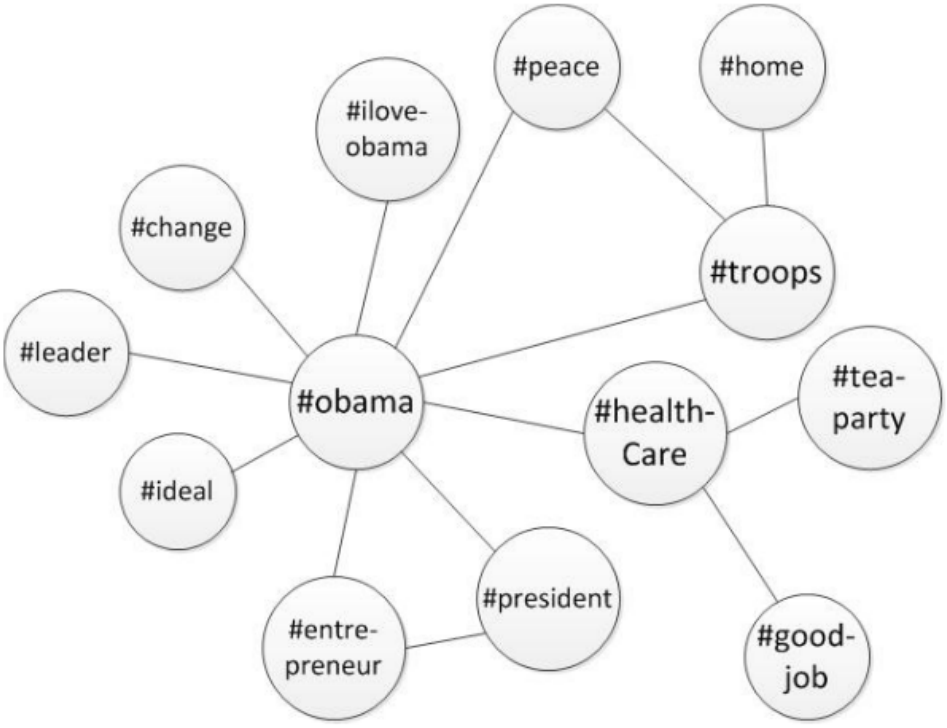


Figura 76. Ejemplo de modelo de grafos de hashtags (Wang, et al., 2011).

Aunque el rendimiento no fue el esperado por sus autores, sí que apreciaron una notable mejora gracias a la clasificación mejorada mediante el significado literal de los hashtags, tal y como se observa en la Figura 77, según lo cual no actualizaban la polaridad del

hashtag que se analizaba en cada instante, sino que se limitaba a ofrecer influencia de sentimiento a los hashtags vecinos.

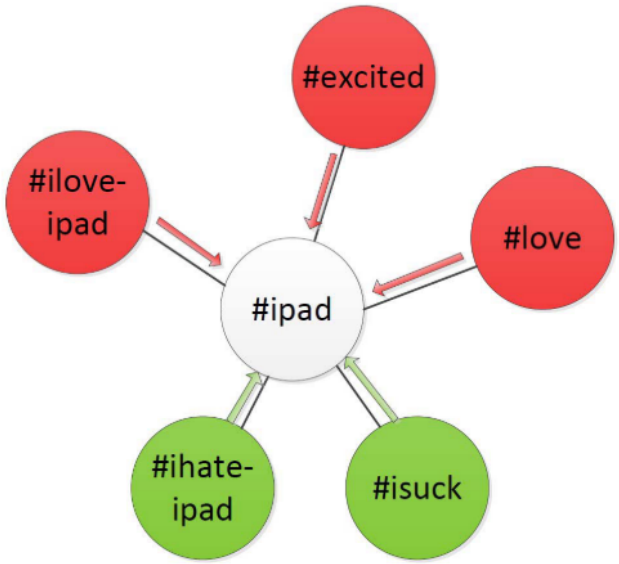


Figura 77. Ejemplo de clasificación mejorada según el significado literal de los hashtags (Wang, et al., 2011).

La limitación de 140 caracteres de los tuits afecta a la manera en que se comunican los usuarios en Twitter, dando lugar al uso de palabras que no son representativas ni sintácticamente consistentes. Es por ello que Kontopoulos et al. (2013) aplicaron al ya clásico análisis de sentimientos por medio de una puntuación de sentimiento, un enfoque en el cual añadían un grado de sentimiento a cada concepto incluido en el tuit, gracias a unas técnicas ontológicas aplicadas mediante la siguiente metodología: creación de un dominio ontológico (una conceptualización compartida especificada, explícita e interpretable por máquinas) y análisis de sentimientos de un conjunto de tuits basándose en los conceptos y propiedades de esa ontología.

Uno de los problemas a los que se ha tenido que enfrentar el análisis de sentimientos ha sido tener en cuenta la ironía o el sarcasmo, ya que transforman la polaridad de aquello positivo o negativo en su sentimiento contrario. González-Ibáñez et al. (2011) estudiaron la posibilidad de detectarlo automáticamente. Sin embargo, no pudieron conseguirlo con un acierto mayor al humano, que ya era bajo de por sí. Barbieri y Saggion (2015) también propusieron un procedimiento que detectara la ironía en inglés, basado en la frecuencia de la cuenta, las diferencias entre la forma de hablar y escribir, sentimientos, ambigüedad, intensidad, sinónimos y estructura. Actuó más bien con un clasificador de tuits según su temática. No obstante, al cruzar los dominios sí obtuvieron un 99% de nivel de confianza, por lo que su modelo podría capturar la ironía si se descarta el dominio.

El análisis de sentimientos en Twitter también puede ser aplicado al estudio de la repercusión de ciertos eventos públicos en el sentimiento de los usuarios. Thelwall et al. (2011) analizaron este fenómeno y descubrieron que hay una fuerte correlación entre eventos que provocan un pico de participación en Twitter y la detección de un sentimiento negativo en ella, y sin embargo no se producía una correlación tan fuerte en el caso del sentimiento negativo. Esto lleva a la interesante conjetura de que, si se produce un evento, es muy probable que produzca efectos de sentimiento negativo antes que positivo, incluso en el caso de eventos que no tengan una connotación negativa, como son los Oscars o los Juegos Olímpicos. La conclusión a la que llegaron fue que Twitter era un medio utilizado sobre todo para impulsar objetivos personales preexistentes al evento (como la creación de humor). Un ejemplo fue el de Tiger Woods, sobre el que solo un 13% de los usuarios expresaron una opinión mientras que el resto se limitaron a analizar, informar, expresar desinterés o cinismo o llevarlo al terreno del humor.

Hansen et al. (2011), por su parte, consideran que la herramienta básica para medir la viralidad es el retuit, y analizando el sentimiento de los tuits comprobaron que un sentimiento negativo aumenta la viralidad en los tuits pertenecientes al ámbito de las noticias, pero no es así en los tuits que no son de noticias, por lo que añaden que “si quieres ser citado: ¡habla bien a tus amigos o postea malas noticias al público!”. Ferrara y Yang (2015) realizaron otro estudio y comprobaron que el contenido negativo se propaga más rápidamente, aunque el positivo alcanza una audiencia mayor. Los eventos, si son esperados, es más probable que susciten sentimientos positivos mientras que si son inesperados provocarán mayoritariamente un sentimiento negativo.

Aunque, siendo más específicos en el ámbito de las noticias, Berger y Milkman (2012) analizaron un conjunto de artículos de New York Times y descubrieron que los sentimientos que más fomentan la viralidad son aquellos que transmiten una excitación fisiológica, sea negativa o positiva. Esto significa que los artículos que transmiten emociones que activan, como el asombro (positivo), enfado (negativo) o ansiedad (negativo) son más virales que aquellos que transmiten emociones que desactivan como la tristeza.

2.2.4. Análisis de tendencias en Twitter

El tercer apartado de los indicadores de esta tesis consiste en el análisis de la tendencia de los términos relacionados con un contenido publicado en Twitter, por lo que resulta necesario comprender en qué consiste este análisis, cómo se difunde la información, cómo se detectan tendencias, las características inherentes de las noticias de última hora y los estudios relacionados con la predicción de tendencias en esta red social.

2.2.4.1. La difusión de la información

Construir modelos de patrones de comunicación en una red social siempre ha sido un desafío, ya que es difícil obtener datos a gran escala de una red social, y más aún obtener datos completos de las dinámicas de comunicación en una red social a través del tiempo. En el capítulo anterior se han mencionado investigaciones relacionadas con la comunicación que se produce debido a un evento. Sin embargo, normalmente este tipo de comunicación se produce durante un corto espacio de tiempo, en un entorno mucho más amplio de comunicación natural y sistémica que circula de manera continuada en la red social. Si se mira este fenómeno de una manera acumulativa, el patrón de fondo nos permite ver información de la comunicación diaria que se expande a través de la red.

Kossinets et al. (2008) consideran interesante, pues, estudiar el potencial informativo de los diferentes nodos, ya no solo por el camino más corto sino por la frecuencia de la comunicación. En el ejemplo de la Figura 78, el nodo B recibe la información más reciente de A vía el nodo C, y no directamente del nodo A.

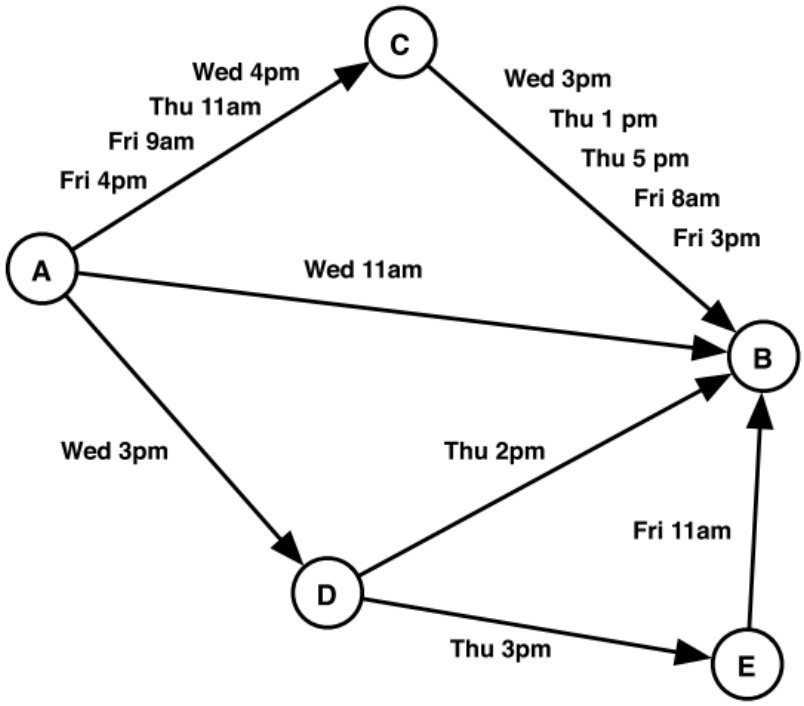


Figura 78. Ejemplo de frecuencia de intercambio de información entre nodos (Kossinets, et al., 2008).

Otras características para tener en cuenta en la diseminación de la información y la adopción de nuevos productos e innovaciones son la influencia de los usuarios y el efecto de boca a boca. Kempe et al. (2015) estudiaron estos efectos de red utilizando modelos de influencia ya conocidos y el comportamiento de cascada, según el cual ciertos usuarios recomiendan un contenido a sus amigos y estos, a su vez, pueden verlo y volver a recomendarlo, generando una tendencia. Alejándose de la suposición de otros estudios de que todos los parámetros del modelo son invariantes entre las observaciones de cada cascada de información, tuvieron en cuenta que la influencia que ejerce un nodo dependerá del contexto de aquello que comparte.

En una red social se publica una gran cantidad de información y, sumado a las publicaciones diarias de los usuarios de carácter más personal, provoca que sea necesario dedicar un gran esfuerzo en filtrar y detectar las temáticas y los eventos más relevantes. Esto se produce según Aiello et al. (2013) mediante la monitorización y obtención de información de fuentes sociales, que debe enfrentarse a desafíos como la fragmentación de la información y el ruido presente en el contenido creado por los usuarios, la publicación en tiempo real, la ráfaga de eventos y su tiempo de resolución. También existe el desafío de la distinción entre interacciones y consumo de información, ya que lo uno no implica lo otro, aunque pueda parecer incongruente. De hecho, Crossfield (2016) cita a Michael Ducker, director de productos de grupo de Twitter, que advierte de que “los botones para

tuitear cuentan el número de tuits que han sido tuiteados con la URL exacta especificada en el botón. Esta cuenta no refleja el impacto en Twitter de la conversación sobre tu contenido - no cuenta respuestas, tuits con citas, variantes de tus URLs, ni refleja el hecho de que algunas personas que tuitean estas URLs pueden tener muchos más seguidores que otras". Algo que reafirmó Haile (2014), CEO de Charbeat, que realizó un estudio de 10.000 artículos compartidos en redes sociales y no encontró ninguna relación entre lo que se comparte y la atención que se le presta.

Tal y como definió Evan Williams, uno de los fundadores de Twitter: "Lo que tenemos que hacer es proveer a la gente de la mejor, la más actual y la más relevante información posible. Pensamos que Twitter no es una red social, sino una red de información. Le dice a la gente aquello que le importa al mismo tiempo que está ocurriendo en el mundo." Es por ello por lo que Twitter cambió la pregunta "¿Qué estás haciendo?" por "¿Qué está pasando?" a la hora de elaborar un tuit en su plataforma, según Rudat et al. (2014). Twitter ha dado desde el inicio mucha importancia a las tendencias, las cuales se originan inicialmente desde términos emergentes y aumentan su frecuencia de uso en un intervalo de tiempo. Y los usuarios también, puesto que adaptan su comportamiento comunicativo a los intereses de su audiencia, aunque a la hora de compartir información sigue siendo importante para ellos el valor informacional y otros criterios inherentes de los mensajes.

Atefeh y Kreich (2015) explican en su estudio sobre las técnicas de detección de eventos en Twitter que, dependiendo de la información disponible sobre un evento, la detección de eventos puede clasificarse en técnicas específicas y no específicas. De esta manera, las técnicas no específicas analizan la señal temporal de los hilos de Twitter en busca de explosiones o tendencias para detectar casos de eventos en el mundo real. Las técnicas específicas, en cambio, dependen de la información y características específicas del evento, como el lugar, el tiempo, el tipo y la descripción, aportadas manualmente antes del análisis. Las técnicas no específicas son de mucha utilidad para descubrir eventos desconocidos, como noticias de última hora, eventos emergentes o temáticas generales. Las técnicas específicas suelen estudiar eventos ya conocidos de antemano, como los eventos sociales.

El procesamiento y las características de la difusión de información pueden ser analizados desde un punto de vista micro y macro, según Qi et al. (2018). La parte micro se refiere al usuario en concreto y la influencia hacia los demás, y si la estructura de la red es determinista, se puede estudiar las características de la red para encontrar los nodos más fuertes y con relaciones cercanas entre ellos. La parte macro pretende explorar el patrón de difusión en el tiempo, observando factores como el volumen de información, el instante

en el que ocurre el pico del evento, la actividad de los usuarios, la probabilidad de retuit, interferencias con otras plataformas y otros eventos y el ruido al mostrar los cambios de velocidad en el tiempo.

Por otro lado, en un experimento para buscar evidencias de contagio complejo de información en Twitter, Mønsted, et al. (2017) indicaron que las dinámicas de propagación pueden obedecer a un modelo de contagio simple o complejo, siendo este último el que representa más fielmente la probabilidad de propagación de la información en las redes sociales. De esta manera, la probabilidad de propagación depende en mayor medida del número de fuentes únicas de información más que en el número de exposiciones, siendo más efectivo por tanto que haya muchas fuentes únicas que una sola, aunque las exposiciones sean las mismas.

2.2.4.2. Detección y clasificación de tendencias

Caro et al. (2010) propusieron un proceso de cinco pasos para detectar tendencias: la extracción y formalización del contenido generado por los usuarios como vectores de términos con sus respectivas frecuencias; definición de un grafo de autores activos y cálculo de su autoridad, modelo del ciclo de vida de cada término, teniendo en cuenta la autoridad el usuario; selección de términos emergentes para clasificar las palabras clave según su estatus de vida; y creación de gráficos temáticos navegables con enlaces entre los términos emergentes y sus términos concurrentes.

Yang y Counts (2010) destacaron en su estudio las tres principales propiedades de la difusión de la información, en un intento de predecirlas en Twitter:

- Velocidad: cómo se verán influenciados los seguidores e interactuaron con la información, ya sea retuiteando, respondiendo o mencionando el tuit inicial con sus propios tuits. Por tanto, se resume en el análisis de si alguien reacciona hacia el tuit de algún modo y cuándo lo hará.

- Escala: cuántos usuarios seguidores de la fuente de información hablaron sobre la misma temática, es decir, usuarios enlazados en primer grado de seguimiento.

- Rango: cuánta influencia ha tenido el tuit según la distancia hasta el receptor en la cadena de difusión, siguiendo los grados de seguimiento hasta llegar al mayor.

Twitter en concreto se caracteriza además por la publicación en tiempo real de muchísimo contenido de sus usuarios, por lo que cualquier enfoque que lo tenga en cuenta se enfrenta a desafíos como el tratamiento efectivo del lenguaje informal, la detección de eventos sin

contar con un número predefinido de eventos y la escalada al gran volumen de datos de Twitter (Kumar, et al., 2014).

En un estudio de tendencias en redes sociales de Asur et al. (2011) comprobaron que el crecimiento de las tendencias en Twitter sigue una distribución logarítmica, y la mayoría de estas no tienen una larga duración. Las que sí la tienen, en cambio, presentan una persistencia que sigue una distribución geométrica. No obstante, al contrario de lo que se podría suponer, el número de seguidores o la ratio de tuits de los usuarios no son atributos que creen tendencias por sí solos, sino los retuits de otros usuarios que estén relacionados con el contenido que se esté compartiendo.

Ardon et al. (2013) realizaron un estudio de 5,96 millones de temáticas con un conjunto de datos de 10 millones de usuarios y 196 millones de tuits, y se centraron en dos aspectos: la red topológica de usuarios seguidores y seguidos y la localización geoespacial. Investigando el efecto de los usuarios que iniciaban una temática y la popularidad de esta, llegaron a la conclusión de que los usuarios con un gran número de seguidores tenían un impacto fuerte en dicha popularidad, la cual se podría producir gracias a conjuntos disjuntos de usuarios que hablan sobre ello. Forman así un gran componente que rompe las barreras regionales de manera agresiva. Esto último es importante, ya que la proximidad y noción de comunidad siguen siendo factores significativos que contribuyen a la popularidad de un tema.

A la hora de investigar la difusión de la información, Guille y Hacid (2012) introdujeron un modelo basado en características sociales, semánticas y temporales que pretendía comprobar si dicha difusión era dirigida por la estructura de comportamientos sociales y locales de los usuarios que participaban en el proceso. Utilizaron el principio AsIC y la regresión logística bayesiana para inferir probabilidades de difusión dependientes del tiempo. Al final, comprobaron que podían predecir correctamente la dinámica de la difusión, pero sin embargo no podían predecir el volumen de los tweets que se iba a generar en la propagación. Otro estudio que aplicó características sociales fue SBIDM, de Mozafari y Hamzeh (2015), que consideró el efecto de la interacción con los vecinos y el consumo de los medios principales, como la televisión y la radio, a la hora de estudiar la difusión de la información en las redes sociales, modelándola así según la vida social de los usuarios.

Marcus et al. (2011) estudiaron la identificación automática de picos de alta actividad en Twitter, etiquetándolos con el texto de los tuits que formaban parte de dicho pico. Crearon una plataforma llamada TwitInfo, que exploraba en tiempo real Twitter en busca de eventos relacionados con una consulta de búsqueda y trata de buscar sub-eventos y términos

importantes dentro de éstos, además de analizar los sentimientos de los usuarios hacia dichos subeventos y su distribución geográfica. El algoritmo de detección de picos llegó a identificar el 80-100% de los picos etiquetados manualmente.

Con la intención de medir la congestión del tráfico y los accidentes de coche, D'Andrea et al. (2015) estudiaron asimismo la detección de eventos en tiempo real mediante Twitter, procesando los tuits, aplicando técnicas de minería de texto y clasificando los tuits según si hablaban del tráfico o no. Obtuvieron una precisión del 95,75%, siendo capaces también de discernir si un evento de tráfico se debía a una causa externa o no. Azam et al. (2015) también aplicaron técnicas de minería de texto para luego aplicar Latent Dirichlet Allocation (LDA) para extraer las características y los términos clave de los tuits en un modelo de espacio vectorial. Kurniawan et al. (2016) analizaron también la congestión del tráfico comparando algoritmos de aprendizaje automático como Naive Bayes (NB), Support Vector Machine (SVM) y Decision Tree (DT), y con SVM obtuvieron 99,77% y 99,87% de precisión clasificatoria en datos balanceados y no balanceados respectivamente, por lo que proponen el estudio de los datos sociales como un mecanismo alternativo para detectar anomalías en el tráfico.

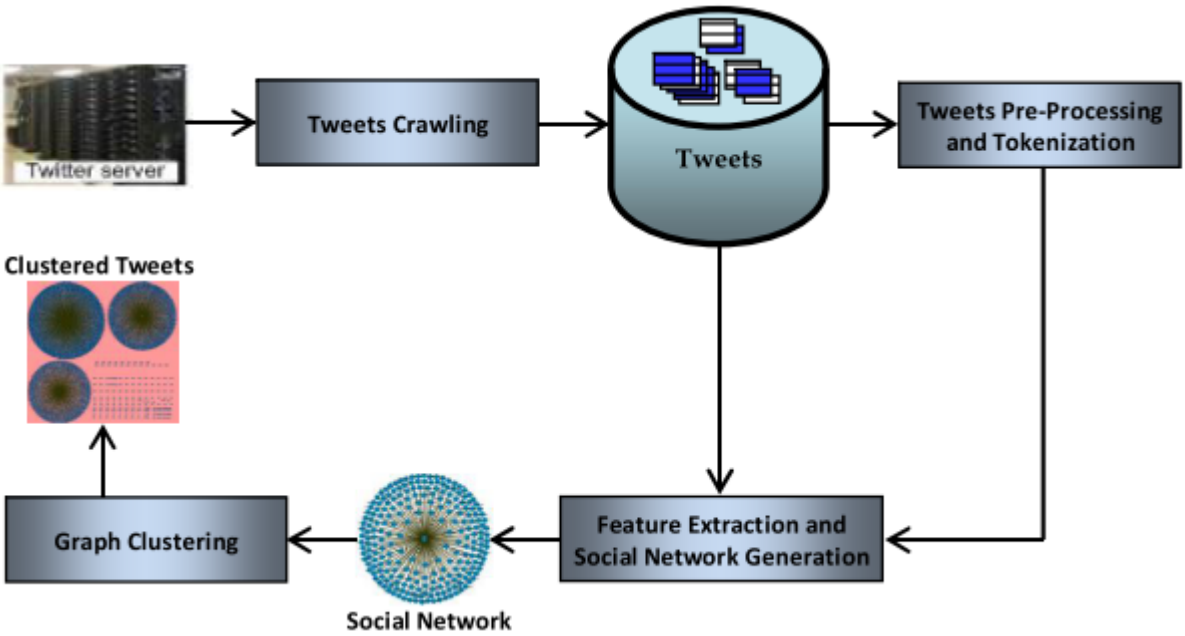


Figura 79. Flujo de trabajo de la técnica de minería de datos (Azam, et al., 2015).

Según McMinn y Jose (2015), la mayoría de los procedimientos de detección de eventos en tiempo real eran demasiado lentos o solo podían detectar eventos de un determinado tipo. Por ello, propusieron clasificar los tuits según las entidades presentes en ellos y detectar explosiones y seleccionar grupos para extraer los grupos que estén relacionados con los eventos del mundo real. Con una complejidad computacional menor y la posibilidad

de una ejecución en tiempo real, obtuvieron datos mejores a los de otros estudios y comprobaron que los sustantivos ayudan a determinar la memoria y los verbos la precisión.

Por otro lado, Zubiaga et al. (2015) afrontaron la clasificación de las tendencias usando tipologías y sin necesitar datos externos. Las tipologías propuestas son: noticias, eventos en curso, memes y actos de conmemoración. Para ello se sirvieron de los patrones asociados a quince características sociales de cada tipo de tendencia: profundidad de retuits, ratio de retuits, número de hashtags, longitud de los tuits, signos de exclamación, preguntas, enlaces, repetición de temáticas, respuestas, velocidad de propagación, diversidad de usuarios autores, diversidad de usuarios que retuitearon, diversidad de hashtags, diversidad de lenguaje y diversidad de vocabulario. Thelwall y Cugelman (2017) apostaron por el método Resonating Topic para buscar automáticamente temáticas exitosas, categorizando los tuits más exitosos y estudiando la organización fuente para identificar nuevas estrategias que puedan derivar en un mayor éxito, así como otras organizaciones para encontrar brechas potenciales de estrategia. Para categorizar los tuits más exitosos, se sirvieron de las siguientes variables: la autoridad del usuario, el número de seguidores del usuario, la temática del tuit, la actualidad del tuit para tener en cuenta fechas o noticias de última hora y el aspecto novedoso del tuit, si lo hay.

Guille y Favre (2014) utilizaron las menciones presentes en los tuits para detectar eventos y mejorar la robustez del sistema frente al ruido del contenido en Twitter. Para ello, crearon un modelo llamado MABED que solo depende de los tuits en sí y estima dinámicamente el periodo de tiempo a estudiar, en vez de depender de un periodo fijo predefinido, lo cual es una ventaja, pero tiene el inconveniente de limitar la detección de eventos duplicados.

Los hashtags también pueden ayudar a detectar eventos y clasificarlos. Cui et al. (2012) propusieron un modelo para ello basándose en tres atributos de los hashtags: su inestabilidad para el análisis temporal, la posibilidad de que incluyan memes como manera de distinguir entre eventos sociales y temáticas virtuales o memes, y la entropía de autoría para encontrar los autores que más contribuyen en el evento.

A la hora de clasificar dichas tendencias, Lee et al. (2011) explican en su estudio que se suele utilizar un enfoque basado en las palabras de los tweets que forman parte de ellas. No obstante, suele verse perjudicado por el ruido producido por la jerga y el límite del número de caracteres de la plataforma. Otro enfoque de precisión significativamente mayor es la clasificación mediante la estructura de la red social. De esta manera, se detectan usuarios influyentes en la temática a la que pertenece la tendencia, se averigua qué temáticas similares utilizan estos usuarios y a qué categoría pertenecen.

También ha tenido su lugar en la academia la investigación de procesamientos de menor coste computacional, como es el caso de TwitterNews+ presentado por Hasan et al. (2016) con un espacio y tiempo de procesamiento constante. Además, utiliza filtros que extraen eventos de noticias del conjunto de candidatos con la intención de detectar también eventos con una explosión pequeña de pocos tuits y descartar eventos triviales.

Por otra parte, Gaglio et al. (2016) estudiaron los datos en tiempo real y realizaron la detección de eventos analizando la corriente de tuits mediante ventanas dinámicas con un tamaño según el volumen de tuits y el instante de tiempo, una selección dinámica de términos y un conjunto de palabras clave que se va refinando con el tiempo para incluir nuevas posibles tendencias, tal y como se ve en la Figura 80. Aplicaron esta metodología junto con la ya existente SFPM para estudiar la 2014 FIFA World Cup, con cinco métricas: temática, precisión de palabras clave, temática de palabras clave, precisión y redundancia), y apreciando la posibilidad de detectar aspectos sociales de los eventos, como lesiones de jugadores o errores de los árbitros.

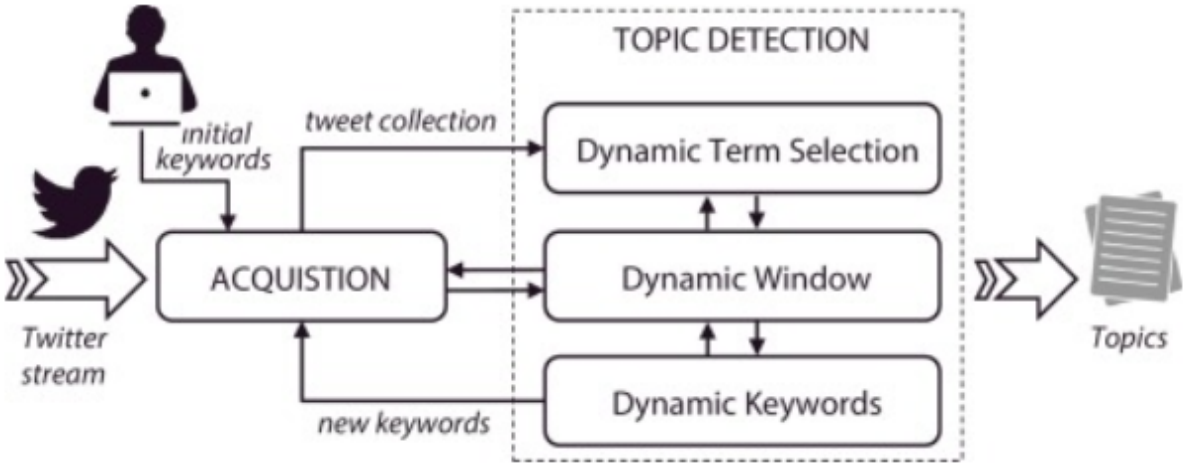


Figura 80. Procedimiento de la detección de tendencias (Gaglio, et al., 2016).

2.2.4.3. Noticias de última hora

La mayoría de las tendencias en Twitter (85% aproximadamente) eran de noticias de última hora o noticias de un mayor recorrido (Kwak, et al., 2010), publicadas por fuentes de medios tradicionales y retuiteadas en Twitter generando las tendencias (Asur, et al., 2011). Es por ello por lo que una de las líneas de investigación favoritas de sistemas de recomendación era, según Phelan et al. (2009), la de recomendación de noticias a los usuarios, con la intención de ofrecer las noticias más relevantes para un usuario en un momento dado. El enfoque más utilizado en su momento era la detección de tendencias tras el análisis de un consumo masivo de información por parte de los usuarios, lo cual

suponía en sí una limitación ya que solo tras producirse dicho consumo podrían ser detectadas.

Las principales temáticas de las tendencias procedentes de noticias son los eventos de música o religión, eventos mediáticos, políticos, de interés humano y deportes. Wilkinson y Thelwall (2012) compararon los principales temas de interés entre distintos países de habla inglesa y analizaron así las diferencias en el uso de estas tendencias a nivel internacional, lo cual es una oportunidad para investigar en campos como problemas de salud o conflictos que afecten a varios países a la vez y comprobar cuál es el impacto que está teniendo en las redes sociales de cada uno de ellos.

La detección de tendencias es de gran valor para periodistas y analistas, ya que podrían obtener de esta forma nuevas noticias que evolucionan con una gran rapidez, aunque también es muy interesante para profesionales de marketing y empresas de análisis de opinión. Tratando de satisfacer este tipo de necesidades, Mathioudakis y Koudas (2010) desarrollaron TwitterMonitor, un sistema que identifica tendencias en tiempo real en Twitter. Primero, identificaba palabras clave “explosivas”, es decir, palabras clave que de repente aparecen en tuits en una ratio inusualmente alto. Después las agrupaba en tendencias basadas en sus ocurrencias simultáneas, definiendo por lo tanto una tendencia como un conjunto de palabras clave explosivas que aparecen frecuentemente juntas en tuits. Tras identificarlas, extraían información adicional de los tuits que pertenecen a esa tendencia.

Sankaranayanan et al. (2009) desarrollaron un sistema de procesamiento de noticias llamado TwitterStand que captura tuits que corresponden a las noticias de última hora. Las fuentes no son conocidas por adelantado y pueden ser numerosas, y los tuits no se publican de acuerdo con un calendario u horario, ya que se producen en tiempo real y tienden a tener ruido producido por una ratio de producción tan alto. Un ejemplo es el de la Figura 81, en el que se pudo observar un relativo aumento de la actividad en Twitter sobre la enfermedad y muerte de Michael Jackson más de una hora antes de que los medios de noticias convencionales informaran de ello.

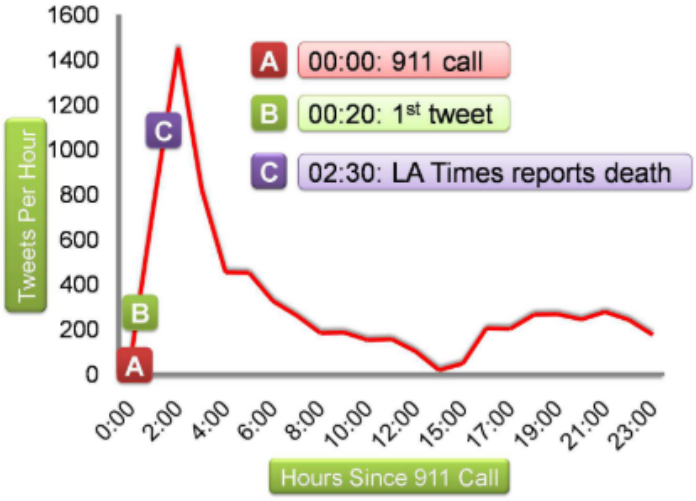


Figura 81. Nº de tuits por hora en las primeras horas tras la muerte de Michael Jackson (Sankaranarayanan, et al., 2009).

2.2.4.4. Procesamiento semántico

El procesamiento del contenido de los tuits también puede ayudar a la hora de analizar tendencias en Twitter. Un sistema propuesto por Li et al. (2012) y diseñado para detectar y analizar eventos, en este caso en el ámbito de los desastres naturales, es TEDAS, cuya arquitectura se puede ver en la Figura 82. Consiste en detectar nuevos eventos, analizar el patrón espacial y temporal del evento e identificar la importancia de este. Su modelo de clasificación busca eventos relacionados con desastres naturales y, para distinguirlos, tiene en cuenta características del contenido (palabras clave dentro de cada tuit y el formato de este), características del usuario (detectando su autoridad y por tanto credibilidad mediante la verificación de la cuenta, el número de seguidores, la edad de la cuenta y el número de tuits publicados) y características del uso (cuántos tuits similares hay en un rango de tiempos y localizaciones, y cuántos tuits hay que contengan los mismos hashtags).

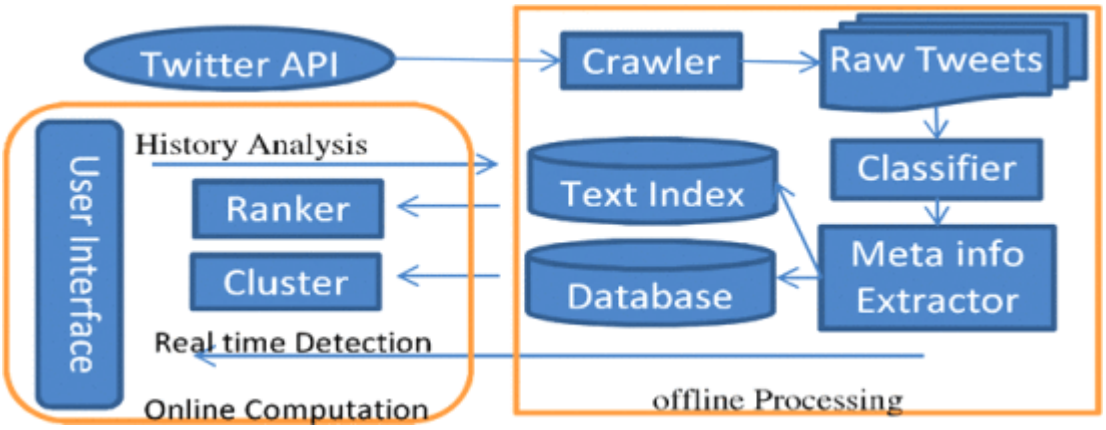


Figura 82. Arquitectura del sistema de TEDAS (Li, et al., 2012).

Alvanaki et al. (2011) propusieron enBlogue, un planteamiento basado en una monitorización continuada de Twitter y RSS de diversas fuentes para detectar cambios repentinos en las correlaciones de etiquetas, pudiendo ser personalizada según los intereses del usuario. Consistía en seleccionar etiquetas fuente según su popularidad y volatilidad, crear pares de etiquetas en las que al menos un de ellas sea una etiqueta fuente, estudiar su correlación y detectar cambios repentinos (no predecibles) en dicha correlación de pares de etiquetas. Un ejemplo esquematizado de cambio repentino sería el de la Figura 83.

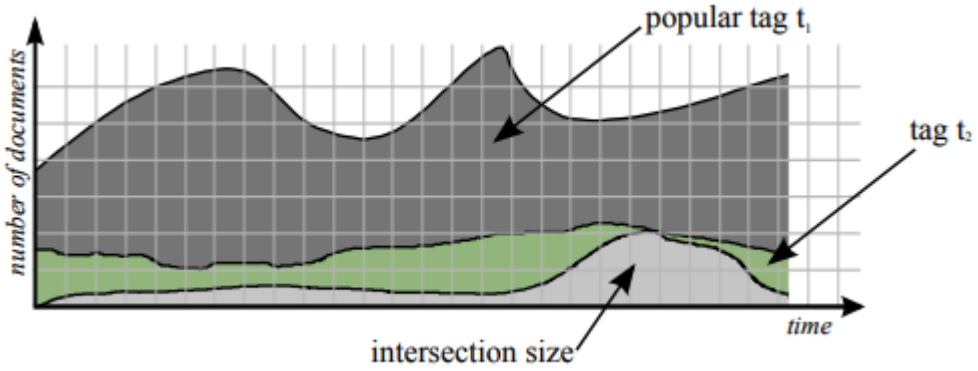


Figura 83. Cambio repentino en la correlación entre dos etiquetas (Alvanaki, et al., 2011).

Una de las primeras integraciones de procesamiento semántico en la detección de tendencias en Twitter la llevaron a cabo Okazaki y Matsuo (2008). Tomando como referencia la información sobre terremotos, trataron de predecirlos, o más bien de avisar sobre ellos lo más pronto posible, gracias a la información en tiempo real de los numerosos usuarios de Twitter. En la Figura 84 se puede apreciar una comparativa de la difusión de información por parte de los medios masivos, la difusión de información útil e inútil por parte de los usuarios en una red social y esta misma, pero con un filtro semántico que ayude a discernir qué información es de utilidad para el usuario:

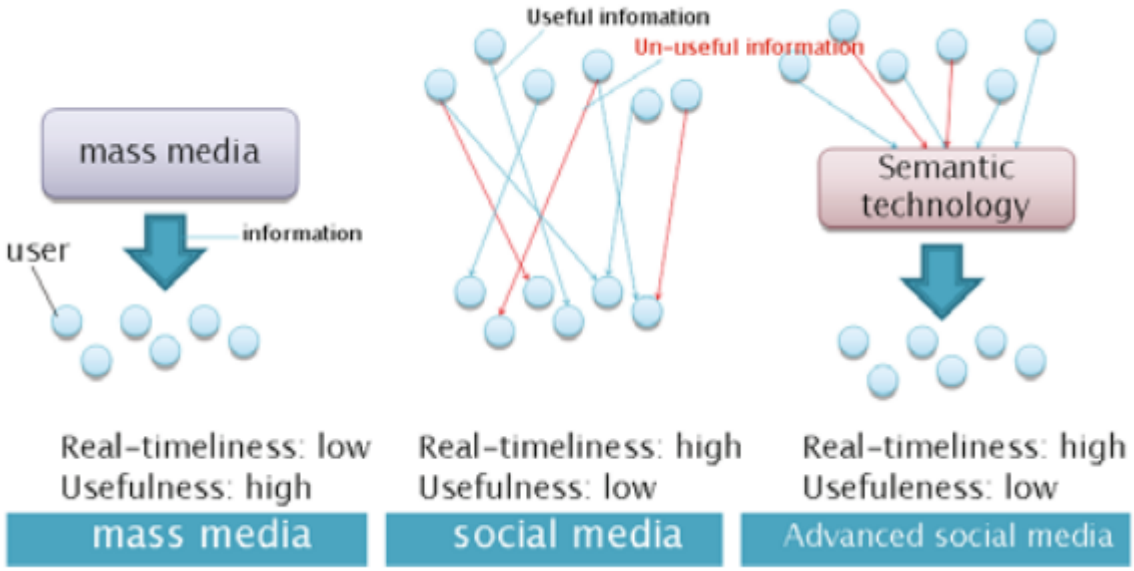


Figura 84. Comparativa de comunicación según su fuente y si se filtra o no su semántica (Okazaki & Matsuo, 2008)

Abel et al. (2011) también estudiaron el enriquecimiento semántico de las noticias y, correlacionado con las actividades de los usuarios de Twitter, mejoraba la precisión de las recomendaciones de noticias de manera significativa.

Por otro lado, Cai et al. (2015) estudiaron las imágenes de Twitter para averiguar en qué medida se podría utilizar dicha característica en la detección de eventos, junto con el texto, la localización, la fecha y hora y el uso de hashtags. Para ello, crearon un procedimiento, que se puede ver en la Figura 85, en el cual recolectaban los datos, detectaban los eventos, les hacían seguimiento y, por último, los representaban de una manera visual. No obstante, las imágenes en Twitter pueden contener mucho ruido, lo cual hace complicada la extracción de las imágenes relevantes para el evento que se quiere analizar. Para eliminarlas, utilizaron un filtro de dos capas y definieron tres características para seleccionar las imágenes representativas: relevancia visual, el número de veces que una imagen ha sido tuiteada; coherencia visual, el número de vecinos cercanos determina si es coherente con el evento y las probabilidades de ser una imagen representativa; y la singularidad, puesto que si tienen contenidos diversos las imágenes son más informativas que si tienen contenidos duplicados.

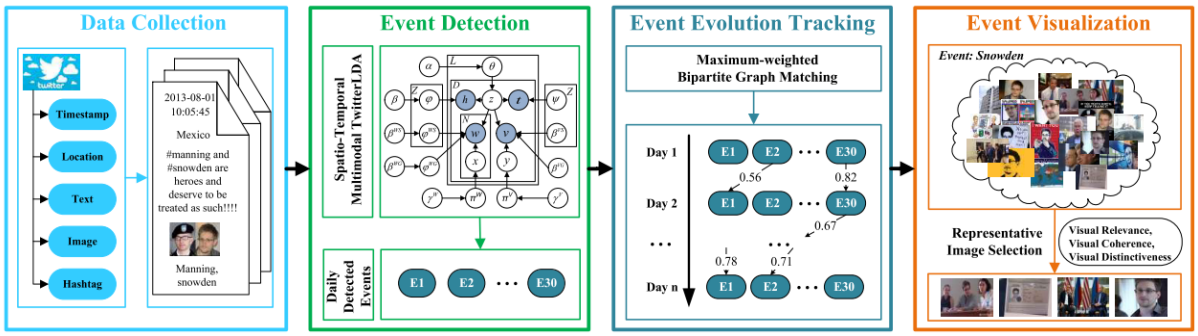


Figura 85. Detección, seguimiento y visualización de eventos en Twitter (Cai, et al., 2015).

Pero no todo lo que se comparte en la red social es estrictamente información, sino que también se puede producir tendencias de emociones. Al parecer, la expresión individual de emociones también depende de las expresiones individuales de su red social, algo que Coviello et al. (2014) estudiaron teniendo en cuenta la influencia de la lluvia en el contenido emocional de los tuits. Comprobaron que también afectaba a los tuits de seguidores de otras ciudades en las que no estaba lloviendo.

Los usuarios pueden tener una susceptibilidad alta o baja al contagio de emociones, siendo dicha susceptibilidad aplicable por igual a las emociones tanto positivas como negativas, según Ferrara y Yang (2015). Este contagio se produce debido a que los sentimientos presentes en los contenidos publicados están relacionados linealmente con los estímulos emocionales a los que el usuario ha estado expuesto. Tsugawa y Ohsaki (2017) analizaron 4,1 millones de tuits para averiguar la relación entre esta polaridad de sentimientos y la viralidad, y comprobaron que los mensajes negativos provocan más retuits (un 20-60% más) y de manera más rápida (25% más) que los mensajes positivos y neutrales, pero sin embargo la difusión recurrente es menos frecuente en los mensajes negativos que en los demás.

2.2.4.5. Predicción de tendencias

Un enfoque realmente interesante es el de la predicción de popularidad de mensajes. Este análisis predictivo tiene aplicaciones en muy diferentes dominios, por lo que Kursuncu et al. (2019) proponen un paradigma como el de la Figura 86 a dos niveles de predicción: encontrar señales de bajo nivel (grano fino) de los tuits individuales para construir bloques de dominios independientes, y combinar estas señales para predecir resultados y acciones de alto nivel (grano grueso) de dominios específicos. Obtuvieron buenos resultados con eventos como elecciones, violencia con armas, consumo de drogas, etc.

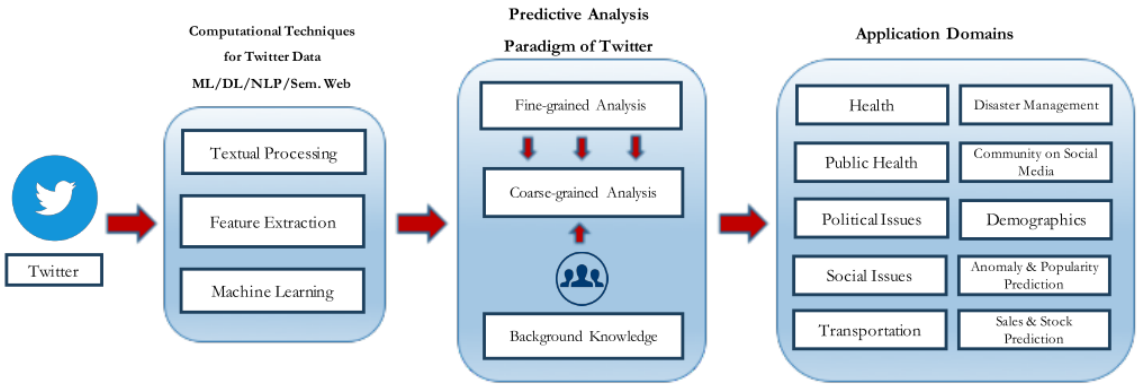


Figura 86. Propuesta de predicción de eventos (Kursuncu, et al., 2019).

Hong et al. (2011) propusieron un método que medía el número de futuros retuits con la intención de averiguar qué factores influencian la propagación de información en Twitter. Lo hicieron entrenando un clasificador binario previamente con ejemplos de mensajes positivos y negativos para medir sus retuits, y entrenando otro clasificador multiclase para predecir el rango de volúmenes de futuros retuits del nuevo mensaje. Los resultados sugerían que se podría obtener la predicción tanto de si el tuit se retuiteará y en qué medida.

Lu y Yang (2012) tomaron prestado MACD, una técnica de análisis de bolsa, para predecir tendencias en Twitter. MACD convierte medias móviles exponenciales en un oscilador de momento para obtener la media móvil más larga de entre todas las medias móviles más cortas, consiguiendo de todo ello: el seguimiento de tendencias y el momento. El procedimiento consiste en monitorizar palabras clave de temáticas de noticias en Twitter, computar dos medias móviles diferentes en tiempo real, definir el concepto de momento de la tendencia eliminando la media móvil del periodo más largo al del más corto y, por último, usar el momento para predecir tendencias de esas mismas temáticas en el futuro.

Pero no solo es interesante predecir las tendencias, sino también el periodo activo de estas. Huang et al. (2020) consideran el periodo activo de la evolución de la popularidad como la duración del tiempo en el que el contenido recibe una continua atención por parte de los usuarios. Para ello, es necesario enfrentarse a dos desafíos: un análisis multifactorial y la elección del instante de tiempo a partir del cual contabilizar el periodo activo. Para lidiar con ambos desafíos, un estudio reciente de Huang et al. (2020) utilizó un sistema basado en DNN (*Deep-Neural-Network-based*) para encuadrar los factores dinámicos y estáticos, y con la activación del periodo activo en un nivel mínimo que la popularidad siempre alcance, y en comparación con otros métodos anteriores como SVR (Hu, et al., 2017) y SpikeM (Myers, et al., 2012), alcanzaron unos objetivos superiores.

2.2.4.6. Otros tipos de análisis

La efectividad de los hashtags depende de la libertad de los usuarios a la hora de decidir si utilizarlos o no y qué hashtags incluir en sus tuits (Ma, et al., 2014). Es por ello por lo que la recomendación de noticias no es la única que se ha estudiado en la literatura. Otros autores como Zangerle et al. (2011) también han desarrollado procedimientos para recomendar los hashtags más apropiados para el usuario con la intención de promover su uso y evitar la utilización de hashtags sinónimos. El usuario introduce un tuit y el sistema le recomienda los hashtags más interesantes para ser incluidos en él. Las recomendaciones siguen los siguientes pasos: se encuentra tuits similares al tuit introducido en el conjunto de datos ya rastreado; se detecta el conjunto de hashtags utilizados en los mensajes más similares; y se clasifica dicho conjunto de hashtags según el número de veces que aparece en los datos rastreados, el número de veces que aparece en los mensajes similares y el valor de similitud entre estos tuits similares y el tuit introducido.

Otra aplicación en este ámbito es la investigación de audiencias, que consiste en analizar cómo se comunica parte de la audiencia de un medio con los creadores de dicho medio. Gracias a ella se puede medir el tono de la reacción del público ante eventos y noticias en tiempo real (Deller, 2011). Siguiendo esta lógica, Congosto et al. (2013) hicieron un estudio del comportamiento de la audiencia de los Goya 2013 en Twitter, y comprobaron que la audiencia audiométrica no es la misma que la social salvo en ocasiones puntuales, y no presentan tendencias parecidas. En ocasiones incluso crecen en sentido opuesto. Sí que observaron que el análisis de la audiencia social complementa al de la audiencia audiométrica, ofreciendo datos sobre los espectadores como la tipología de los usuarios que participan en la conversación sobre el evento. Los tres tipos presentes son los medios de comunicación, los personajes públicos mediáticos y los personajes no mediáticos.

Relacionado también con la detección de usuarios spam del capítulo anterior está el análisis de contenido spam que pretende ser tendencia y promocionar enlaces a través de las granjas de enlaces o *link farming*. Se trata de una estrategia consistente en la obtención de una gran cantidad enlaces hacia una web en una red social u otras webs. Según Ghosh et al. (2012), la intención a menudo no es solo de incrementar el tráfico sino también la influencia percibida. Para detectar estas granjas de enlaces, estos autores publicaron un estudio en el que aplicaron una metodología consistente en elaborar una gran muestra representativa de *spammers* y la conectividad y actividad de estos *spammers* y los usuarios que están conectados a ellos. Sorprendentemente, descubrieron que los usuarios más legítimos, populares y activos de Twitter eran los que más se involucraron en actividades

de granjas de enlaces. Una manera de penalizar esta actividad sería, como proponen los autores, la de penalizar a los usuarios que sigan a usuarios *spam*.

Es por ello por lo que resulta interesante analizar la credibilidad de los tuits, además de su influencia o éxito en la creación de tendencias. Para que pueda formar parte del análisis de tendencias que se realice en tiempo real, el análisis de credibilidad también debería poder hacerse en tiempo real, consideran Gupta et al. (2014), quienes presentaron el primer trabajo de investigación propuesto para ello, que asignaba una puntuación de credibilidad a los tuits de un usuario en tiempo real. Hamidian y Diab (2019) aportaron un procedimiento mediante el que se detectan los rumores y, después, los clasifican en un solo paso mediante RDC en seis tipos: no es un rumor, apoya un rumor, niega un rumor, cuestiona un rumor, neutral y tuits indeterminados; o en dos pasos, en el primero separan los que no son rumores, los indeterminados y el resto, y el segundo clasifican el resto en los otros cuatro tipos antes mencionados, teniendo en cuenta para ello características del contenido, de los conjuntos de palabras presentes en el tuit, la forma de hablar, análisis de sentimiento, emoticonos, detección de entidades, extracción de eventos, el tiempo de publicación y propagación, las respuestas, retuits, ID de usuarios, uso de hashtags y uso de URLs. Con ello, obtuvieron mejores resultados que en otros procedimientos de la academia hasta la fecha.

2.2.5. Publicidad digital en la web

El primer apartado de los indicadores de esta tesis incluye también algunos parámetros relacionados con la publicidad instalada en la web que se va a estudiar. En el estudio actual se ha limitado a los anuncios automáticos de Google AdSense[35], por lo que se justifica la necesidad de definir qué es la publicidad web, cómo se optimiza, cuáles son los sistemas más conocidos y cómo funcionan, además de cómo se gestionan el fraude y el bloqueo de anuncios.

2.2.5.1. La publicidad en la World Wide Web

La publicidad, según Martín de Antonio (2000), es un proceso de comunicación en el que se producen seis características: se trata de un proceso específico de comunicación, transmitiendo un mensaje con un objetivo y a través de un medio concretos; es impersonal; es una comunicación pagada por el anunciante, por lo que este controlará la extensión y características de sus anuncios; la intención de la publicidad es la de tener el mayor alcance posible; puede anunciar productos tanto tangibles como intangibles y servicios; trata de informar sobre un producto, servicio u organización, influyendo así en la compra o aceptación de este.

Históricamente se ha comprobado que la industria de la publicidad siempre se ha adecuado a las nuevas tecnologías, como fueron en su momento la radio y la televisión. La diferencia entre Internet y los medios anteriormente mencionados, señala Lei (2000) es que éstos son plataformas eminentemente de entretenimiento, mientras que Internet es un medio basado en la información.

La publicidad web apareció por primera vez en 1994 (Bruner & Gluck, 2006). El nacimiento de la World Wide Web introdujo también el comercio electrónico, definido por Kiani (1998) como el intercambio de información, servicios, bienes y pagos. En un principio, muchos departamentos de marketing utilizaron Internet basándose en un modelo de comunicación tradicional, para lo cual la academia exploró la necesidad de modelos de comunicación bidireccionales y nuevos conceptos que aprovecharan las oportunidades de este medio, que consistían principalmente en la direccionalidad, la flexibilidad y la accesibilidad. Kiani (1998) propuso un modelo en el que se tiene en cuenta los patrones de comunicación entre la compañía, los consumidores y otras compañías en un entorno interactivo.

[35] https://www.google.es/adsense/

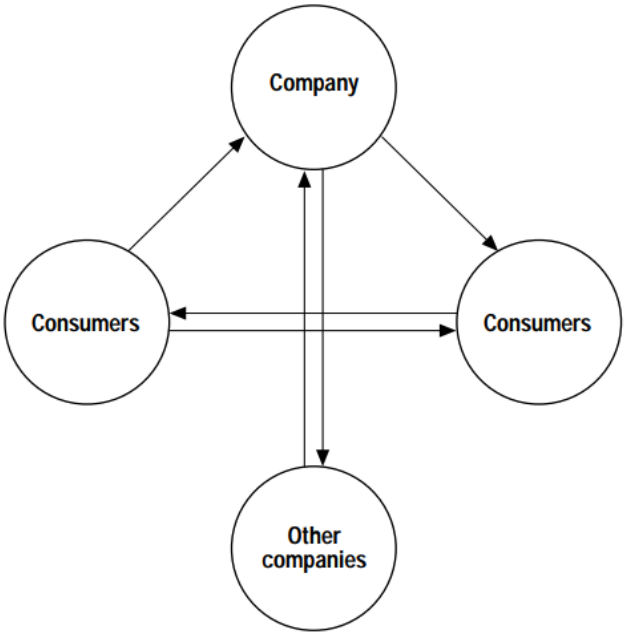


Figura 87. Modelo de comunicación online (Kiani, 1998).

Para ello se sirvió de un estudio de 95 de las compañías Fortune 500 de Kierzkowski, McQuade et al. (1996), que aporta un modelo conceptual con los cinco factores esenciales del marketing digital:

- Atraer usuarios, gracias a los enlaces desde otros sitios web y otros mecanismos de marketing.
- Captar el interés y la participación de los usuarios, de manera que se consiga una interacción o transacción.
- Retener a los usuarios y asegurar su retorno, manteniendo los contenidos actualizados y/o creando nuevos contenidos que sea modificable con el tiempo.
- Aprender de sus preferencias, así como las características demográficas, actitudes y comportamientos de los usuarios. Esto se puede realizar mediante emails, encuestas, cuestionarios o procesos de registro, así como la medición de los clics y transacciones.
- Proveer interacciones personalizadas de manera que el producto o servicio se adecúe a cada cliente en cada momento en forma de sugerencias de otros productos o un servicio personalizado. Un objetivo que es posible manteniendo una relación bidireccional en la que aprender de cada uno.

Las características principales de la publicidad en internet son, tal y como las describió Gómez Nieto (2016), las siguientes: con un público objetivo joven, un alto número de

soportes en forma de una gran variedad de sitios web, mucha flexibilidad de contratación, un sistema de medición de audiencias de baja fiabilidad, una comunicación interactiva como punto fuerte y una accesibilidad total.

Berthon, Pitt et al. (1996) también desarrollaron un concepto que llamaron eficiencia de conversión para analizar la eficiencia con la que el comercio electrónico opera, gracias a recomendaciones basadas en la conciencia, el contacto o visita, la compra y la recompra. Sukpanich y Chen (1998) propusieron un modelo de medición de la efectividad de la publicidad web basado en tres conceptos: la conciencia cognitiva, basada en si los usuarios han notado los anuncios; la preferencia, aprovechando la ventaja adaptativa de internet a las necesidades del usuario frente a los medios tradicionales adaptados a una audiencia masiva; y el comportamiento intencionado, como la intención de aprender, compartir información y comprar.

La publicidad dirigida, también conocida como publicidad comportamental, ha demostrado un gran potencial de efectividad con el paso del tiempo, aseguran Chen y Stallaert (2014) en su estudio al respecto. Consiste en utilizar información del comportamiento del usuario desde su navegador para seleccionar los anuncios que se le mostrarán. Llega a duplicar los ingresos de publicidad frente a la publicidad no dirigida. Este aumento de ingresos no siempre se produce, ya que depende del grado de competición y de las valuaciones de los anunciantes, por lo que se aprecia un efecto de competición y de propensión o tendencia. El efecto de competición consiste en el nivel de competición que haya entre las marcas que se dirigen al mismo público, mientras que el efecto de propensión o tendencia consiste en la ratio de clics en anuncios dirigidos. La publicidad comportamental, por tanto, será más efectiva que la no comportamental si el efecto de propensión o tendencia domina al efecto de competencia.

Macias (2003) considera que el factor de interactividad también es muy importante en el proceso, ya que influye positivamente en la percepción de marca y de los anuncios, y tiene una relación directa y positiva con la comprensión y la intención persuasoria. Para ello, también es necesario que el usuario entienda fácilmente el anuncio interactivo, lo cual influye positivamente en el proceso, por ejemplo, mediante la comunicación de los beneficios del producto o servicio. La intención es crear una experiencia interactiva, en la que el usuario realiza acciones libremente en tiempo real, lo cual según Carrillo y Castillo (2005) ofrece unas posibilidades extra que la publicidad digital pero no interactiva no tiene. Este tipo de experiencias animan al usuario a controlar de manera activa el proceso, crean una mayor sincronía entre reacciones diferentes y permite la multidireccionalidad al poder

vincular a más de un usuario a la vez. Todo ello contribuye a que el anuncio adquiera una mayor complicidad con el usuario, fusionando si es posible lo audiovisual y la interactividad.

Ante la necesidad de anunciarse en Internet para atraer el tráfico necesario para el comercio electrónico, las webs se dividieron en los años 90 según Brain (2002) en las pertenecientes al comercio electrónico y las que ofrecían contenido audiovisual y que generaban sus ingresos mediante la publicidad en su web. Como forma de llevar tráfico de las segundas a las primeras mediante la publicidad, se advirtió un crecimiento importante en el número de webs que realizaban campañas consistentes en ventanas emergentes, anuncios con música o movimiento, etc.

Además, Brain (2002) explica que se ofrecen lugares de la web en forma de banners, con la intención por parte de los comercios de hacer conocida su marca, obtener tráfico de pago e incluso, en ocasiones, obtener ventas directas del mismo banner. Estos banners se ofrecen en una variedad de formas y tamaños, y mostrándose en lugares como la cabecera de la página, la barra lateral, flotando por encima del contenido o, como antes se ha mencionado, en forma de ventanas emergentes. Según Yu y Stotlar (2000), se trata de la herramienta de marketing directo más efectiva, en parte gracias a que produce unidades cuantificables, y es el método de publicidad predominante en la Web suponiendo un 54% de las ganancias del sector de publicidad digital en el año 2000.

Hoffman y Novak (2000) distinguieron dos contenidos publicitarios: los banners, entendidos estos como pequeños, típicamente rectangulares y con imágenes gráficas y con poca información que inviten al visitante a clicar; y las comunicaciones objetivo, que podrían ser una página web, aplicaciones Java, medios de transmisión en tiempo real o formularios. Los primeros serían considerados exposiciones pasivas de publicidad mientras que los segundos serían de tipo activo.

De este modo, no solo se consiguen clics sino también aumentar la presencia de marca de aquello que se anuncia, obteniendo los mismos anuncios en la versión impresa y digital unos resultados igualmente efectivos al respecto (Gallagher, et al., 2001).

El cálculo de la rentabilidad, entendida como ROI (*Return On Investment*) se expresa en porcentaje, siendo ROI = (ganancia obtenida - inversión) / inversión, tal y como indica Oliva Rodríguez (2014). Los banners de publicidad usan los formatos IAB (*Interactive Advertising Bureau*) como estándar, entre los que se incluyen los tamaños presentes en el siguiente gráfico:

Figura 88. Formatos IAB de banners de publicidad (Oliva Rodríguez, 2014).

McCandless (1998) indicó en su estudio que en 1997 ya se invirtieron 906,5 millones de dólares en anuncios en Internet en Estados Unidos, ayudándose de características inherentes de la publicidad en internet como que cada anuncio es servidor a un usuario en concreto, su efectividad puede ser medida de manera directa y el entorno tiene una naturaleza de tiempo real en la que la adaptación a la retroalimentación de los usuarios puede ser casi inmediata. Ese mismo año, el 85% de los periódicos online incluían publicidad digital, aunque en el 65% de los casos los clientes no eran los mismos para las versiones en papel y digital (Neuberger, et al., 1998). Briggs y Hollis (1997) realizaron un estudio mediante el cual comprobaron que la publicidad web podría obtener mejores resultados que la tradicional en televisión o medios escritos, incluso sin tener en cuenta el *click-through*. Goldsmith y Lafferty (2002), en cambio, preguntaron a 329 estudiantes, y llegaron a la conclusión de que el medio online suscita más ventajas, pero también más

inconvenientes, lo cual podría ser motivado porque en Internet las opiniones se polarizan más y los anuncios no son tan bien valorados ni se recuerdan tanto como en la televisión. También comprobaron que, si los visitantes valoraban bien una web, también valoraban positivamente los anuncios en ella.

Sin embargo, el contenido impreso, por características propias de la entrega de cada medio, obtiene una puntuación más alta de retención en la memoria que el contenido digital, según comprobaron Shyam Sundar et al. (1998) al estudiar su comparación. Las características que podrían influenciar en ello son la visualización total de la página impresa frente a la parcial de una página web, las expectativas como lector y como usuario y que se ha apreciado una necesidad de esfuerzo mayor por parte de los anuncios online para captar la atención. Esto último se debe a que, según Benway (1998), se ha producido un efecto de "ceguera" con respecto a la atención que captan los anuncios en las webs, incluso dándose el caso de que los banners situados en los lugares más predominantes captan menos la atención que aquellos localizados más abajo o cercanos a otros enlaces.

Yuan et al. (1998) dividieron los tipos de publicidad según si era directa o indirecta, agrupada con contenido no publicitario o no agrupada, con recompensa o sin ella. En el caso de las webs corporativas, al ser publicidad indirecta, son los clientes los que toman las decisiones sobre cuánto y qué anuncios ven. En el caso de la publicidad digital directa, los usuarios normalmente consumen contenido no publicitario y, de paso, ven los anuncios de manera agrupada, cuyas características controla la página web que los sirve. Por último, las recompensas se pueden dar de manera automática en el caso de la publicidad digital según su navegación o acciones, como podría ser un descuento en el precio de un contenido digital.

El público latino no ha sido una excepción, porque solo las compañías de Estados Unidos invirtieron una suma de 1,7 billones de dólares para anunciarse al público hispánico en 1998, según informaron en su estudio Korgaonkar et al. (2001). Sus principales motivaciones para el clic son sus percepciones de una información de producto útil, una mejora de roles sociales, placer o hedonismo, promoción del materialismo y que sea una información en la que puedan confiar.

Si ya se había observado una diferencia entre los hombres y las mujeres en las actitudes y creencias hacia la publicidad en los medios tradicionales, también se observó que estas diferencias se producían de manera significativa en la publicidad digital. En un estudio de Wolin y Korgaonkar (2003) se comprobó que los hombres presentaban actitudes y creencias generalmente más positivas hacia Internet que las mujeres. Los varones

presentaban una intención más funcional y de entretenimiento que las mujeres, quienes preferían navegar con la intención de comprar por Internet.

En cuanto a las campañas multinacionales, An (2006) comprobó que existe una diferencia cultural en el uso de conceptos como información, la estrategia de aserción simbólica, el rendimiento y una señal de oferta especial, así como la cantidad de información en un anuncio o el número de personas que aparezcan en este. Así se observó analizando el contenido de anuncios de 33 marcas multinacionales de Corea y Estados Unidos, por lo que, si la intención es realizar una campaña de publicidad web internacional, sería conveniente prestar atención a los valores que existen en los contextos culturales de los mercados en los que se vaya a participar. Puesto que según Fu y Wu (2010) una cultura de gran contexto exhibe una estructura cognitiva más compleja, es necesario que los diseñadores de anuncios tengan en cuenta el contexto cultural, directo o indirecto, al que pertenece el público objetivo, así como el método de comunicación que prefiere, directo o indirecto.

La publicidad digital ha demostrado una tendencia claramente ascendente. Solamente en el último trimestre de 2003 se invirtieron 2.200 millones de dólares en Estados Unidos, principalmente de banners en páginas web (Amiri & Menon, 2006). En 2005 y tras un aumento del 30% por año en los dos últimos años, el gasto anual en publicidad digital era de 12.500 millones de dólares, siendo el 25% de estos de compañías del Fortune 500 (Bruner & Gluck, 2006). En 2007, el beneficio ya habría alcanzado los 21.100 millones de dólares, más de la mitad de éstos de publicidad web en sí, viéndose reforzada por el uso de sistemas de anuncios web que organicen los anuncios según los factores que motivan una mayor efectividad (Zhang & Kim, 2008). En 2008, la publicidad en Internet casi había duplicado sus números desde 2005 pese a la recesión económica, debido a la cual su crecimiento se ralentizó (Kirchhoff, 2009). En 2009, la industria de la publicidad alcanzó los 65.000 millones de dólares en todo el mundo, repartiéndose de la siguiente manera: 18.000 millones en Europa Occidental, 22.700 millones en Estados Unidos, 2.000 millones en América Latina, 2.000 millones en China y 720 millones en Rusia (Shanahan & Kurra, 2011). En el primer trimestre de 2011 los ingresos de publicidad digital en Estados Unidos alcanzaron los 7.300 millones de dólares (Yuan, et al., 2012). En 2013, la publicidad digital en Estados Unidos generó un 17% más que el año anterior alcanzando los 42.780 millones (Kumar, 2016). En 2014, la publicidad tanto digital como convencional supuso el 1,07% del PIB español, subiendo cinco centésimas (Gómez Nieto, 2016). En 2016, los beneficios de la publicidad digital alcanzaron 72.500 millones solo en Estados Unidos, siendo la segunda tipología con más beneficios tras los anuncios de televisión (Hussain, et al., 2018). En

2018, la publicidad digital en Estados Unidos generó 107.500 millones de dólares, un 22% más que en 2017 (IAB, 2019).

Según Farhi (2007), los periódicos digitales de Estados Unidos recaudaron 1.200 millones de dólares en 2003, y en 2006 incrementaron esa cifra hasta los casi 2.700 millones de dólares, aunque ya por entonces demostraron una importante ralentización en el crecimiento de su audiencia online, de solo el 2,3% entre 2006 y 2007. Algunos estudios como el de Kumar y Sethi (2009) se han centrado también en investigar una fórmula combinada de contenido de pago y anuncios en los periódicos digitales, ya sea con un precio de suscripción variable o con ambos dinámicos.

Por otro lado, Mediavilla y Abuín Vences (2007) trajeron a colación la denominación de Web 2.0 o Web Social de Tim O'Reilly de 2004 como la segunda generación de webs o micromedios entre los que se incluyen los blogs, podcasts, videocasts y wikis, y apuntan principalmente al grupo de "nativos digitales", personas menores de 35 años, con perfil de comprador online y que generan también contenidos en internet. Estas tipologías de webs utilizan también en ocasiones otros métodos de publicidad digital como la mención patrocinada de productos, publicidad en el contenido enviado a los suscriptores, etc.

El carácter multicanal de Internet facilita que los usuarios también se vean expuestos a los anuncios a través de distintos tipos de dispositivos. A los ordenadores se sumaron los llamados dispositivos móviles: tablets y smartphones. Según Ghose et al. (2013), en 2012, nueve de cada diez personas usaban múltiples pantallas a la hora de realizar una tarea. Analíticamente nos podemos encontrar con conversiones en un canal que en realidad se han producido gracias al clic en un anuncio en otro canal, cuya analítica de CTR (ratio de clic) se ve perjudicada debido a eso. Es muchísimo más probable, opinan estos autores, que los usuarios cliquen en un anuncio desde un dispositivo móvil, lo cual está provocando que las empresas incrementen el presupuesto para publicidad en dispositivos móviles. La ratio de conversión cruzada entre canales es 2,7 veces mayor de móvil a web que al revés, aunque sigue siendo positivo cuando los anuncios están presentes para todos los dispositivos, ya que se produce un efecto de refuerzo al haber visualizado el mismo anuncio en dispositivos diferentes, sobre todo produciéndose la conversión finalmente en el formato web. En 2012 ya circulaban 2.100 millones de anuncios para dispositivos móviles, y Dihn Le y Ho Nguyen (2014) estudiaron qué factores predicen mejor las actitudes hacia la publicidad móvil. Descubrieron que los factores que más ayudan a ello son la credibilidad o confianza y, en segundo lugar, el entretenimiento. En cambio, otros factores como la necesidad de información y la irritación no ayudan a predecir las actitudes

hacia la publicidad móvil, algo que contrasta con los resultados de otros estudios anteriores.

2.2.5.2. Optimización de la publicidad web

Con el crecimiento del tamaño y número de los sitios webs en Internet, creció asimismo la necesidad de optimizar la localización y programación en el tiempo de los anuncios. Para ello, dos conceptos claves según Aggarwal et al. (1998) son las visualizaciones de anuncios y los clics en éstos. Se considera que éstos han sido expuestos cuando se han servido al usuario, y que han sido cliqueados cuando un usuario ha cliqueado en el enlace de éstos. Las dos variables principales a la hora de medir la efectividad de un anuncio son el coste del número de exposiciones (CPM, coste por mes por cada mil impresiones), el coste por clic (CPC) o ambos. Teniendo en cuenta estos conceptos, estos autores destacaron las siguientes complejidades:

- Distribución de ratio de acceso a las páginas web, ya que algunas son mucho más accedidas que otras.
- Equidad, puesto que la localización del anuncio tendrá una implicación directa en su exposición, siendo esta la visualización directa por parte del usuario.
- Programación dinámica, que asegure que todos los anuncios mantienen un porcentaje mínimo de exposición.
- Tiempo del día, ya que es un factor que influye en la cantidad de clics en determinados anuncios.
- Clasificación de contenidos, puesto que los anuncios deben ser apropiados para el contenido de la web en la que se muestran.
- Decisiones dependientes del cliente, según información de la propia IP, como la localización del usuario, datos demográficos, etc.

Para asignar los anuncios a cada web, Aggarwal et al. (1998) recomendaron servirse de las variables de la ratio de Clic/Exposición y el factor de recurrencia con el que el usuario ha accedido la misma página web, buscando desde el punto de vista de los sistemas de anuncios el del flujo de coste mínimo.

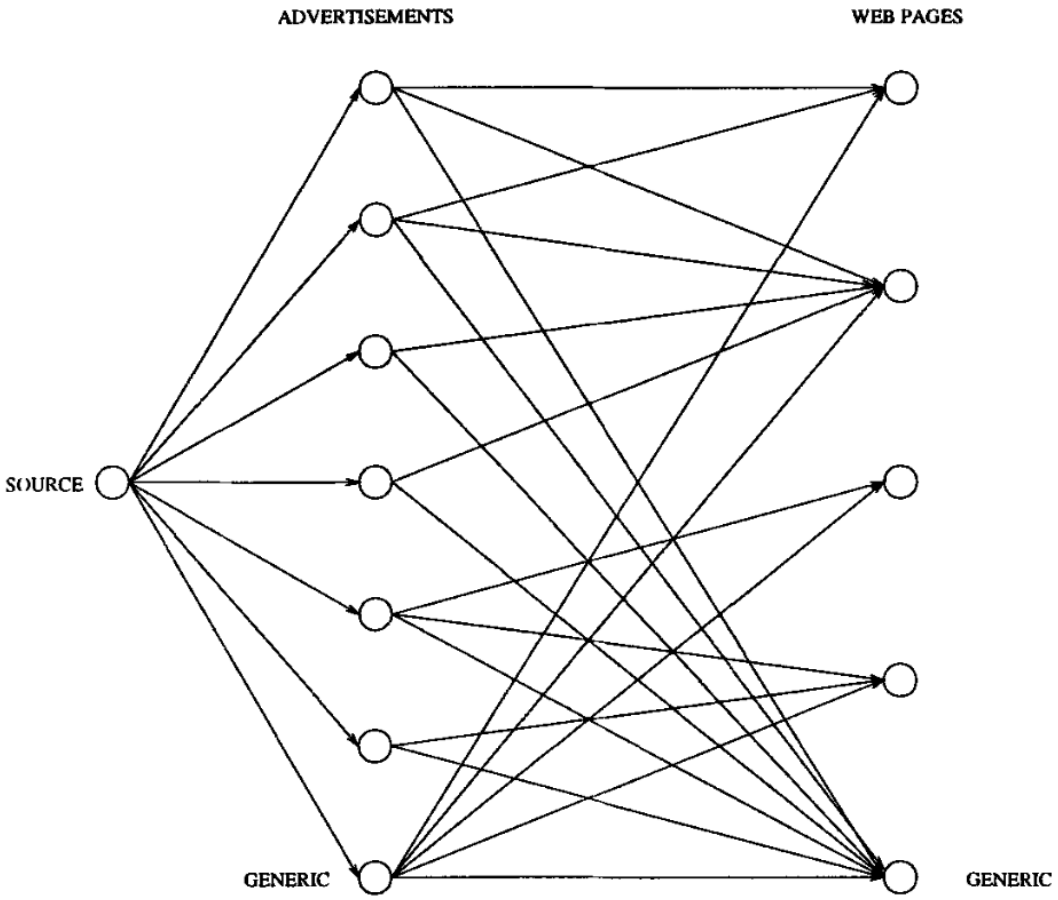


Figura 89. Modelo de asignación simple de anuncios a páginas web (Aggarwal, et al., 1998).

Para crear campañas, según Bruner y Gluck (2006), DoubleClick recomendaba que se haga uso de las cuatro siguientes indicaciones:

- Establecer objetivos claros.
- Segmentar las audiencias para identificar cuáles responderán mejor ante los anuncios.
- Optimizar las creatividades y los medios iterativamente con lo que se aprenda de los resultados.
- Revisar y evaluar en qué medida ha cumplido la campaña con los objetivos, para aprender qué se podría mejorar en el futuro.

Para ello, en su estudio sobre la efectividad de la optimización de los anuncios, Bruner y Gluck (2006) llegaron a la conclusión de que hay que tener en cuenta el alcance y la frecuencia de los anuncios, que tienden a seguir la siguiente distribución:

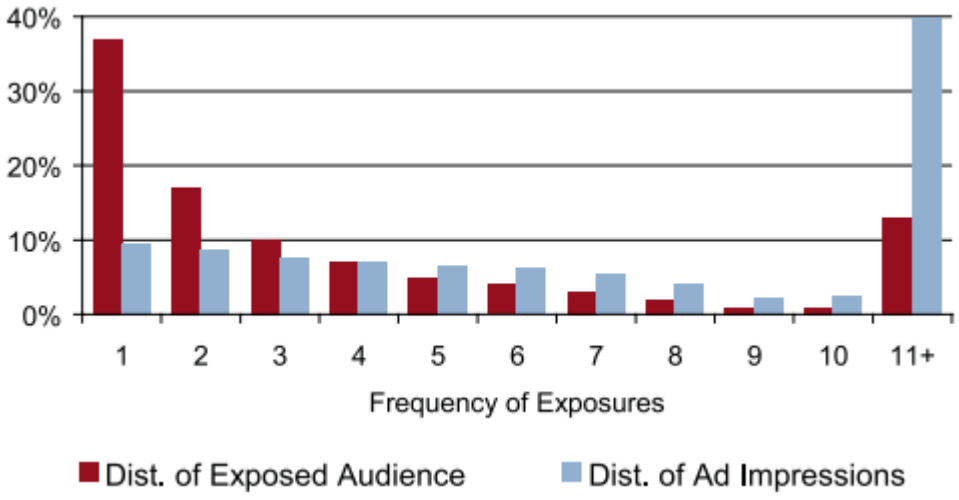


Figura 90. Distribución de audiencia expuesta a los anuncios e impresiones de éstos según el número de exposiciones (Bruner & Gluck, 2006).

En la Figura 90 se puede observar que, a mayor frecuencia de exposiciones, disminuye el porcentaje de la audiencia que ve los anuncios. Por otro lado, el tamaño de los anuncios también influye, ya que los tamaños más grandes obtienen una ratio de clics (CTR) y una intención de compra mayores, como se puede ver en las siguientes figuras:

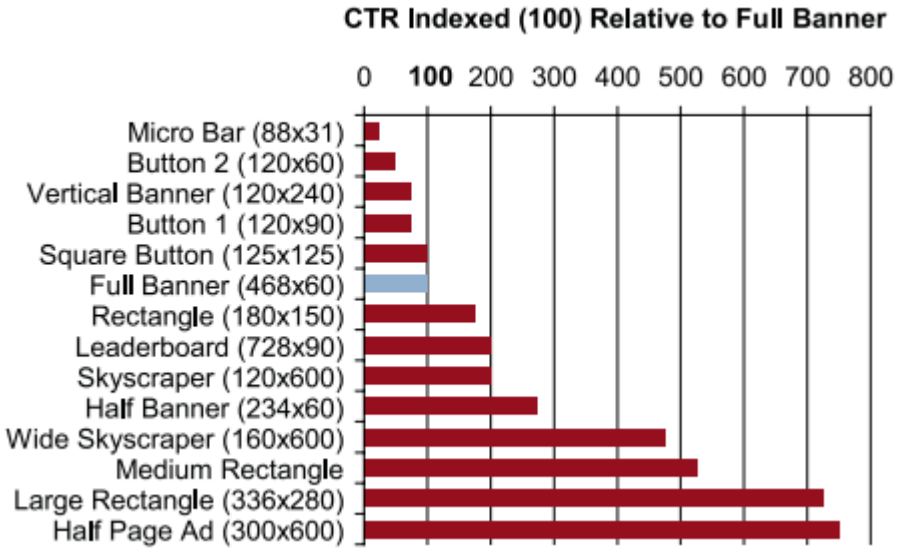


Figura 91. CTR según el tamaño de los anuncios (Bruner & Gluck, 2006).

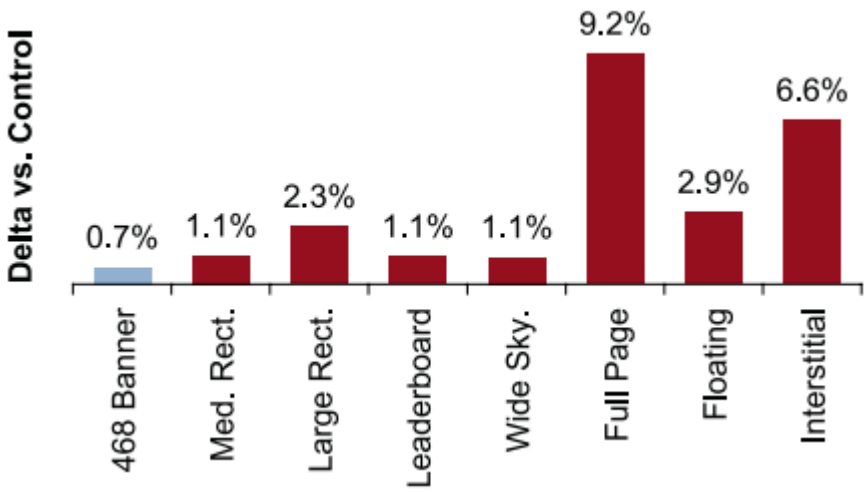


Figura 92. Intención de compra según el tamaño de los anuncios (Bruner & Gluck, 2006).

Por otro lado, Bruner y Gluck (2006) también observaron que la tipología de los anuncios es un factor determinante en el CTR y la intención de compra provocada por éstos. En los gráficos X y X se puede observar que los de tipo intersticio o flotante tienen un CTR más de diez veces mayor al de otros tipos como la imagen estática o incluso los anuncios expandibles. Del mismo modo, los anuncios de imagen estática (JPG) o dinámica (GIF) fueron los que peor intención de compra obtuvieron, seguidos de los anuncios de medios enriquecidos (servidos por proveedores especializados) y, superados por éstos, los de vídeo:

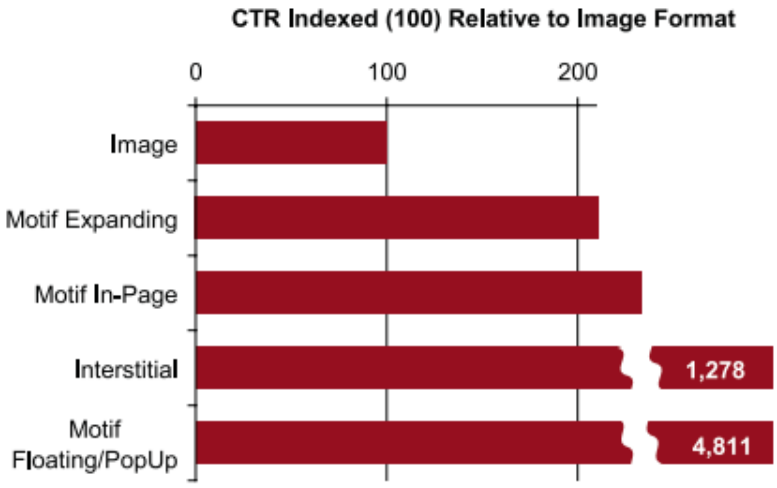


Figura 93. CTR según el tipo de anuncio (Bruner & Gluck, 2006).

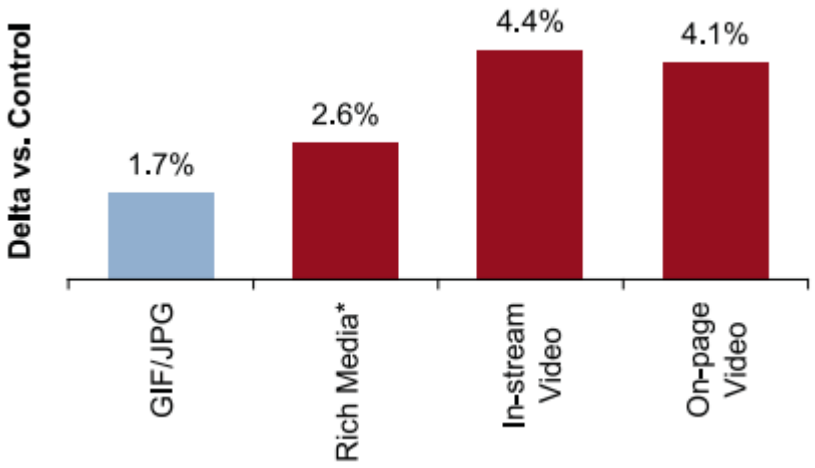


Figura 94. Intención de compra según el formato de anuncio (Bruner & Gluck, 2006).

Algunas de estas características se corroboraron en otro estudio, de Robinson et al. (2007), en el que tras analizar una muestra de 209 banners llegaron a la conclusión de que los parámetros que potenciaban la efectividad son: un mayor tamaño, la ausencia de incentivos promocionales y la presencia de información del sector al que se dirigen, en su caso el del juego online. Sin embargo, otras características como la animación, el uso de frases de acción o la presencia de la marca o logo no provocaron un mejor CTR. De hecho, al contrario de lo que los investigadores esperaban, comprobaron que los mensajes largos conseguían un mejor CTR.

El modelo CPC también es conocido como PPC (pago por clic), y fue inicialmente creado para medir la publicidad en los resultados de búsqueda, pero ha acabado aplicándose a todo tipo de anuncios web, según Shanahan y Kurra (2011). Reduce el riesgo para el anunciante, ya que este no debe pagar por impresiones que no estén creando un valor directo, como es la obtención de tráfico hacia el enlace del anuncio. Este riesgo, en cambio, lo comparte con el dueño de la web, que debe asumir sin ingresos la publicación de anuncios sin clic.

Hay otros modelos para anuncios como el PPI (pago por impresión) y el PPA (pago por acción) (Wu, et al., 2013), pero Villuendas Solsona (2018) aclara que en la publicidad digital se suele hablar de:

a) CPL (coste por cliente, *Cost Per Lead*), que consiste en un contrato mediante el cual el anunciante paga por cada usuario que cliquee en el anuncio y efectúe una acción concreta, como puede ser el registro o la suscripción a un boletín. La intención es recopilar datos de los usuarios para convertirlos posteriormente en clientes.

b) CPA (coste por adquisición o compra, *Cost Per Adquisition*), con el que el anunciante solo paga por aquellos usuarios que cliqueen en el anuncio y realicen posteriormente una compra.

c) Pago fijo cada cierto tiempo, en el que el anunciante paga un precio fijo por el anuncio durante un periodo de tiempo determinado y en una web entera o un conjunto de páginas concreto.

Un modelo basado en CPC (coste por clic) depende directamente de la CTR (ratio de clic). Sin embargo, el CTR no se puede saber a priori, por lo que según Li et al. (2010) los sistemas de anuncios deben usar un histórico de CTR para tomar las decisiones iniciales. Otro problema que sucede es que los anuncios con más impresiones y un mayor CTR se mostrarán más a menudo, con lo cual obtendrán más impresiones (y quizá más clics) en el futuro. Lo mismo ocurre de manera inversa en los anuncios con menos impresiones y un menor CTR, aunque también puede ocurrir con anuncios con pocas impresiones y un CTR alto, ya que se puede considerar que no han tenido buenos resultados y en realidad es debido a un histórico muy breve. Es por ello por lo que hay estudios como el de Li et al. (2010) que proponen que no solo se explote aquello que ya ha demostrado ser eficiente, sino que también se aporte recursos para explorar nuevas opciones que puedan suponer el descubrimiento de mejores anuncios a largo plazo.

Pese a que no hay un límite virtualmente en la cantidad de información que se puede incluir en Internet, sí que hay un límite en el tamaño de las pantallas, lo cual desemboca en un paradigma de desafío tanto para diseñadores web como para la publicidad web (Schumann & Thorson, 2010, p. 163), por lo que el tamaño de los anuncios adquiere relevancia de cara a su efectividad. Google Adsense, uno de los sistemas de anuncios más conocidos, aseguró que los tamaños que mejor funcionan no solo son grandes, sino también anchos, siendo los tamaños que mejor funcionaban el de 336x280, 300x250 y 160x600. Sigel et al. (2008) hicieron un estudio con banners de tres tamaños diferentes: 160x600, 728x90 y 300x250, y los mostraron en dos webs con 10 millones de impresiones en total. Comprobaron que el tamaño 160x600 tuvo un CTR mayor en ambos sitios web, y tuvo el mayor grado de interacción en uno de ellos, mientras que en el otro lo consiguió el formato 300x250.

No solo es importante el tamaño, sino también la posición o localización donde se coloque el anuncio dentro de la página web. Se ha estudiado en muchas ocasiones la relevancia del anuncio con respecto al contenido global de la página, pero no tanto la relevancia del anuncio con respecto al contenido que se visualice cerca de este. Wu et al. (2013)

propusieron un enfoque en el cual tener en cuenta el contexto global y local del anuncio. De esta manera, obtuvieron una mayor exactitud a la hora de elegir los anuncios y una mayor estabilidad semántica, ya que el contenido y los anuncios circundantes tenían una relación más fuerte. Generalmente, según Davis (2006, p. 21), los anuncios que se pueden ver sin necesidad de desplazarse por la página web (*above the fold*) funcionan mejor que los anuncios situados más abajo, un factor todavía más para tener en cuenta en pantallas más pequeñas, como monitores antiguos de resoluciones de 800x600.

Otra recomendación es la de Li et al. (2010) de aprovechar los espacios en blanco que existen en el 90% de las páginas web para incluir anuncios. Aportaron un sistema que automáticamente detecta lugares que podrían incluir anuncios relacionados semánticamente con la página web y el estilo de esta. Esta necesidad surge debido a que muchos sistemas de anuncios web, como el más popular de todos Google AdSense, obligan a los editores a colocar los anuncios en lugares concretos de la web, cambiando para ello la estructura original de las páginas y definiendo esa posición y estilo manualmente.

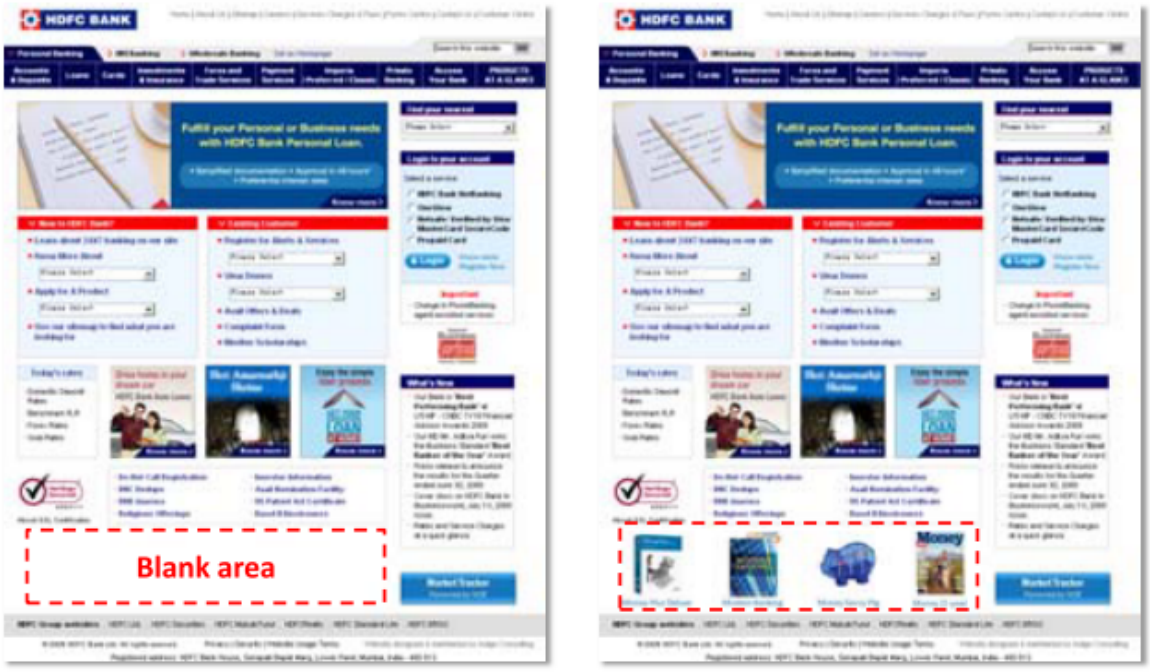


Figura 95. Colocación de anuncios en lugares vacíos de una página web (Li, et al., 2010).

Para ello, Li et al. (2010) debieron comprobar el porcentaje de webs que tenían un espacio vacío en el cual introducir cada formato de tamaño de Google AdSense de una manera no intrusiva. Se puede observar en las siguientes gráficas que el número de webs con ese espacio en blanco disponible disminuye conforme el tamaño de este aumenta:

Ad type	Format	Size
Half Banner	234x60	14,040
Button	125x125	15,625
Small Rectangle	180x150	27,000
Banner	468 x 60	28,080
Vertical Banner	120 x 240	28,800
Small Square	200 x 200	40,000
Square	250 x 250	62,500
Leaderboard	728 x 90	65,520
Skyscraper	120x600	72,000
Medium Rectangle	300 x 250	75,000
Large Rectangle	336 x 280	94,080
Wide Skyscraper	160x600	96,000

Figura 96. Formatos de anuncio por tamaño de Google AdSense (Li, et al., 2010).

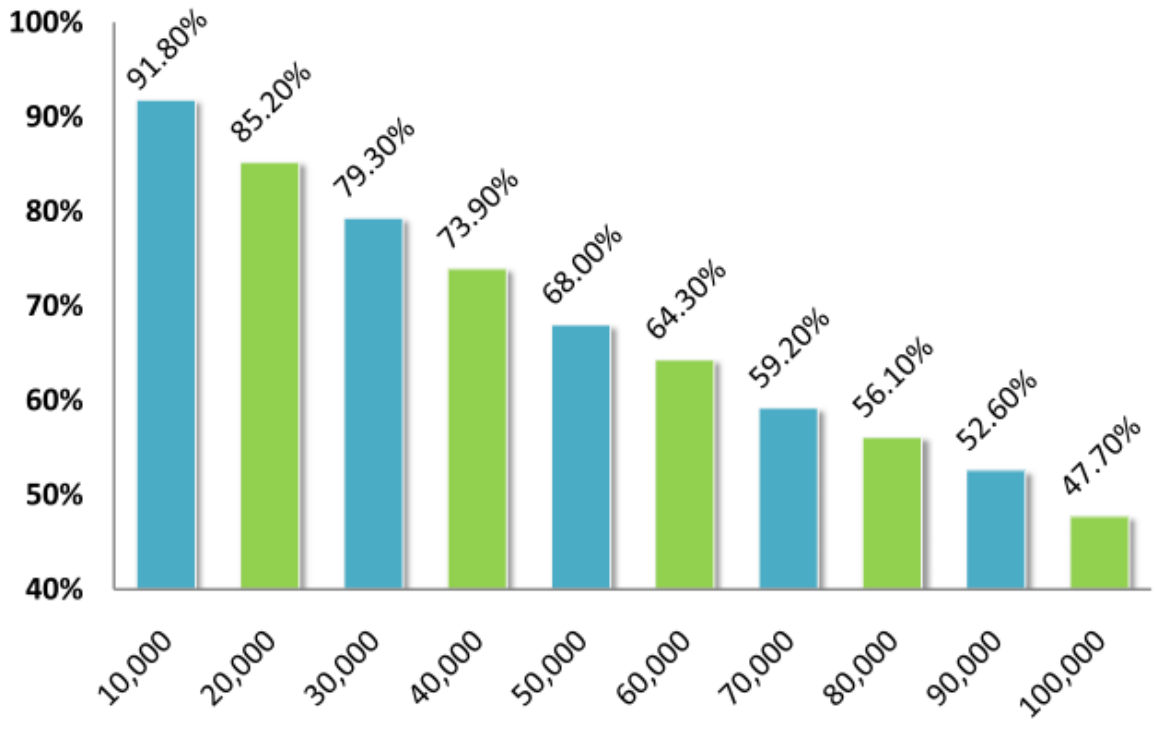


Figura 97. Porcentaje de webs con espacio en blanco, en franjas de 10.000 píxeles (Li, et al., 2010).

Según Farhan y Yousaf (2016), los cinco factores que ayudan a que los anuncios web sean más atractivos para los posibles consumidores son: la localización, la presentación, el contenido, la presencia de una celebridad y la duración, siendo la localización el factor más

importante de todos, seguido de la presentación según el tamaño de la fuente, el estilo de la fuente, el color, etc. El contenido debe tener la información suficiente pero también ser humorístico y llamar a una respuesta por parte del usuario. Las celebridades aumentan la efectividad, y la duración debe ser la suficiente para que el usuario le preste atención.

Los métodos de medición de la efectividad de la publicidad como el CPM y el CTR no tienen en cuenta las características únicas de interacción del medio ni cómo los consumidores procesan los estímulos que les provocan los anuncios. Es por ello por lo que Chatterjee (2001, p. 209) propuso otras características para tener en cuenta como son: el tipo de exposición, el tipo de atención por parte del usuario, los procesos de comunicación que provocan y los resultados de comportamiento en el usuario. Wolin et al. (2002) añadieron factores relacionados con las creencias del usuario en cuanto a la información del producto, el placer hedonista y la imagen y rol social. Características como el nivel de confianza, su posición frente al materialismo e incluso su nivel de educación y de ingresos pueden influir negativamente en su actitud frente a la publicidad web. Según Mahmoud (2014), estos factores se repiten en su estudio de los consumidores sirios, ya que comprobó que las creencias sobre la publicidad digital eran multidimensionales teniendo en cuenta el valor informacional, de entretenimiento, rol social, materialismo, confianza, corrupción e irritación.

Zhu et al. (2010) comprobaron que los clics estaban sesgados debido a que también dependen del orden de presentación, la reputación de la web, el navegador del usuario, la hora local del clic, etc. Por ello, propusieron un modelo de clic llamado General Click Model (GCM), basado en una red bayesiana que emplea un método de Propagación de Expectación, realizando una inferencia bayesiana aproximada. Permite ventajas de aprendizaje y una mejor generalización, lo cual los llevó a obtener mejores resultados, especialmente para consultas de cola. Se asume que los usuarios escanean las URLs de arriba hacia abajo, y lo hacen de manera que se sigue el gráfico de la Figura 98. También argumentaron que la mayoría de los modelos existentes en 2010 podrían ser reducidos a GCM con una asignación de parámetros diferentes según el caso.

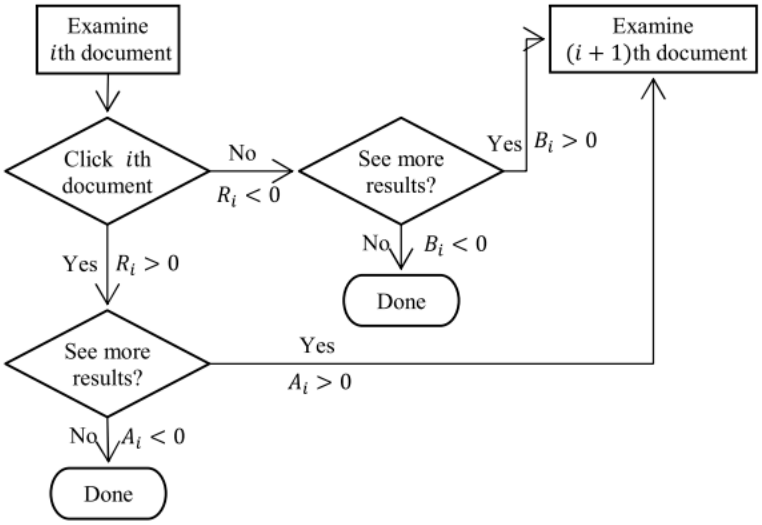


Figura 98. Gráfico de usuario del General Click Model propuesto por Zhu et al. (2010)

La literatura ha tratado en varias ocasiones la relación de la publicidad digital con la memoria. Yoo (2008) argumentó en un estudio que los consumidores que habían visto anuncios web experimentaban un aumento en la memoria implícita y una actitud más favorable hacia la marca promocionada, sin tener en cuenta el nivel de atención que le dieron a esos anuncios. Otro detalle muy interesante que observó fue que los anuncios también afectan al subconsciente de los usuarios, ya que, aunque estos no recordaran haberlos visto, sí que se mostraban más partidarios de tener en cuenta aquello que se promocionaba en ellos. Esto, según el autor, se debe a que los usuarios se sirven de una mezcla de información que procesan consciente y subconscientemente. Esta conclusión se vio ratificada en otro estudio realizado por Yoo (2009), en el que observó que, aunque los usuarios no presten atención de manera directa a los anuncios, estos obtenían beneficios a nivel de marca de forma subconsciente, siempre y cuando los anuncios fueran congruentes con el contenido de la página web. Todo ello apoya la noción de que la métrica de impresiones también importa, más allá del CTR en sí.

En cuanto a los detalles del anuncio que favorecen la memoria, Kong et al. (2019) analizaron si el recuerdo del anuncio se veía reforzado según si se utilizaba una imagen, un texto y el precio del producto mediante el seguimiento de la mirada y otras medidas para la memoria. Llegaron a la conclusión de que la mejor combinación era utilizar los tres a la vez, seguida de utilizar el texto y el precio. Además, comprobaron que el género del usuario afecta también a este proceso, ya que los hombres recordaban mejor los anuncios sin imágenes. La intención de la navegación es otro factor para tener en cuenta, puesto que, si es la de recolectar información, el enfoque ayuda a la memoria también de los anuncios presentes junto a la información.

Los usuarios, además, interpretan de manera diferente un anuncio web diseñado para un producto que el de una marca, según un estudio de van Steenburg (2012). La necesidad de conocimiento afecta a la memorización del anuncio, por lo que un usuario que esté dispuesto a un esfuerzo cognitivo mayor para procesar información les dará una mayor atención a los banners de producto, mientras que los que no estén dispuestos prestarán una mayor atención a los banners de marca. Esto puede ayudar a la selección de qué anuncios mostrar en cada tipo de página, ya que páginas como la de inicio o una secundaria podrían mostrar un mayor beneficio en publicidad si incluyen anuncios de marca, mientras que los anuncios de producto podrían ser más productivos para usuarios que han navegado en más páginas de la misma web.

Según Ka Po y Kong (2006), los factores que influyen son la percepción del usuario hacia la publicidad; el contenido del anuncio, así como la idea que comunica y la involucración del usuario con esta; y la marca en sí. Curiosamente, propuso como factor las emociones que suscita el anuncio en el usuario, y esa hipótesis no se vio validada en su estudio. Las anteriores características buscan influenciar la actitud del usuario hacia la publicidad de cara a una intención de compra, como se puede ver en el siguiente gráfico:

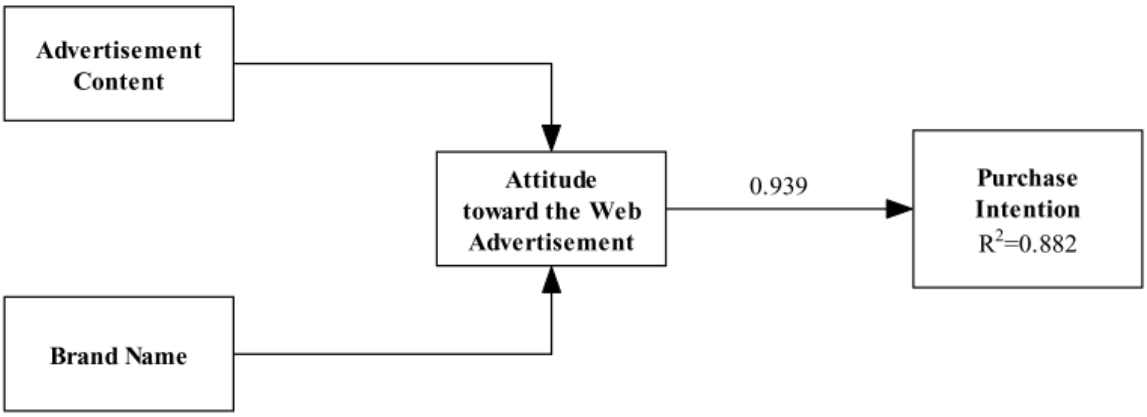


Figura 99. Modelo de factores que fomentan la intención de compra (Ka Po, 2006).

Para Rifon (2002), el nivel de confianza en sí es un factor vital para la efectividad de la publicidad, ya sea por la propia credibilidad de la web, la relevancia del anuncio con respecto al contenido de la web, la credibilidad del anunciante según sus anuncios y su marca, así como las intenciones de compra. Las tarifas publicitarias, en cambio, no tienen porqué corresponderse con la credibilidad ya que no están directamente relacionadas. Las tarifas publicitarias dependen más de variables como el número de visitantes y el CTR.

La frecuencia en la que deberían mostrarse los anuncios es un factor importante para optimizar las ratios de clic o la recaudación, según Tomlin (2000), por lo que este aportó

un modelo mediante el cual realizar esta optimización, buscando también la evitación de situaciones en las que los anuncios se muestran a un grupo demasiado ajustado de usuarios debido a ecuaciones excesivamente específicas o restrictivas. Esto lo consigue mediante la aleatorización teniendo en cuenta también la probabilidad de CTR. En un estudio de Lee (2010) se observó que si se incluyen anuncios con una gran frecuencia de exposición, empeora la capacidad de reconocer los anuncios por parte de los usuarios, aunque no provoca efectos negativos en la memoria, la actitud y el comportamiento de los usuarios. En otro estudio de Hussain et al. (2018), se llegó a la conclusión de que la frecuencia es un factor para tener en cuenta para anuncios estáticos o con un enfoque emocional, mientras que es mejor disminuirla cuando se busca aumentar el alcance mediante ventanas emergentes o un enfoque más racional.

Langheinrich et al. (1999) comprobaron que los anunciantes también necesitaron aprovechar la adaptabilidad de la red para aprovechar la identificación del usuario y sus características y hábitos online. Para estudiarlo, dividieron los enfoques de la publicidad digital en este sentido en cuatro categorías: generalizada, sin enfocarse a un público objetivo en concreto; editorial, centrándose en una web o temática; dirigida, con un público objetivo según las características propias del usuario; y personalizada, según la navegación y las interacciones del usuario con la web.

Según Cao (1999), la efectividad de la publicidad web se puede estudiar según los diferentes actores que participan en la misma: los consumidores, la publicidad en sí, el producto o servicio anunciado, el medio y el entorno. Estos actores influyen en un proceso cuya medición presenta los siguientes problemas: la especificación de objetivos, la adquisición de datos, el desarrollo del modelo de datos y los procesos analíticos. Los principales beneficios que buscar mediante la optimización de la publicidad digital son: un coste bajo, una respuesta rápida, la unificación de todo tipo de requerimientos, la extensibilidad y la reusabilidad.

Desde el punto de vista del dueño de la página web, el desafío está en conseguir que la navegación del usuario finalice en un comportamiento de cliqueo, aseveran Hofacker y Murphy (2000). Esto se produce mediante tres fases en las que el usuario presta atención al banner, experimenta interés o curiosidad por el contenido de este y lo cliquea. El problema es que cuantos más enlaces haya presentes en la página web, el usuario le prestará menos atención a cada uno, por lo que resulta necesario optimizar la localización de los banners en la página web. Hofacker y Murphy (2000) realizaron un estudio variando la frecuencia y localización de los banners en una o varias páginas, y comprobaron que ambos factores, así como el tamaño, contenido, animación y color, dependen del objetivo

del banner, puesto que si el banner está pensado para ser cobrado mediante impresiones, busca una ganancia acumulada con la visualización repetida, mientras que si es por clics, la localización no será tan crucial ya que, al efectuarse el pago solo si se cliquea, solo afectará al volumen del tráfico obtenido y no a su coste con respecto a las páginas vista.

Es por ello por lo que los responsables de los sitios web deberían escoger con cuidado la inclusión de la publicidad sin romper la congruencia entre el banner y el contenido de la web, puesto que, según Newman et al. (2004), si se realiza de manera correcta puede beneficiar la efectividad de la publicidad y proveer de mayores ingresos. En el caso de hacerse incorrectamente, puesto que el banner y la página web son incongruentes, podría perjudicar a la marca del sitio web e incluso también a la marca del anunciante. La irritación producida por la publicidad afecta a los resultados de esta, según han podido comprobar Saadeghvaziri et al. (2013) en su estudio sobre la publicidad web. La competición por la atención del usuario puede ser molesta bien porque suscite fastidio, ofensa o incluso la sensación de manipulación. Por ello es recomendable invertir tiempo y dinero en fomentar creencias y emociones que sean positivas, de manera que se produzcan actitudes positivas y un comportamiento favorable de consumo.

Schumann et al. (2014) comentan que es también habitual informar de la razón por la que se incluye publicidad en medios digitales. Pero la justificación es más efectiva si va acompañada de una comunicación proactiva de que se trata de una relación recíproca. Es, por tanto, recomendable también recordar a los usuarios el carácter gratuito de los servicios o contenidos que hay en su web a la hora pedir permiso para utilizar una publicidad dirigida o comportamental con ellos.

Chatterjee et al. (2003) modelaron el flujo de clics de los usuarios en los anuncios de una web en la que era obligatorio el registro para ver los contenidos. Observaron que la variabilidad en los clics no era tan significativa entre usuarios diferentes, pero sí entre sesiones diferentes del mismo usuario. Los primeros anuncios que se visualizan en una sesión tienen mayores probabilidades de ser cliqueados, mientras que conforme avanza la sesión, se produce un efecto negativo y no lineal. Las características que aumentan la probabilidad de clic en los anuncios en la sesión actual, estudiando diferentes sesiones, fueron las siguientes: intervalos de tiempo mayores entre sesiones, más exposiciones a los anuncios en sesiones anteriores y más tiempo desde el último clic. Sin embargo, la probabilidad de clic descendía con el aumento del número de sesiones del mismo usuario.

Para potenciar el interés o curiosidad por el contenido de la publicidad web, Menon y Soman (2002) consideran conveniente utilizar una estrategia basada en: la generación de

curiosidad destacando el desconocimiento de algo, la utilización de pistas que guíen al usuario hacia la resolución de la curiosidad, el tiempo suficiente para tratar de resolver la curiosidad, dar confianza en que se ofrecerá la información que la resuelva, y el uso de técnicas que ayuden a la elaboración y aprendizaje de los usuarios, más allá del mero clic en el anuncio, y así poder medir la efectividad de la publicidad.

Todo ello debe de tener en cuenta también el nivel de intromisión de la publicidad en el flujo de navegación, según Parsons et al. (2000). Es cierto que la inclusión de publicidad no tiene por qué provocar que el usuario consuma menos contenido en el medio, de hecho, puede ser un punto a favor para este especialmente si la publicidad tiene una relación fuerte con el contenido en sí. Sin embargo, si es demasiado intrusiva, empeora la efectividad de la publicidad en sí, ya que provoca irritación y molestia al visualizar el contenido, especialmente si este tiene características activas como parpadeo, ocupación del área activa y animaciones, añaden Lewandowska (Tomaszewska) & Jankowski (2017). Estas últimas no parecen invasivas a no ser que vayan acompañadas de una intensidad o parpadeo de gran frecuencia.

Por otro lado, la facilidad de segmentación de la publicidad también ha suscitado la posibilidad de introducir banners muy especializados y otros más genéricos. Sin embargo, se ha observado un descenso generalizado del CTR con el paso del tiempo, por lo que la tendencia ya en 2004 era hacia la inclusión de la publicidad en un número cada vez menor de webs de una mayor audiencia (Rausell Köster, 2004).

La segmentación se puede realizar basándose en el concepto de *Behavioral Targeting*, una técnica que según Yan et al. (2009) mejora la efectividad de las campañas investigando el comportamiento de los usuarios. Teniendo en cuenta esto, en su estudio dicho autor comprobó que los usuarios que cliquean el mismo anuncio tienden a comportarse de la misma manera en la web, y sus comportamientos influyen más si son a corto plazo que a largo plazo. Gracias a esta técnica, se puede obtener una mejora de hasta un 670% de los anuncios.

Para realizar la mejor selección de sitios web para una campaña de publicidad digital, Ngai (2003) realizó un estudio en el que aplicó un proceso de jerarquía analítica (AHP) mediante el cual evaluó los sitios web en base a cinco criterios: la ratio de impresiones, el coste mensual, la adecuación de la audiencia a la campaña, la calidad del contenido y la apariencia de este, haciendo especial énfasis en el diseño y la usabilidad.

2.2.5.3. Sistemas de anuncios web

Un sistema de anuncios web procesa las interacciones con los anuncios de las campañas de publicidad dadas de alta en su plataforma. Estas interacciones son principalmente las peticiones de selección de anuncios, las peticiones de los datos de cada anuncio y las redirecciones de las peticiones resultantes del clic del usuario. Todo ello basado, según Langheinrich et al. (1999), en métodos de aprendizaje automático con los que incrementar el valor del espacio de inventario, aumentar el control por parte de los anunciantes sobre los objetivos de sus anuncios al mismo tiempo que maximizar la efectividad de estos, y tratar de no molestar a los usuarios gracias a ofrecer banners personalizados y adaptados al contexto.

Un anunciante puede obtener de un sistema de anuncios web la posibilidad de pagar porque se incluya su banner en una serie de sitios web, además de optar por otros métodos alternativos como intercambiar banners con otras webs o pagar directamente a estos sitios web, especifican Amiri y Menon (2006) en su estudio. Puesto que los anuncios compiten por el mismo espacio en un sitio web, es importante realizar una programación efectiva de éstos, actualizando los anuncios de manera regular en intervalos cuya frecuencia supone una problemática que ha sido estudiada en varias ocasiones en la academia.

Kazienko y Adamski (2007) propusieron un sistema llamado AD ROSA como método para personalizar automáticamente los anuncios en las páginas de una web, respetando la privacidad del usuario al no guardar sus datos en una base de datos. Este sistema utiliza técnicas de minería web y calcula factores como el contenido más adecuado de la web, la probabilidad de clic, la política de publicidad y la prevención de aburrimiento, al mismo tiempo que permite la programación de los anuncios para el usuario.

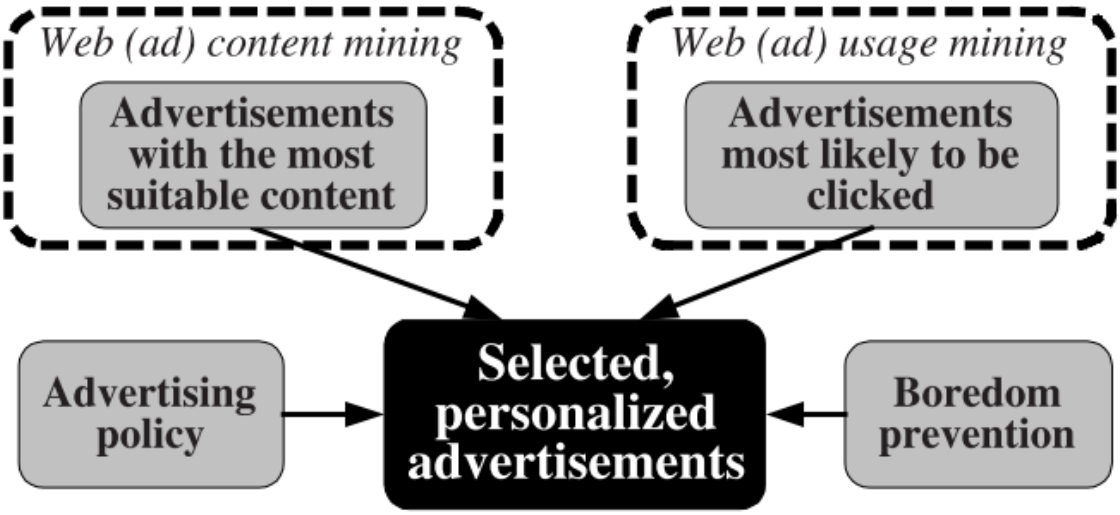


Figura 100. Esquema de técnicas involucradas en el sistema AD ROSA (Kazienko & Adamski, 2004).

Ribeiro-Neto et al. (2005) investigaron diferentes estrategias con las que relacionar los anuncios con páginas web. Cinco de ellas trataban de asignarlos a una página web de manera directa, obteniendo un aumento de precisión del 60%. Las otras cinco expandieron la relación semántica más allá de las propias palabras clave con nuevos términos, obteniendo un 50% extra de precisión con respecto a las anteriores. En suma, se observó que la publicidad dirigida al contenido de la propia página web era algo práctico y factible.

Los sistemas de anuncios web, responsables de hacer estas vinculaciones entre anuncios y páginas web, siguen un modelo descrito por Vratonjic et al. (2010) en el cual los anunciantes se suscriben al sistema que se encargará de publicar automáticamente anuncios en páginas web relacionadas. Estos anuncios son almacenados en servidores de anuncios. El sistema de anuncios tiene contratos con webs que quieren mostrar los anuncios en sus páginas. Cuando el usuario visita una web que contiene anuncios, su navegador descarga el contenido de la página y un código que se ejecuta en su máquina y se conecta a uno de los servidores de anuncios, realizándole una petición de un anuncio. El servidor escoge y le envía el anuncio para mostrar que más se ajusta a los intereses del usuario y las características del contenido, de manera que se pueda maximizar el potencial de ingresos de publicidad. Todo este modelo se puede observar en la Figura 101:

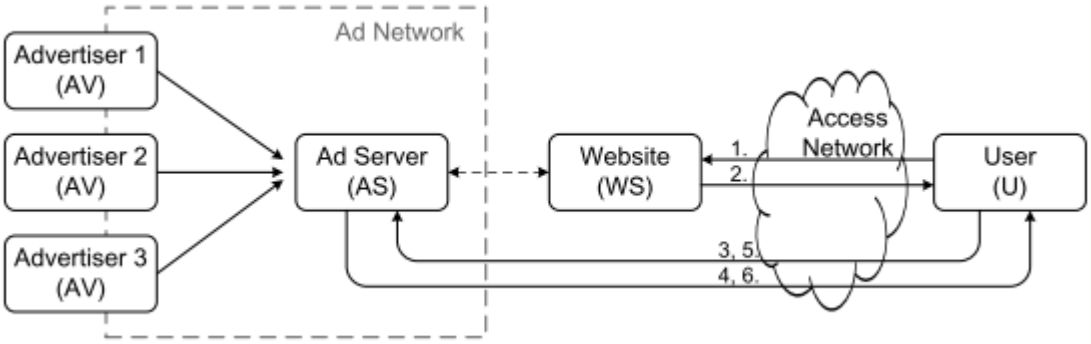


Figura 101. Modelo de arquitectura de un sistema de entrega de anuncios (Vratonjic, et al., 2010).

En el procedimiento anterior se destaca que el servidor de anuncios escoge el anuncio que visualizará el usuario en tiempo real. Para ello se sirve de las palabras clave que los anunciantes han elegido para su campaña, el contenido de cada página web e incluso el historial de navegación del usuario a través de las cookies, que son archivos temporales con información que se almacenan en el navegador del usuario.

Esto último se conoce como segmentación por comportamiento (*Behavioral Targeting, BT*). Chandramouli et al. (2012) aportaron además un modelo llamado TiMR en el que combinaron un sistema que procesa datos orientados al tiempo y un marco de reducción de grupos (M-R). Argumentaron que muchas peticiones analíticas suelen ser fundamentalmente temporales, por lo que su modelo permite peticiones temporales a una escala de conjuntos de datos de tamaño enorme y sin conexión. Obtuvieron así una mayor escalabilidad, un mejor rendimiento, un menor uso de memoria y tiempo de aprendizaje y un mejor CTR que otros modelos propuestos hasta entonces por la academia.

Debido a la gran variedad de necesidades de publicidad en Internet, surgieron redes de publicidad o *Ad Networks* que conectan a los anunciantes con los sitios web en los que poder incorporar anuncios, de forma que según Oliva Rodríguez (2014) actúan como intermediarios adaptándose a diversos tipos de negocio: según el contexto del contenido de la web, el comportamiento del usuario, los resultados de los anuncios sin importar la web donde se muestran, la venta directa, una temática específica o un inventario exclusivo.

Con el advenimiento de la Web 2.0, también llamada Web Social 2.0, en la que el sistema potenciaba sus características de interacción constante y mudable, se popularizó el uso de programas para anunciantes como Google AdWords[36], mediante los cuales se creaba campañas de publicidad web con banners interactivos tanto en páginas webs como en los principales buscadores (San Millán Fernández, et al., 2008). Google AdWords es un

[36] https://ads.google.com/intl/es_ES/home/

programa creado en 2000 por Google y el hecho de que se adecuara a las necesidades puntuales de los usuarios ha permitido a la publicidad online, según Arroyo Valencia y Ceron Mayor (2009), diferenciarse de la tradicional, aprovechando el análisis en tiempo real de las campañas para llegar a la audiencia de una manera efectiva.

En abril de 2003, Google adquirió Oingo y lo renombró como Google AdSense, cambiando el panorama de los anuncios web ya que según Yuan et al. (2012) muchas otras empresas no tardaron en imitar el proyecto: Yahoo! Publish Network, Microsoft adCenter[37] y Advertising.com Sponsored Listings entre otros. Se adaptaron a un entorno de medios enriquecidos permitiendo tipologías de anuncios en formato de imagen, vídeo y audio con información geográfica. En 2016, Google AdSense seguía dominando foros y discusiones, siendo considerado por muchos como la manera más sencilla de ganar dinero online (Ekanem, 2016, p. 3).

La relación entre Google AdSense y Google AdWords es muy estrecha, explican Desnica et al. (2014). Mientras AdWords presta un servicio mediante el cual los anunciantes crean anuncios para el buscador de Google en sí y páginas web de Internet, AdSense los distribuye entre los editores de dichas webs para que sus usuarios los visualicen e interactúen con ellos. Una relación que se puede observar en la Figura 102:

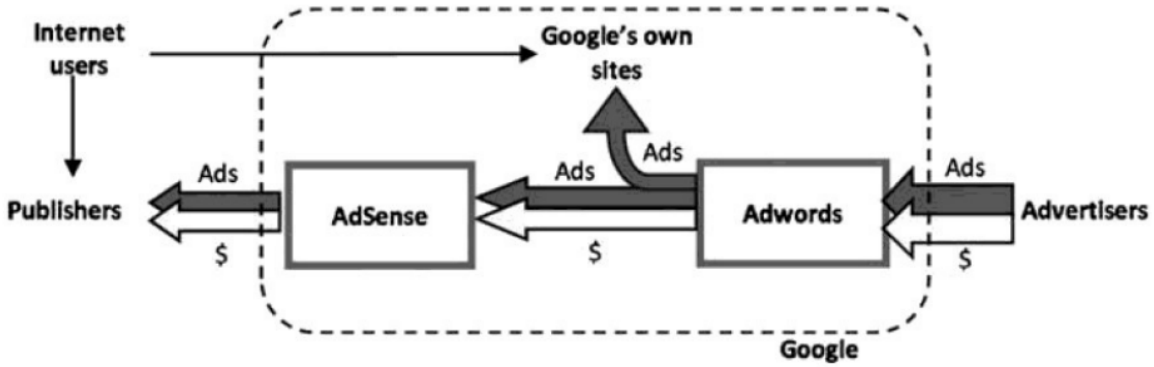


Figura 102. Diagrama de conexiones entre Google AdWords y AdSense (Desnica, et al., 2014).

En 2014, Google AdSense ya estaba presente en 4 millones de webs, dirigidas por 2 millones de editores. Desnica et al. (2014) indican que la plataforma permite una segmentación por el idioma principal de las webs, por geolocalización de los usuarios, por palabras clave del contenido, por intereses según el historial de navegación de los usuarios e incluso información relevante de su navegación en herramientas de Google como el buscador. A la hora de seleccionar los anuncios a mostrar en cada página y localización,

[37] https://ads.microsoft.com/

estos compiten según la relevancia con respecto al contenido, y el formato del anuncio en sí, normalmente elegido por el editor de la web, siendo los más comunes los de texto básico, imágenes, vídeos o banners flash.

Google AdSense utiliza, según Boone et al. (2010), la técnica de la publicidad dirigida a los usuarios según el contexto y la semántica del contenido de la página web visualizada. Las ratios de los anuncios se basan en la popularidad del enlace y el número de lectores que visualicen el anuncio. Google AdSense es, por tanto, la primera gran plataforma de publicidad contextual, y consiste según Wu et al. (2013) en las siguientes cuatro partes: el anunciante, el editor, la plataforma o sistema de anuncios y los usuarios. El anunciante se registra en el sistema de anuncios, el usuario navega por Internet y las páginas web incluyen contenido del editor y anuncios del sistema de anuncios que tenga instalado en su web.

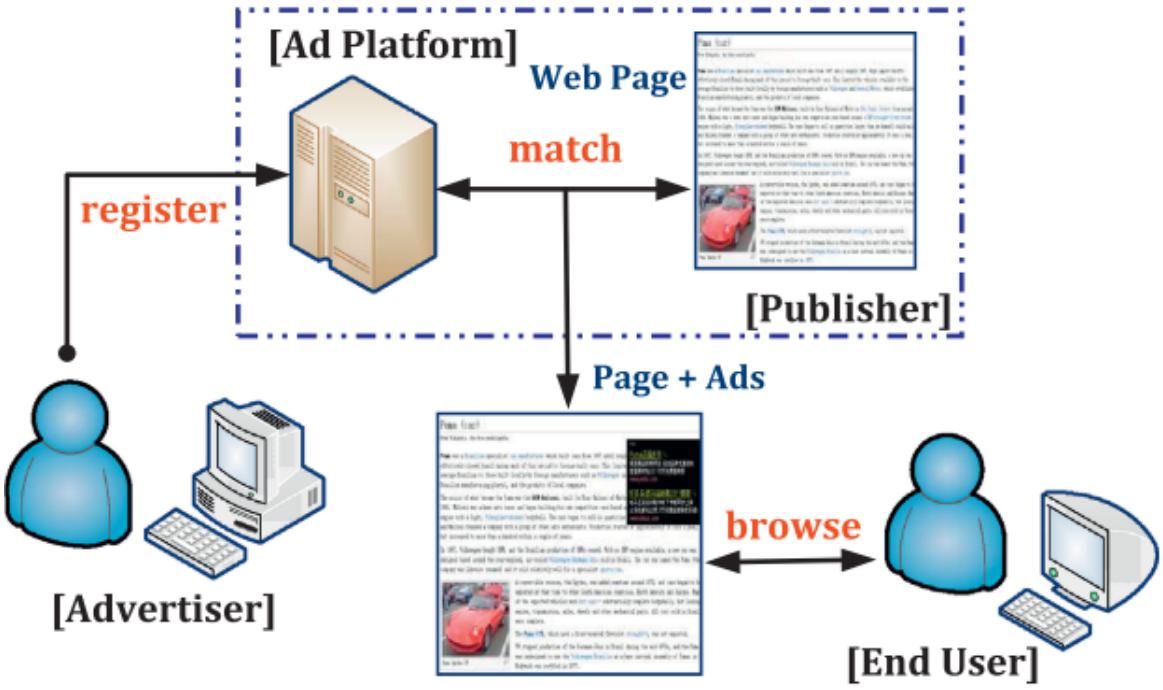


Figura 103. Plataforma de publicidad contextual (Wu, et al., 2013).

Sin embargo, Suber (2006) señala que los editores no saben a priori qué anuncios se pondrán en su web, ya que el algoritmo de Google es el responsable de la selección y muestra de la publicidad basándose en las palabras clave más relevantes de cada página. Lo bueno de que sea la publicidad la que se adapte al contenido y no al revés, es que así no se pone en tela de juicio la intención del contenido en sí, puesto que los editores no podrán tratar de manipular el mensaje para agradar a un anunciante en concreto. Google

AdSense tampoco toma decisiones editoriales ni de revisión por pares, por lo que se eliminan posibles conflictos de interés. La plataforma sí que ofrece la opción de bloquear anuncios de temáticas concretas, para evitar así que se visualicen anuncios que puedan provocar un conflicto en los lectores de la web.

En 2005, según Yuan et al. (2012), se sumaron a la lista de sistemas de anuncios web en tiempo real otras plataformas como ADSDAQ[38], AdECN, DoubleClick Advertising Exchange, adBrite y Right Media Exchange, pero con otro modelo: su objetivo era agregar múltiples sistemas de anuncios web a la vez de manera que se nivelara la oferta y demanda de cada mercado. Todo ello inició un nuevo ecosistema de la publicidad en Internet en el que cada rol aporta un bien a otros roles y recibe dinero a cambio. Los anunciantes pagan por incluir anuncios y cobran de aquello que venden a los usuarios. Los usuarios pagan a los anunciantes por lo que compran y aportan su atención a los contenidos de los editores. Estos ponen anuncios en sus webs tanto de anunciantes directos como de sistemas de anuncios y cobran por ello. Por último, los sistemas de anuncios cobran a los anunciantes los anuncios publicados, y pagan a los editores por ponerlos. Todo ello se puede observar en la siguiente gráfica:

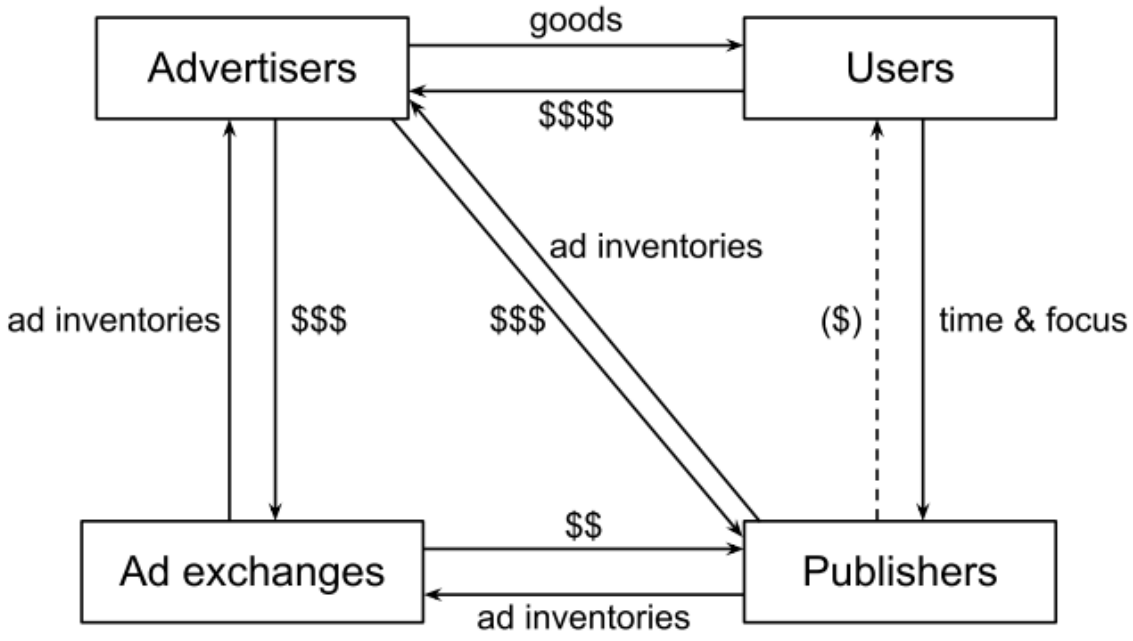


Figura 104. Ecosistema simplificado de la publicidad en Internet (Yuan, et al., 2012).

Fox et al. (2009) especifican en su estudio sobre los anuncios de AdSense que hay dos tipos de sistemas de anuncios o *Ad Servers*: los de soporte, que contabilizan a nivel

[38] http://adsdaq.digital/

particular las impresiones, los clics, el inventario vendido, la rotación de formatos, etc., en un sitio web en concreto; y las de anunciante, que miden los usuarios de una campaña, los datos de los diferentes soportes, la actividad post-clic y post-impresión, etc.

Los blogs, weblogs o como se le suele llamar a menudo, la blogosfera, se posicionaron rápidamente como una de las principales categorías de webs que apostaban por los anuncios de este tipo, explican Fox et al. (2009). Es por ello por lo que Google adquirió Blogger en 2003 y permitió que se pudieran incluir anuncios de Google AdSense en los blogs de esa plataforma. Mientras tanto, el número de blogs no hacía más que crecer: en 2006 ya había 50 millones en Internet, creciendo a un ritmo de 175 mil nuevos blogs al día. Ofreciendo beneficios por un esfuerzo leve, AdSense permitía a los editores centrarse en el esfuerzo de aumentar el tráfico y la calidad de sus contenidos. El objetivo: convertir contenido en dinero. No obstante, Google no tardó en ofrecer consejos para optimizar los anuncios ya que no todo se resumía en un mayor tráfico, sino también en cómo conseguir que los usuarios cliqueen más en los anuncios. Cambios de color, de tipología y de tamaño con la intención de retenerse más en la memoria del usuario y llamar a su atención. Fox et al. (2009) realizaron un estudio con anuncios AdSense en el que comprobaron que lo más importante era captar la atención utilizando contrastes altos y, en segundo lugar, colocar el anuncio lo más arriba posible.

Desnica et al. (2014) presentan algunos de los inconvenientes de Google AdSense: el precio de los anuncios fluctúa según el nivel de competencia entre los anunciantes, lo cual desestabiliza el flujo de ingresos. También ha habido quejas sobre la visualización de anuncios irrelevantes o repetición de los mismos anuncios hasta provocar irritación en los usuarios, consecuencia de que los editores no tengan un control directo en los anuncios que se mostrarán en sus páginas web. Por otro lado, la existencia de bloqueadores de anuncios en forma de extensiones de navegador y otros tipos de software, un tema del que se hablará en el siguiente apartado, hacen disminuir el número de usuarios a los que se alcanza con los anuncios.

Los beneficios de los sistemas de anuncios web suelen ser un porcentaje del precio por clic que los anunciantes pagan en las plataformas. *The New York Times* reportó que Google pagaba en 2006 a los editores el 78,5% de los ingresos de la publicidad (Comm, 2006). Según otro medio digital, a finales de 2013 se pagaba el 68% de los ingresos a los editores, siendo el resto considerado una comisión que cobra Google por su servicio (Maier, 2013). A la pregunta de si un blog puede generar ingresos suficientes respondieron Chen y Chen (2010, p. 10) en un estudio, en el que comprobaron que sí se puede, pero que la cantidad de beneficios depende de la originalidad, el profesionalismo, la variedad

del público objetivo y la frecuencia de actualización de los contenidos. Estos factores no solo generan un mayor CTR sino también un mayor tráfico, aunque en ese estudio no tomaron en cuenta la interrelación entre ambas variables.

Con la intención de unir el tráfico social con beneficios de Google AdSense, Aimiuwu et al. (2012) realizaron un estudio empírico y exploratorio en el cual observaron que, en el caso de Facebook, sí que ayuda a aumentar el tráfico hacia una web siempre que el contenido tenga una gran tendencia de búsqueda, mientras que aumenta los beneficios de anuncios AdSense independientemente de la tendencia del contenido.

2.2.5.4. El fraude y el bloqueo de anuncios en la publicidad web

Ya en 1999 existían formas de abusar de los mecanismos publicitarios de manera que se produjeran de manera fraudulenta los eventos por los cuales se cobraba por incluir publicidad en una web. Anupam et al. (1999) introdujeron en la academia una metodología mediante la cual provocar un ataque de inflación de clics en los anuncios de una web, virtualmente indetectable en su momento y muy efectivo a la hora de que los clics de referencia fueran contabilizados. La intención de los investigadores era la de llamar la atención sobre la debilidad de esta forma de rentabilizar los anuncios y su migración a otras como el pago por venta o por conversión.

Vratonjic et al. (2010) han comprobado que el fraude de clics ha aumentado con el paso de los años, pero también han surgido otros tipos de fraude relacionados con los anuncios web. Es el caso de los sistemas infectados que realizan modificaciones en los anuncios a la hora de servirlos al público, alterando incluso los que aparecen junto a los resultados de búsqueda de Google. Esta inyección de anuncios se puede producir mediante un ataque de “contaminación”, es decir, inyectando una línea de código en otras webs de manera que se añada al código Javascript que se ejecuta y muestre anuncios que generen ingresos a quien haya perpetrado el ataque. La inyección de anuncios también se puede producir de una manera dirigida, es decir, infectar la máquina del usuario de manera que se muestren anuncios en los lugares con más CTR de una página o buscador. Otros tipos de acciones podrían ser el bloqueo de anuncios y la suplantación de identidad o *phishing* de manera que se consiga información de los usuarios que en realidad iba dirigida a otra web. Todos estos ataques pueden suponer un descenso en los ingresos de la publicidad y la pérdida de confianza de los usuarios.

Es por ello por lo que López Jiménez & Martínez López (2010) defienden que la publicidad comportamental, es decir, la que se adapta al comportamiento y características del usuario, debe ajustarse a las directrices de las normas del país y comunidad política a la que

pertenezca el medio, haciendo respetar la privacidad de datos de los usuarios y evitando cualquier vulnerabilidad de esta.

Los usuarios pueden en ocasiones bloquear las interrupciones de publicidad en el flujo de navegación mediante extensiones o programas, por lo que Yuan et al. (1998) consideran interesante plantearse en qué consiste la decisión de bloquear o no la publicidad web. La publicidad en sí puede ser buscada de manera directa por el usuario si desea realizar una compra o eliminar el coste de la información de tercera mano, buscando un beneficio que justifique el coste de la búsqueda de diferentes alternativas. Al fin y al cabo, argumentan Yuan et al. (1998), la publicidad directa aporta beneficios y costes al usuario. Los beneficios podrían ser el valor de la información e incluso el valor de entretenimiento, así como reducir el coste de búsqueda de más información para la compra de algo que le pueda interesar. Sin embargo, estos beneficios vienen con un coste de interrupción que puede perjudicar su experiencia de navegación, por lo que, si el coste de buscar una alternativa que bloquee los anuncios es menor al de tolerarlos, puede que la decisión se incline hacia la primera opción.

Anderson (2005) realizó un estudio al respecto y observó que el 38% de los usuarios de internet utilizaban algún bloqueador de publicidad web, por lo que la publicidad online ha ido perdiendo efectividad con el tiempo. Esto ha hecho más necesario todavía que tanto los anunciantes como los propietarios de webs sean más conscientes de la forma en que se incluyen los anuncios, de manera que traten de reducir la inconveniencia de éstos en la navegación.

En 2018, Miroglio et al. (2018) observaron que esta tendencia seguía en aumento, estimándose que 600 millones de dispositivos utilizaban un bloqueador de anuncios a finales de 2016, un 30% más que en 2015. Apreciaron también una diferencia en el comportamiento de los usuarios que tienen instalado un bloqueador de anuncios y los que no. Aquellos que bloquean los anuncios web pasan mucho más tiempo navegando (un 28% más) y consumen muchas más páginas vistas (un 15% más) que el grupo de control en el estudio realizado por Miroglio et al. (2018), por lo que concluyeron que el bloqueo de anuncios mejora la calidad de navegación de los usuarios, y les sale rentable a nivel de coste de instalación de la funcionalidad frente al coste de oportunidad de no visualizar los anuncios en las páginas web. Lo cierto es que, sin anuncios, el tiempo de carga, la navegación y la privacidad se ven beneficiados, sobre todo en los dispositivos móviles. Sin embargo, si se trata de su vía principal de ingresos, los responsables de los sitios webs pueden tratar de impedir el uso de bloqueadores de anuncios mediante la limitación de

acceso y la petición de que desactiven el bloqueador para su web en concreto, una funcionalidad que suelen facilitar las extensiones de navegador que bloquean anuncios.

2.3. Conclusiones del estado de la cuestión

Como se ha podido ver en el estado de la cuestión, especialmente en lo referente al análisis de tendencias en Twitter, se aprecia una carencia en la literatura a la hora de investigar la predicción del éxito de un contenido concreto. Hay muchos trabajos académicos sobre el análisis de tendencias y la detección de eventos, incluso con la intención de predecir tendencias. A la hora de estudiar el éxito, se suele realizar el análisis de la web completa, de la cuenta de la red social o de una publicación o serie de publicaciones en sí.

El factor innovador de la tesis reside en que la mayoría de las publicaciones consultadas hasta la fecha ha analizado el éxito online de una web completa gracias a herramientas de analítica web como Google Analytics y Piwik (Álvarez Intriago, et al., 2016). Sin embargo, no se han descubierto investigaciones sobre la medición desde el punto de vista de un contenido en concreto, y menos aún sobre la predicción de sus parámetros analíticos.

Tampoco se han encontrado investigaciones que integren dichos conocimientos en la elaboración de una metodología cibermétrica que optimice la estrategia de publicación de contenidos y su compartición en las redes sociales. Periódicos como El Confidencial sí que utilizan robots con informes automáticos de tráfico mediante los cuales estudian el comportamiento de navegación de la audiencia y tratan de detectar patrones, pero otros como eldiario.es no permiten el acceso a estas métricas a los periodistas, enviando semanal y mensualmente informes a los jefes de sección y directores. El objetivo en el caso de eldiario.es no es utilizar los datos para seleccionar las temáticas, sino para tener una visión general del resultado que ha obtenido el medio online (Tascón Gabella, 2018).

El interés de esta tesis es el de estudiar esta incógnita, diseñando una metodología que ayude en la toma de decisiones de la estrategia de contenidos de un medio digital. Una metodología que aúne dimensiones de análisis como son:

- La analítica web, ya que gracias a ella se obtendrá la información necesaria para analizar los contenidos dentro del marco de la web estudiada,
- La analítica de publicidad web, que aportará datos relacionados con la visualización e interacción con los anuncios integrados en la web,
- La analítica de tendencias en Twitter, que permitirá estudiar el consumo de las temáticas relacionadas con los contenidos en Twitter.

De esta manera, se tienen en cuenta estos ámbitos de aplicación a la hora de desarrollar una estrategia de contenidos para un proyecto digital. Se trata de predecir los indicadores

que se consideren de éxito de manera que se enriquezca así la toma de decisiones del equipo editorial.

Las ventajas de incorporar una metodología de estas características en el proceso editorial son numerosas, como la optimización de recursos a los que destinar la elaboración de un artículo, la elección de elegir o descartar un contenido, la optimización del contenido para atender a la tendencia a la que podría formar parte e incluso la optimización de la comunicación digital en las redes sociales, como es el caso de Twitter.

3. Metodología

3.1. Antecedentes y marco metodológico

A finales de 2009 se creó Hello Friki[39], un sitio web con contenidos informacionales online de entretenimiento, en el cual se pretendía dar cabida a las noticias más destacadas de ámbitos como el cine, las series de televisión y la literatura. La intención de los directores era de transformar en material productivo el tiempo empleado en el ocio, así como dar cabida al proceso creativo propio de la escritura periodística.

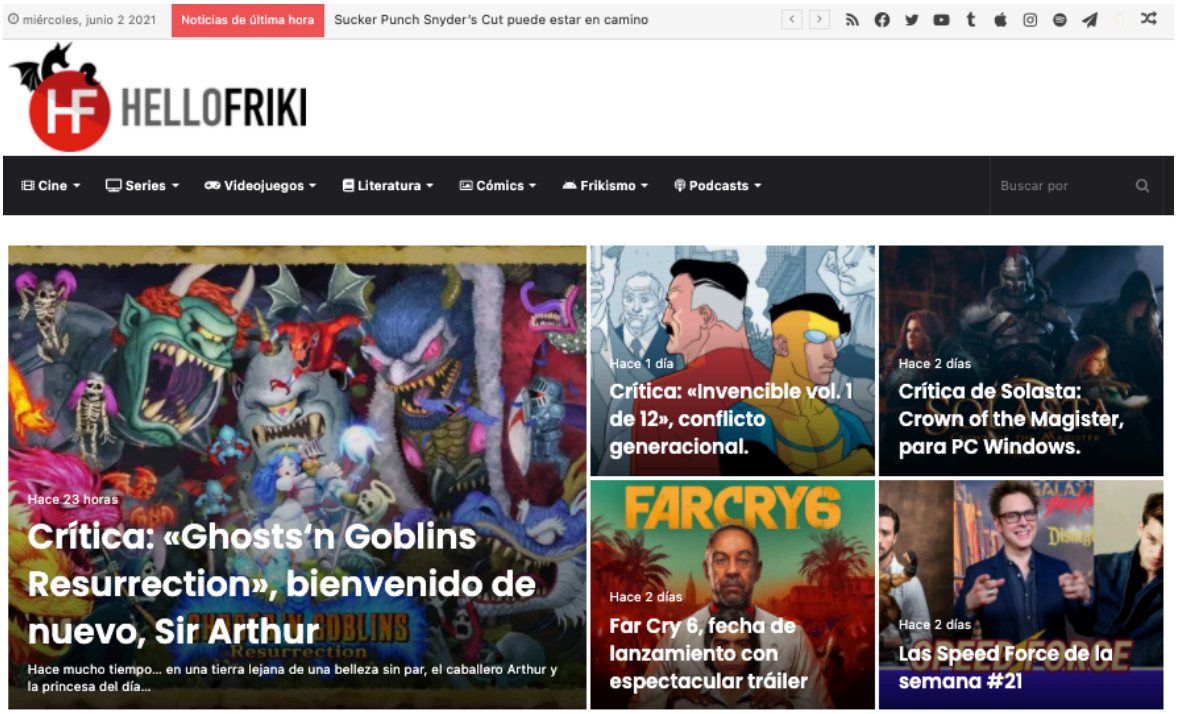


Figura 105. Captura de la página principal de la web de Hello Friki, realizada el día 2 de junio de 2021

Se trata de un medio de comunicación online cuya principal intención es solucionar la necesidad de información en unas temáticas que proveen diariamente de un amplio número de novedades en forma de noticias, así como la posibilidad de análisis, reportajes, entrevistas y muchos otros formatos de información. Todos estos contenidos se encuentran bajo el ámbito de las secciones de cine, series, videojuegos, literatura y cómics.

Además, se acompañó de un podcast o programa de radio online que daba cabida al análisis exhaustivo en forma de contenido audiovisual de las mismas temáticas, en un formato que no se suele encontrar en los programas convencionales de radio. De este modo se apuesta por un contenido que es tanto textual como audiovisual.

[39] https://www.hellofriki.com/

En un principio se basó en sistemas PHP de noticias como Cutenews[40], pero en 2012 se tomó la decisión de migrar la web a Drupal[41], un sistema de gestión de contenidos (*Content Management System*, CMS) modular multipropósito. Dicha migración dio lugar a la elaboración del trabajo final de la carrera Ingeniería en Informática del autor de esta tesis en la Universidad Politécnica de Valencia, titulado *Portal web de artículos on-line basado en Drupal*, leído el 27 de julio de 2012.

En este trabajo se describió de manera detallada el proceso de migración y desarrollo del sistema de contenidos resultante, así como los tipos de usuarios, contenidos, interfaz y diseño gráfico.

Esta decisión resultó no ser la óptima debido a que incrementó el tiempo de desarrollo y mantenimiento de un sistema demasiado complejo, tanto en el consumo como sobre todo en la elaboración de los contenidos del medio digital.

Es por ello por lo que entre agosto de 2013 y mayo de 2014 se realizó una migración de la web a WordPress, se crearon nuevos contenidos, se apostó por nuevas temáticas como los cómics y los videojuegos y finalmente se aplicó *email marketing* y otros métodos de marketing digital. Todo este trabajo fue recopilado en el Trabajo Final de Máster dentro del marco del Máster Universitario de Gestión de la Información en la Universidad Politécnica de Valencia, con el nombre de *Aplicación de una estrategia de optimización de la gestión de un portal de contenidos*, leído el 30 de junio de 2014.

Entre los métodos llevados a cabo, uno de los más destacados fue la optimización de Facebook, Twitter y Google+, principales redes sociales mediante las cuales se difundían los contenidos del sitio web para lograr un mayor alcance de éstos.

En el caso de Twitter, red social en la que se ha centrado la atención en esta tesis, se seguía originalmente una estrategia basada en un elevadísimo número de tuits (publicaciones en Twitter) publicados cada unos pocos minutos, repitiendo los últimos artículos de manera automática.

Se decidió, en su lugar, por un ritmo de publicación mucho menor con una calidad mayor en cada tuit, incluyendo contenido multimedia para aumentar la atención de los usuarios y con el objetivo de fidelizar a la audiencia ofreciendo una mejor experiencia a la hora de poder mantenerse al día de las publicaciones del medio digital.

[40] https://cutephp.com/cutenews/upgrade.php
[41] https://www.drupal.org/

La publicación de los tuits, con una imagen incrustada y un mensaje personalizado, se programó entre las 15-18h y las 21-0h, considerados períodos del día con mayor posibilidad de tráfico según las estadísticas de Google Analytics de la web.

El resultado fue de un aumento de aproximadamente 100 seguidores al mes y la conclusión, a través de la comparativa entre la estrategia utilizada en Facebook y Twitter en este caso real, de que “Twitter permite mucho más la viralización de los contenidos que Facebook, y además obtiene una participación más entusiasta de los seguidores.”

A continuación, se comenzó a preparar la presente tesis bajo el marco del Doctorado de Industrias de la Comunicación y Culturales, en la cual se pretendía estudiar y optimizar la comunicación en las redes sociales de manera que se aumentará el alcance en todas estas a través de la mejora tanto de la comunicación como de la publicación de los contenidos.

Sin embargo, conforme avanzaba la investigación, se tomó la decisión de ceñirse a Twitter debido a la mayor cantidad de seguidores del medio digital, su carácter informativo y la capacidad de su API para cumplir con las expectativas de la metodología.

En 2016 se hicieron varias tentativas de cambio en la programación de la comunicación en Twitter, con el objetivo de que sirvieran como primeras puestas de contacto y búsqueda de problemas que pudieran dificultar la programación de la metodología final. Para ello, se puso en marcha un algoritmo que fue perfeccionado a lo largo de los meses de marzo y abril de 2016.

En ese momento, el medio contaba con más de 5.000 artículos de las siguientes tipologías: noticias, críticas, reportajes, entrevistas, artículos de estrenos y programas del podcast. Debido a que el 47,07% de las visitas procedían de España, el horario en el que se ejecutaba el algoritmo se adaptó al español, siendo entre las 8h y 1h del día siguiente.

La primera versión capturaba los últimos artículos publicados en cada tipología y los colocaba en una cola, publicándolos de manera aleatoria a un ritmo de 1 tuit por hora y eliminándolos de esta. Cuando se vaciaba la cola, se volvía a generar para continuar el mismo proceso. Se puso en marcha entre el 8 y el 30 de marzo de 2016, y se produjo un 15,21% más de visitas que en el periodo anterior.

La segunda versión fue similar a la anterior, pero aumentando el ritmo a dos tuits por hora. Se puso en marcha entre el 31 de marzo y el 18 de abril de 2016, y se produjo un 13,62% más de visitas que en el periodo anterior.

Sin embargo, en el caso de esta segunda versión, se observó un aumento de quejas por parte de algunos seguidores de que los tuits se repetían. El problema principal es que el ritmo de publicación no era el suficiente para que la renovación de cada ciclo del algoritmo permitiera una renovación asimismo de los contenidos que se compartían en Twitter. Además, si no se publicaban suficientes artículos de una misma tipología para satisfacer la demanda del algoritmo, se podrían compartir enlaces de artículos ya desfasados.

Por ello, la tercera y última versión incorporó un número menor de artículos, discriminando tipologías que no se publicaban continuamente (como las entrevistas) e incorporando filtros temporales para asegurar la compartición de noticias consideradas de actualidad (siete días anteriores). En el caso de las críticas, los reportajes y los programas del podcast, se intercalaban los últimos publicados con alguno al azar de todos los demás a modo de intentar aprovechar los ya publicados un tiempo atrás. A los programas de podcast más antiguos se les añadía una coletilla con la intención de añadir naturalidad a los tuits.

Además, se incorporó un filtro que no permitía la publicación de un tuit si este había sido publicado en las 6 horas anteriores, y los tuits se ordenaban por prioridad, número de veces que han sido compartidos, la fecha de publicación del último tuit y la fecha de publicación del artículo en sí.

Se puso en marcha del 19 de abril al 12 de julio, y se produjo un 10,76% más de visitas que en el periodo anterior.

Gracias a estos experimentos preliminares, se estudió en más profundidad las limitaciones de la API de Twitter, las consecuencias de versiones anteriores a la final y se determinó la necesidad de añadir más complejidad a dicha metodología para que tuviera en cuenta otros factores: el cálculo de la prioridad de los tuits en base al éxito de los hashtags, el número de lecturas del artículo, el éxito de los tuits anteriores, el análisis de los *influencers* y los competidores, la adición de otros contenidos de humor, tener en cuenta otros públicos como los lectores latinoamericanos, calcular el horario de mayor audiencia para publicar los artículos más exitosos en esa franja horaria...

Todo ello derivó en la necesidad de conocer qué artículos deberían tener una mayor prioridad tanto en la estrategia de contenidos como en la de comunicación en las redes sociales. Tras haber realizado un acercamiento a este último con los experimentos anteriormente mencionados, se decidió evaluar la problemática desde una perspectiva más amplia y centrarse en una metodología que ayudará a discernir la prioridad de cada artículo, para que luego sirviera de ayuda en las decisiones posteriores sobre dicho contenido.

El propio diseño de esta metodología daba cabida para una tesis. Una metodología que prometía ser más compleja conforme más se profundizaba en sus posibilidades y los resultados del experimento. Es por ello por lo que se decidió elegir el título de la tesis *Diseño de una metodología de cálculo del éxito para la optimización de contenidos web*, centrándose además en la tipología de las noticias por ser más susceptibles a las novedades, las tendencias y los intereses de los usuarios en el tiempo.

3.2. Selección de los datos de estudio

El análisis de resultados se ha realizado tomando como universo el conjunto de noticias publicadas en el medio digital. Dicha elección se ha justificado en el apartado 2.2.4.3, en el que se plantea el universo de las noticias de última hora como uno de los más interesantes a la hora de realizar el análisis de tendencias: el 85% de las tendencias en Twitter son de noticias de última hora o noticias de un mayor recorrido (Kwak, et al., 2010).

Los datos seleccionados en el estudio actual engloban el 100% de las noticias de última hora publicadas por el medio digital en el marco temporal elegido, dando lugar al análisis completo de la estrategia de contenidos en la tipología mencionada.

También se ha realizado el estudio para poblaciones consistentes en el 100% de las noticias de última hora publicadas de una temática concreta y con suficiente tamaño para permitir sacar conclusiones. En el caso de uso de esta tesis son: cine, series y videojuegos. Además, debido a que los resultados del análisis auguraban un mejor resultado si se producía una selección más restrictiva, se hace también el estudio del 100% de las noticias de última hora cuyo contenido principal es un tráiler.

Al haber elegido poblaciones del 100% de los elementos disponibles de una tipología concreta, se asegura la representatividad de la muestra, ya que engloba al 100% de la población. En el caso de reproducir este estudio con una muestra de menor porcentaje, se tendría que realizar teniendo en cuenta criterios probabilísticos mediante un muestreo aleatorio, ya sea simple o sistemático, o un muestreo de probabilidad, ya sea por conglomerados o estratificado (por secciones o temáticas principales, por ejemplo). La ventaja de realizar un muestreo aleatorio o de probabilidad es la de tener en cuenta la validez estadística de población, es decir, en qué medida se pueden extrapolar las conclusiones a toda la población. Si la población es tan grande que las características del estudio (marco temporal, cuotas de las API, etc.) no permiten una muestra al 100%, es importante asegurar la representatividad y la practicabilidad, teniendo especial cuidado con los riesgos de muestreo: la aceptación y el rechazo incorrectos. Incluso los métodos de muestreo anteriormente mencionados, que minimizan el error del proceso de muestreo, están sujetos al azar y pueden dar lugar a sujetos no representativos de las conclusiones generales por casualidad.

Es por ello por lo que, en ese caso, se debería calcular la desviación estándar muestral para expresar así la variabilidad de la población de la muestra y si esta es alta, aumentar el tamaño de la muestra para que el error del proceso de muestreo disminuya. En el caso

de uso de esta tesis no es necesario debido a que el único muestreo que se ha realizado consiste en el 100% de las noticias de última hora publicadas de temáticas concretas.

Los datos analíticos de las noticias de última hora publicadas en un medio digital sirven como base para el análisis del contenido según unos objetivos seleccionados de antemano. Esos objetivos delimitan el estudio según las variables que se consideren como éxito, ya que se tratará de predecirlas para ayudar en la toma de decisiones de la estrategia de contenidos.

Esta predicción se realizará mediante una fórmula que se extrae de una regresión lineal múltiple y/o de una regresión de Poisson o una regresión binomial negativa. Para poder realizar estas regresiones, los datos tienen que cumplir ciertas condiciones estadísticas. Durante dos meses se extraerán datos de entrenamiento, tras lo cual se obtendrán datos de test para validar las ecuaciones de predicción.

Por ello, la investigación se ha dividido en dos fases:

- Fase 1: Análisis de los datos de entrenamiento. Se ha realizado una medición según los indicadores propuestos en el apartado 3.4, utilizando como población el 100% de las noticias de última hora publicadas en el medio digital durante 2 meses. Esto permite ver una panorámica general del consumo de contenidos en el medio digital, tras lo cual se ha realizado un análisis de las variables de éxito para comprobar su correlación y si es necesario estudiar todas ellas. Tras ello, se ha procedido a comprobar las condiciones estadísticas de la regresión lineal múltiple. A continuación, se ha realizado un análisis de componentes principales con el objetivo de reducir la dimensionalidad para identificar y eliminar los indicadores irrelevantes, mejorar el rendimiento computacional y reducir la complejidad tanto de la medición como del análisis posterior. Gracias a todo ello, se han realizado los estudios de regresión lineal múltiple para cada una de las variables de éxito, extrayendo las ecuaciones de predicción. Por último, se ha realizado también el filtro de alta correlación de las variables de predicción y, con las resultantes, se ha realizado la regresión de Poisson o la regresión binomial negativa de las variables de éxito, extrayendo así las ecuaciones de predicción. Se ha repetido el proceso para el 100% de las noticias de última hora de cine, series, videojuegos y tráileres.
- Fase 2: Análisis de los datos de test. Se ha realizado una medición según los indicadores propuestos en el apartado 3.4, utilizando como población el 100% de las noticias de última hora publicadas en el medio digital durante un mes. En esta fase se ha aplicado las ecuaciones de predicción obtenidas en la Fase 1 y los datos

reales, realizando una comparación de ambos con la raíz del error medio cuadrático para obtener, por tanto, la validez de la predicción.

3.3. Selección de temáticas

En esta investigación se ha considerado una temática como una palabra clave o conjunto de palabras clave relacionadas temáticamente. Las temáticas principales son las secciones de contenido del medio en el que se va a llevar a cabo la metodología, de manera que enmarcan las temáticas más específicas de cada contenido en concreto. Por ejemplo, en el caso de uso en el que se ejecutan los experimentos de esta tesis, las temáticas principales serían el cine, las series de televisión, los videojuegos, la literatura y los cómics. Además, se han seleccionado otras temáticas principales por la tipología del contenido: noticias cuyo contenido principal es un tráiler y noticias que traten de la temática de los superhéroes, popular en el caso de uso. Estas temáticas principales sirven para segmentar los tipos de contenido, pero son demasiado genéricas para su inclusión en el paquete de palabras clave de manera directa.

Para obtener el paquete de palabras clave a analizar se ha tenido en cuenta una serie de conceptos clave relacionados con el contenido que luego podrían ser potenciales etiquetas o tags en la publicación del contenido en el medio. De esta manera se pueden obtener las subcategorías relacionadas con cada contenido con el objetivo de calcular tendencias y el éxito de dichas temáticas en el momento actual. Estos conceptos clave pueden ser de distintos tipos enmarcados en las temáticas principales del medio de comunicación online:

- Títulos de obras: los títulos de las obras relacionadas con el contenido, preferentemente en el idioma del medio, y en versión original si no se dispone de este. En el caso de uso de esta investigación podrían ser películas, sagas de películas, series de televisión, videojuegos, novelas, cómics...

- Nombres de entidades: los nombres de las entidades relacionadas con el contenido. En dicho caso de uso podría tratarse de directores, actores, guionistas, autores, dibujantes, productoras, distribuidoras, canales de televisión, plataformas en línea o servicio de VOD, premios, editoriales, videoconsolas...

- Nombres de eventos: los nombres de los eventos relacionados con el contenido, preferentemente en el idioma del medio, y en versión original si no se dispone de este. En un caso de uso de las temáticas anteriores podría darse ejemplos como festivales, mesas redondas, congresos, convenciones internacionales, ferias...

Es necesario tener en cuenta que estos conceptos clave deben tener una relación directa con el contenido y no circunstancial o de contexto. Por ejemplo, si se trata del nuevo tráiler de una película, un concepto clave sería el título de la película, los actores protagonistas e

incluso el director, pero no necesariamente todos los miembros del reparto y otras entidades relacionadas, ya que son datos de contexto que no aportan tanto valor, al no estar relacionados directamente con el contenido del artículo. Otro ejemplo sería el de una noticia de última hora en la que un actor se ha unido al reparto de una película. En este caso tanto el actor como la película serían conceptos clave relacionados con el artículo, pero no el resto del reparto, la productora o el guionista de la película. Esta selección de términos se ha realizado de manera manual por parte del redactor de cada artículo del medio digital.

3.4. Indicadores

Teniendo en cuenta todos los trabajos de investigación consultados en el apartado 2 del estado de la cuestión, a continuación, se exponen los parámetros e indicadores que participarán en el diseño de la metodología cibermétrica que da lugar a esta tesis.

Este modelo de diseño hará uso de un conjunto de indicadores estructurados en ocho apartados, relacionados con las tres dimensiones en las que se propone estudiar el éxito de un contenido digital y su repercusión en Twitter: la analítica web de los contenidos relacionados publicados por el medio, la analítica de los tuits de los contenidos relacionados publicados por el medio y el análisis de tendencias en Twitter de los contenidos relacionados.

La analítica del contenido en la web se divide en tres apartados:

- Adquisición: Indicadores que ayuden a cuantificar el acceso a la información.
- Comportamiento: Indicadores que aporten datos sobre la navegación.
- Resultados: Indicadores que evalúen la conversión de la información en forma de anuncios Adsense.

La analítica del contenido en la cuenta de Twitter se divide a su vez en dos apartados:

a) Amplificación: Indicadores centrados en el grado en que los usuarios se han hecho eco de los tuits relacionados con el contenido gracias a su compartición.

b) Elogio: Indicadores con información sobre el marcado como favoritos de los tuits relacionados con el contenido.

Por último, el análisis de tendencias de los hashtags y conceptos clave del contenido incluye los dos apartados de la analítica del contenido en Twitter, pero orientados a los tuits de una consulta sobre esos hashtags y conceptos clave. Además, también tiene en cuenta otros dos apartados:

- Autoridad: Indicadores relacionados con la influencia que pudiera tener el autor de cada tuit relacionado.
- Contenido: Indicadores relacionados con el contenido en sí de los tuits relacionados que pudieran afectar a la compartición de estos.

Cada indicador será analizado independientemente de los demás, para luego combinarse con el resto de las variables en un análisis factorial para llegar a un modelo que se propone

como metodología en la fase final de la investigación. De esta manera, se obtendrá una valoración a modo de predicción del posible éxito de un contenido que podrá utilizarse para comparaciones, priorizaciones, optimizaciones y otro tipo de decisiones de estrategia de contenidos y marketing digital.

Se trata, pues, de una aproximación cibermétrica a la predicción del éxito de contenidos online, por lo que es necesario subrayar también la importancia que tienen los factores relacionados con el medio en sí, que siempre pueden influir en las métricas globales y la forma en que se comunican los contenidos. Factores como la usabilidad y el buen funcionamiento del medio, la accesibilidad, el diseño web, las temáticas y otras limitaciones editoriales e incluso la línea de comunicación elegida por el medio como negocio.

Estos parámetros pueden servir para líneas de investigación futuras, ya que aportan información extra que enriquece el concepto de éxito que pueda tener el equipo editorial del medio. La metodología de esta tesis se puede adaptar por tanto a diferentes indicadores, ya que el objetivo es aportar una mayor inteligencia en la toma de decisiones editoriales, que sin duda podría servir de apoyo y empuje en el alcance y *engagement* total de sus contenidos.

3.4.1. Analítica del contenido en la web

Los indicadores relacionados con la analítica web de un contenido digital se han clasificado teniéndose en cuenta el enfoque de Kaushik (2011), dividiéndolas en categorías de Adquisición, Comportamiento y Resultados. Se han tenido en cuenta las métricas básicas propuestas por Alvarez Intriago et al. (2016) junto con algunas de las métricas disponibles de la vinculación de Google Analytics con Google Adsense (Google Developers, s.f.).

Para medir los anuncios de Google AdSense se aportan indicadores como el CTR y el eCPM, los más adecuados para el caso de uso de los propuestos por Gutiérrez Argüello (2013), basándose así en las métricas clave de la publicidad web desde sus inicios, que ya mencionaba Aggarwal et al. (1998), como son los clics (ratio de clic o CTR) y las impresiones.

Se almacenan los datos de todos los artículos relacionados con el que ha sido publicado, de los cuales se obtienen los totales y promedios, pero puesto que el objetivo es estudiar las características de tráfico y consumo de contenidos en la web, para los promedios solo se han tenido en cuenta los artículos que han tenido tráfico en el periodo analizado.

No se ha tenido en cuenta indicadores relacionados con el formato de los anuncios en sí, ya que el caso de uso a analizar utiliza la función de anuncios automáticos de Google AdSense. Esta función analiza el sitio web y coloca los anuncios de manera automática donde calcule una mayor probabilidad de que obtengan un buen rendimiento y generen un mayor número de ingresos. Para ello, comprenden la estructura de cada página web, detectan otros anuncios de Google en la página e insertan los anuncios nuevos automáticamente según características como el diseño, la cantidad de contenido y la presencia de otros anuncios de Google (Google AdSense, s.f.).

Se utilizarán los siguientes indicadores para analizar los contenidos relacionados con una temática concreta publicados por el medio en el que se está aplicando la metodología.

Adquisición	Páginas vistas únicas (total)
	Páginas vistas únicas (media)
	Entradas (media)
Comportamiento	Duración en la página (media)
	Porcentaje de salida (media)
	Páginas vistas por sesión (media)
Resultados	Impresiones de anuncios por sesión (total)
	Impresiones de anuncios por sesión (media)
	Porcentaje de impresiones visibles de anuncios (media)
	CTR de los anuncios (media)
	eCPM: Estimación de ingresos Adsense / 1000 páginas vistas (media)

3.4.2. Analítica del contenido en la cuenta de Twitter

Los indicadores relacionados con la analítica de un contenido digital en una cuenta de Twitter se han clasificado según el enfoque de Kaushik (2011) en Amplificación y Elogio. Así, se han tenido en cuenta dos de los mecanismos de interacción principales de Twitter: los retuits y los favoritos. Se utilizarán para analizar los contenidos relacionados con una temática concreta publicados en Twitter por el medio en el que se está aplicando la metodología.

Amplificación	Retuits (total)
	Retuits por tuit (media)
Elogio	Favoritos (total)
	Favoritos por tuit (media)

3.4.3. Análisis de tendencias

Los indicadores relacionados con el análisis de tendencias de los hashtags y conceptos principales relacionados con el contenido en Twitter se han elegido según las métricas del apartado 3.4.2, siguiendo la lógica descrita por Suh et al. (2010), ya que realizaron un estudio sobre los factores que potenciaban el uso del retuit y, por tanto, fomentaban la difusión de la información. En cuanto a la autoridad del usuario, se ha optado por las métricas que analizan el autor de cada tuit según Thelwall y Cugelman (2017), debido a su enfoque de vincular la autoridad a la difusión. Se utilizarán para medir en qué medida los hashtags y conceptos principales del contenido son tendencia en los últimos siete días (máximo de tiempo que permite la versión estándar de la API de Twitter (Twitter Developers, s.f.).

Amplificación	Tuits (total)
	Retuits (total)

	Retuits por tuit (media)
	Porcentaje de seguidores que han publicado tuits sobre la misma temática
Elogio	Favoritos (total)
	Favoritos por tuit (media)
Autoridad	Seguidores del autor del tuit (media)
	Tuits de la cuenta del autor del tuit (media)
	Edad de la cuenta del autor del tuit (media)
Contenido	Inclusión de una URL en el tuit como 0 / 1 (media)

Este análisis de tendencias se ha realizado en dos instantes diferentes: el mismo día de la publicación del artículo con las temáticas a estudiar, y 14 días después para evaluar la evolución de dicha tendencia en ese periodo de tiempo.

3.5. Instrumentos de recolección de datos

Con la creciente expansión de internet se vuelve cada vez más interesante el análisis de los contenidos digitales, aumentando así la atención que le prestan tanto en el ámbito privado como en el público, en las empresas y en las instituciones de investigación. Por ello, con los años han surgido multitud de herramientas y programas que analizan de manera automática los contenidos online, así como las cuentas de Twitter y sus publicaciones. Muchas de ellas cuentan con una versión gratuita que permite una serie de funcionalidades básicas con ciertos límites, permitiendo superar estos o acceder a otras funcionalidades más avanzadas previo pago de una cuenta premium.

En este caso, debido a la necesidad de integrar datos de dos ámbitos separados como son el de la publicación en la web y el análisis de sus comparticiones y las temáticas relacionadas en Twitter, se ha optado por una programación a medida que accede tanto a la API de informes de Google Analytics v4 como a Twitter Standard API, siempre atendiendo a los límites de estas. Todo ello permite extraer los datos y realizar los cálculos pertinentes con ellos para dar lugar a los experimentos necesarios para el diseño de la metodología.

3.5.1. API de informes de Google Analytics v4

Una de las herramientas más utilizadas a la hora de analizar la web es Google Analytics, aunque utilizar esta para *machine learning* es un método relativamente nuevo debido quizá a su enfoque de datos globales o no agregados (Cristian, et al., 2019). La plataforma consiste en cuatro componentes principales en los que se procesa la información gracias a la interfaz de usuario, las bibliotecas de cliente y las APIs: la recogida de datos sobre las interacciones de los usuarios, la configuración del procesamiento de esos datos, el procesamiento en sí según los datos de configuración y los informes con la posibilidad de ver los datos procesados. Todo este proceso se puede ver en el esquema de la Figura 106 (Google Developers, s.f.).

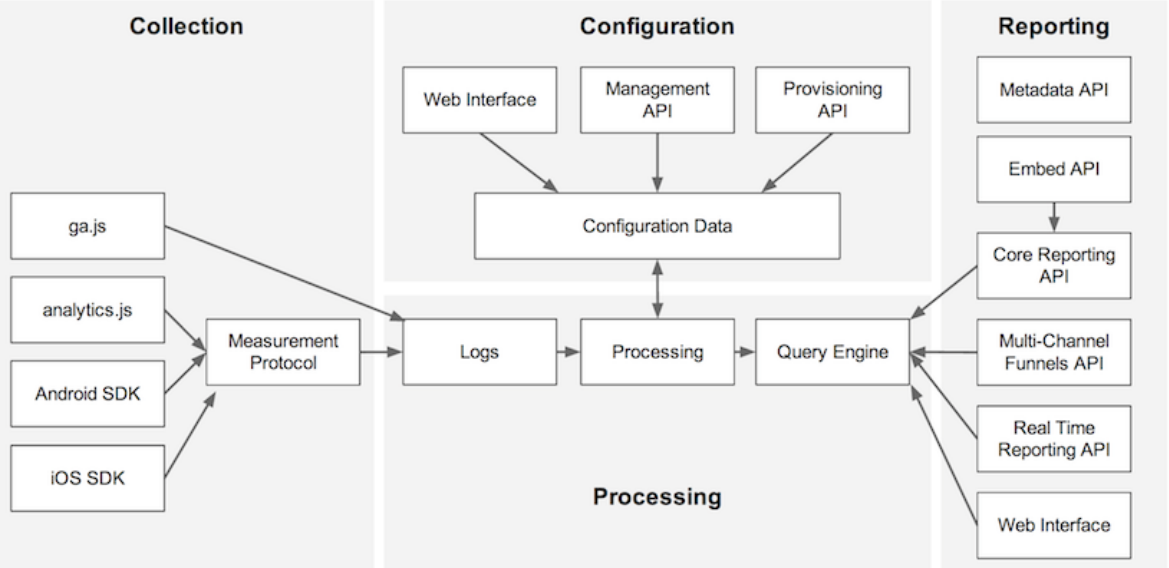


Figura 106. Esquema de los componentes de Google Analytics (Google Developers, s.f.)

En el momento en el que se elabora esta tesis, la versión 4 de la API de informes de Google Analytics es el método programático más avanzado para acceder a los datos de los informes de la herramienta y permite, además, crear paneles personalizados, automatizar tareas complejas de creación de informes e integrar los datos de Google Analytics con otras aplicaciones. Además, ofrece la posibilidad de solicitar no solo métricas integradas sino también combinaciones de métricas con operaciones matemáticas, obtener datos de dos periodos y de varios segmentos con la misma solicitud.

Para ello, la API ofrece una lista de dimensiones y métricas mediante las que se puede comprobar no solo cuáles hay disponibles sino también para identificar las combinaciones válidas entre estas. Se trata de una lista de 510 dimensiones y métricas divididas en 33 agrupaciones que se puede consultar en el anexo 0 (Google Developers, s.f.).

Google Analytics Reporting API v4 exporta los datos en formato JSON, el cual puede luego ser procesado para recoger y clasificar la información de sus componentes principales: datos de usuario, datos de sesión, datos de páginas vistas y datos de Adsense, entre otros. Los datos de usuario representan la información de cada usuario, con un Client ID único (referido al navegador en sí), un dispositivo, un navegador y un tipo de usuario (nuevo o recurrente). Los datos de sesión ofrecen información de cada sesión, como la duración, el número de búsquedas, el número de páginas vistas, etc. Los datos de páginas vistas son datos sobre la información de la navegación de cada sesión o como el tiempo transcurrido en cada página en la que se navega, cuándo se accedió a una página (Cristian, et al., 2019). Los datos de Adsense, por último, extraen información de la integración de Google

Analytics con Google Adsense para ofrecer datos sobre la visualización e interacción con los anuncios.

De esta herramienta se podrá extraer la información que necesitamos para los indicadores del apartado 3.4.1, es decir, la analítica referente al contenido en la web, utilizando como dimensión la URL de la página: ga:pagePath. A continuación, se puede observar el listado de indicadores y la variable de su dimensión o métrica correspondiente en la API de Google Analytics:

Adquisición	Páginas vistas únicas (total)	ga:uniquePageviews
	Páginas vistas únicas (media)	ga:uniquePageviews
	Entradas (media)	ga:entranceRate
Comportamiento	Duración en la página (media)	ga:avgTimeOnPage
	Porcentaje de salida (media)	ga:exitRate
	Páginas por sesión (media)	ga:pageviewsPerSession
Resultados	Impresiones de anuncios por sesión (total)	ga:adsenseAdUnitsViewed
	Impresiones de anuncios por sesión (media)	ga:adsenseAdUnitsViewed
	Porcentaje de impresiones visibles de anuncios (media)	ga:adsenseViewableImpressionPercent
	CTR de los anuncios (media)	ga:adsenseCTR
	eCPM: Estimación de ingresos Adsense / 1000 páginas vistas (media)	ga:adsenseECPM

Las cuotas aplicables a la versión 4 de la API de informes de Google Analytics son las siguientes: 50.000 solicitudes por vista o perfil al día y 10 solicitudes simultáneas por vista o perfil. También permite 2.000 solicitudes por cada 100 segundos y 100 solicitudes por

cada 1000 segundos por usuario. Además, en caso de que una solicitud falle y reciba como respuesta un error 500 o 503, Google Analytics permite 10 solicitudes fallidas por proyecto y perfil a la hora y 50 al día. En el caso de superar esa cuota, la aplicación manda un código de error 403 o 429 y un mensaje en el que informa de que la cuota ha sido superada. Se puede consultar los límites y cuotas en las solicitudes a la API en el anexo 6.2.50 (Google Developers, s.f.).

Para que el proceso no se detuviera frente a un error de este tipo, se ha optado por implementar un proceso de retardo exponencial, mediante el cual se vuelve a intentar la solicitud de manera periódica durante un periodo de tiempo que cada vez es mayor. Es una estrategia de gestión de errores estándar para las aplicaciones de red, muy recomendada e incluso considerada “obligatoria” por la API de Google Analytics para reducir el número de solicitudes necesarias para una respuesta correcta. El proceso de retardo exponencial consiste en implementar un retardo exponencial simple, mediante el cual el proceso se retrasa 1 más un número aleatorio de segundos al encontrarse con una respuesta de error, lo reintenta, si vuelve a recibir una respuesta de error espera dos más un número aleatorio de segundos, lo reintenta, si vuelve a recibir una respuesta de error espera otros cuatro más un número aleatorio de segundos y así sucesivamente, aumentando esta cantidad de segundos de manera exponencial hasta un máximo de cinco intentos (Google Developers, s.f.).

3.5.2. Twitter Standard API

Para extraer la información relacionada con la analítica tanto del contenido en la cuenta de Twitter del medio como el análisis de tendencias en Twitter se utiliza la Twitter Standard API. Se trata de la versión estándar y gratuita de la API oficial de Twitter con posibilidad de hacer consultas básicas de los últimos 7 días, filtrar, obtener una muestra, publicar, mandar mensajes directos, acceder a la información pública de cuentas de usuario, crear y administrar listas, obtener tendencias de una localización específica, etc (Twitter Developers, s.f.). Se puede consultar toda la información de esta API en el anexo 6.4.

Para ello, se necesita una autenticación de tipo OAuth 1.0a, lo cual permite actuar y realizar peticiones API en lugar de una cuenta de Twitter específica. Se pueden encontrar más detalles al respecto en el anexo 6.4.1 (Twitter Developers, s.f.). Con la excepción de la Streaming API, el resto de las funcionalidades de la API de Twitter funcionan según los principios de diseño de REST (Representational State Transfer), devolviendo las respuestas en formato JSON a través de HTTP. Los límites de paginación son de 3200

mensajes en la cronología de usuario, mientras que en otro tipo de cronologías el máximo es de 800 (Twitter Developers, s.f.).

Las peticiones tienen un número máximo en ventanas de 15 minutos, por lo que Twitter recomienda la utilización de caché, la priorización de usuarios activos y la filtración o menor frecuencia de consultas de búsqueda sin resultados (Twitter Developers, s.f.). La información sobre los límites y cuotas en las solicitudes a la API se pueden encontrar en el anexo 6.4.10.

A la hora de analizar los tuits publicados por la cuenta del medio, se ha hecho uso de la función “statuses/user_timeline” que permite hasta 900 peticiones cada 15 minutos como usuario, hasta 1500 cada 15 minutos como aplicación y hasta 100.000 peticiones diarias en general. Devuelve una colección de hasta los 3200 tuits más recientes de un usuario en concreto, incluyendo los retuits nativos de otros usuarios (Twitter Developers, s.f.). Se puede encontrar la información completa sobre la función en el anexo 6.4.6.

A continuación, se puede ver la correspondencia de los indicadores del apartado 3.4.2 con los campos de cada tuit en la respuesta JSON de esta función, según el objeto Tweet especificado en el anexo 6.4.3 (Twitter Developers, s.f.):

Amplificación	Retuits (total)	Tweet / retweet_count
	Retuits por tuit (media)	Tweet / retweet_count
Elogio	Favoritos (total)	Tweet / favorite_count
	Favoritos por tuit (media)	Tweet / favorite_count

En cuanto a la API de Streaming, es decir, la referente al seguimiento de los tuits en tiempo real, consta de un acceso que permite hasta 400 palabras clave y hasta 5000 usuarios a los que se sigue. Tiene la peculiaridad de que los campos de seguimiento de palabras clave, seguimiento de usuarios y localización son utilizados con el operador OR. Esto significa que en el ejemplo de la consulta “track=foo&follow=1234” se devolvería un conjunto de tuits con la palabra clave “foo” o creados por el usuario 1234. Los parámetros (Twitter Developers, s.f.) así como los operadores (Twitter Developers, s.f.) disponibles

para esta función se pueden encontrar en la web oficial, pero para la metodología que se ha utilizado, se ha preferido optar por la API de búsqueda estándar.

La API de búsqueda estándar permite consultas simples de los índices de los más recientes o populares tuits, limitados a los últimos siete días, y se puede encontrar su documentación en el anexo 6.4.7. Twitter Developers advierte de que esta API está enfocada a la relevancia y no a la completitud, puesto que esta última solo se puede obtener con los planes Premium o Enterprise (Twitter Developers, s.f.). Esta funcionalidad permite 180 peticiones en ventanas de 15 minutos si se hace desde el punto de vista de un usuario y 450 peticiones en ventanas de 15 minutos en el caso de una aplicación aprobada por la plataforma. La lista completa de parámetros se puede encontrar en el anexo 6.4.7.3 (Twitter Developers, s.f.), así como los códigos de respuesta posibles de todas las API en el anexo 6.4.9 (Twitter Developers, s.f.).

También se ha utilizado la función “followers/ids”, que devuelve el conjunto de seguidores de un usuario paginado en grupos de 5000 y con un límite de 15 peticiones cada 15 minutos (Twitter Developers, s.f.). Se puede leer la documentación sobre ella en el anexo 6.4.8.

En cuanto a los indicadores del apartado 3.4.3, relacionados con el análisis de tendencias, se puede ver su correspondencia con los diferentes objetos de Tweet (anexo 6.4.3) (Twitter Developers, s.f.) y Usuario (anexo 6.4.4) (Twitter Developers, s.f.) en la siguiente lista:

Amplificación	Tuits (total)	Tweet
	Retuits (total)	Tweet / retweet_count
	Retuits por tuit (media)	Tweet / retweet_count
	Porcentaje de seguidores que han publicado tuits sobre la misma temática	(Seguidores en la búsqueda / Seguidores totales) * 100 Seguidores en la búsqueda: Usuario / id_str que también está en Seguidores totales Seguidores totales: Función followers/ids / user_id
Elogio	Favoritos (total)	Tweet / favorite_count
	Favoritos por tuit (media)	Tweet / favorite_count

Autoridad	Seguidores del autor del tuit (media)	Usuario / followers_count
	Tuits de la cuenta del autor del tuit (media)	Usuario / statuses_count
	Edad de la cuenta del autor del tuit (media)	Usuario / created_at
Contenido	Inclusión de una URL en el tuit (0 / 1)	Tweet / entities { urls }

3.6. Técnicas de procesamiento y análisis de datos

La recolección de los datos se ha hecho, como ya se ha indicado en el subapartado anterior, mediante las API oficiales de Google Analytics y Twitter.

En el caso de Google Analytics, Google Cloud ofrece una serie de bibliotecas cliente para poder llamar a sus API. Estas bibliotecas ofrecen la oportunidad a los desarrolladores de trabajar con ellas de forma óptima sin tener que escribir su propio código estándar. El listado de estas bibliotecas es el siguiente: Go, Java, Node.js, Python, Ruby, PHP, C# y C++ (Google Cloud, 2021). En este estudio se ha utilizado la librería de cliente para PHP oficial de Google APIs, con el requerimiento de utilizar PHP 5.4.0 o mayor (Google APIs, s.f.).

En cuanto a Twitter Analytics, el equipo de Twitter mantiene tres librerías oficiales: JavaScript / Node.js, Python y Ruby. Otras librerías mantenidas por la comunidad son: .NET (ASP, C#, VB), C++, Clojure, ColdFusion, Dart, Elixir, Erlang, Go, Haskell, Java, JavaScript / Node.js, Julia, Kotlin, Lua, Objective-C, Perl, PHP, PowerShell, Python, R, Ruby, Rust, Scala, Shell script (Bash), Swift y TypeScript (Twitter Developers, s.f.). En esta tesis se ha escogido la librería PHP más popular para conectarse a la Twitter OAuth REST API, creada por Abraham Williams y con el requerimiento de PHP 7.2.0 o mayor y dependencias de las librerías OpenSSL y cURL (Williams, s.f.).

Los resultados, obtenidos de las respuestas de las peticiones a las API en formato json, son procesados en PHP mediante funciones programadas a medida que sirven para realizar cálculos e iteraciones con los diferentes indicadores para tratar de obtener una estimación óptima de qué temática tiene más posibilidades de éxito en el instante de su ejecución. Posteriormente, todos los datos se han almacenado en tablas de MySQL, siguiendo la estructura descrita en el apartado.

Se ha llevado a cabo una metodología estadística basada en aprendizaje automático (*machine learning*), dividiéndose en dos fases: una fase de entrenamiento de datos para obtener los cálculos propuestos de predicción, y una fase de test de datos para comprobar la validez del modelo que se ha generado. El conjunto de datos se suele repartir en un 70% de datos de entrenamiento y un 30% de datos de test (Recuero de los Santos, 2020).

Por tanto, se ha realizado una primera fase de recogida de datos de entrenamiento mediante la cual se ha captado la información perteneciente a las noticias de última hora publicadas en un intervalo de dos meses.

Tras realizar un análisis de los indicadores, su variabilidad y correlación, se ha propuesto una serie de ecuaciones que pretenden predecir las variables de éxito que se hayan elegido de entre las del conjunto inicial. El análisis consiste en la selección y ponderación de los indicadores mediante regresiones múltiples, siempre que el modelo haya sido significativo, es decir, que las variables de predicción con las que se cuenta sean capaces de predecir la variable de éxito seleccionada.

Tras obtener las ecuaciones de las variables de éxito, se han validado en un intervalo de tiempo de experimentación de un mes. De esta manera se valida la metodología que se propone en esta tesis comparando las predicciones con los valores finalmente obtenidos en este mes.

Esta comparación se realiza según López (2017) mediante el cálculo del coeficiente de determinación (R cuadrado), que es la proporción de la varianza total de la variable explicada por las regresiones, de las cuales se obtienen las ecuaciones de predicción propuestas en la primera fase.

3.6.1. Tablas de la base de datos

Se ha realizado una obtención de datos de cada noticia de última hora publicada en la web según los indicadores descritos en el apartado 3.4. Todos los datos relacionados se almacenan en una base de datos en MySQL, dividiéndose en las siguientes tablas:

- **tesis_followers:** lista de User ID de los seguidores de la cuenta del medio.
- **tesis_hometimeline:** datos de los tuits publicados por la cuenta del medio que comparten noticias de última hora de la web.
    - status_id: ID del tuit
    - created_at: fecha de publicación
    - text: contenido del tuit
    - path: URL extraída tras procesar la URL acortada en text
    - post_shared: ID del artículo en WordPress que se está compartiendo
    - retweent_count: número de retuits
    - favorite_count: número de favoritos

- **tesis_hometimeline_other:** datos de tuits publicados por la cuenta del medio que no comparten noticias de última hora de la web. Otras tipologías, comparticiones automáticas de Facebook, tuits personalizados sin enlace a un artículo, etc. Con los mismos campos que tesis_hometimeline.
- **tesis_posts:** datos de los artículos publicados por la web y procesados para algún análisis.
    - stats_id: ID del análisis
    - post_id: ID del artículo en WordPress
    - post_date: fecha de publicación del artículo en WordPress
    - post_title: título del artículo
    - path: URL del artículo en la web del medio
    - tags: ID de etiquetas o tags de WordPress relacionadas con el artículo
    - uniquepageviews: páginas vistas únicas
    - entrancerate: ratio de entradas
    - avgtimeonpage: duración promedio de la visita
    - exitrate: ratio de salidas
    - pageviewspersession: páginas vistas por sesión
    - adsense_adunitsviewed: número de anuncios vistos por los usuarios
    - adsense_viewableimpressionpercent: ratio de visualizaciones de los anuncios
    - adsense_ctr: ratio de clics de los anuncios
    - adsense_ecpm: estimación de ingresos de los anuncios por cada 1000 páginas vistas
- **tesis_stats:** datos de un análisis en concreto, realizado a cada noticia de última hora publicada. Los campos con valores estadísticos se pueden calcular de los datos de las otras tablas, pero se guardan los cálculos totales y promedios para un procesamiento posterior más rápido y sencillo.

- id: ID del análisis
- phase: fase de la tesis en la que se ha realizado análisis (ahora mismo todos son 1)
- time: “0” si es en el momento de la publicación, “1” si es 14 días después
- start_date: fecha y hora en la que se realiza la medición el día de la publicación
- end_date: fecha y hora en la que se realiza la medición 14 días después
- main_post_id: ID del artículo publicado a analizar
- main_post_theme: Sección principal del artículo publicado a analizar
- superheroes_theme: “1” si trata sobre superhéroes, “0” si no
- trailer_theme: “1” si tiene tráiler, “0” si no
- name: campo vacío, posibilidad de añadirle un nombre personalizado manualmente
- notes: campo vacío, posibilidad de añadirle notas personalizadas manualmente, como si se ha quitado algún tag manualmente por ser considerado demasiado genérico, pese a que el redactor lo pusiera
- num_articles: número de artículos analizados
- num_articles_with_traffic: número de artículos analizados con tráfico (los que se tendrán en cuenta para el análisis de tráfico)
- num_articles_with_tw_data: número de artículos con datos de cuando fueron compartidos en la cuenta de Twitter del medio
- num_terms: número de términos analizados
- uniquepageviews_total: total de páginas vistas
- uniquepageviews_mean: promedio de páginas vistas
- entrancerate_mean: promedio de la ratio de entradas
- avgtimeonpage_mean: promedio de duración de las visitas

- exitrate_mean: promedio de la ratio de salidas
- pageviewspersession_mean: promedio de páginas vistas por sesión
- adsense_adunitsviewed_total: total de anuncios visualizados
- adsense_adunitsviewed_mean: promedio de anuncios visualizados
- adsense_viewableimpressionpercent_mean: promedio de la ratio de visualizaciones de los anuncios
- adsense_ctr_mean: promedio de la ratio de clics de los anuncios
- adsense_ecpm_mean: estimación de ingresos de los anuncios por cada 1000 páginas vistas
- retweet_count_total: total de retuits
- retweet_count_mean: promedio de retuits
- favorite_count_total: total de favoritos
- favorite_count_mean: promedio de favoritos
- terms_ini_num_tweets: total de tuits sobre los términos en el día de la publicación
- terms_ini_retweet_count_total: total de retuits sobre los términos en el día de la publicación
- terms_ini_retweet_count_mean: promedio de retuits sobre los términos en el día de la publicación
- terms_ini_favorite_count_total: total de favoritos sobre los términos en el día de la publicación
- terms_ini_favorite_count_mean: promedio de favoritos sobre los términos en el día de la publicación
- terms_ini_followers_talking_rate: ratio de seguidores de la cuenta de Twitter del medio que han publicado recientemente un tuit hablando sobre los términos en el día de la publicación

- terms_ini_user_num_followers_mean: promedio de seguidores de usuarios que han hablado de los términos en el día de la publicación
- terms_ini_user_num_tweets_mean: promedio de tuits publicados por los usuarios que han hablado de los términos en el día de la publicación
- terms_ini_user_age_mean: edad promedio en días de los usuarios que han hablado de los términos en el día de la publicación
- terms_ini_url_inclusion_rate: ratio de inclusión de URL de los tuits que hablan sobre los términos en el día de la publicación
- terms_end_num_tweets: total de tuits sobre los términos 14 días después de la publicación
- terms_ini_retweet_count_total: total de retuits sobre los términos 14 días después de la publicación
- terms_ini_retweet_count_mean: promedio de retuits sobre los términos 14 días después de la publicación
- terms_ini_favorite_count_total: total de favoritos sobre los términos 14 días después de la publicación
- terms_ini_favorite_count_mean: promedio de favoritos sobre los términos 14 días después de la publicación
- terms_ini_followers_talking_rate: ratio de seguidores de la cuenta de Twitter del medio que han publicado recientemente un tuit hablando sobre los términos 14 días después de la publicación
- terms_ini_user_num_followers_mean: promedio de seguidores de usuarios que han hablado de los términos 14 días después de la publicación
- terms_ini_user_num_tweets_mean: promedio de tuits publicados por los usuarios que han hablado de los términos 14 días después de la publicación
- terms_ini_user_age_mean: edad promedio en días de los usuarios que han hablado de los términos 14 días después de la publicación
- terms_ini_url_inclusion_rate: ratio de inclusión de URL de los tuits que hablan sobre los términos 14 días después de la publicación.

- **tesis_terms:** datos de los términos (tags) relacionados con los artículos procesados.
    - stats_id: ID del análisis
    - time: “0” si es en el momento de la publicación, “1” si es 14 días después
    - term_id: ID del término (etiqueta) en WordPress
    - name: Nombre del término
    - slug: URL del término
    - num_tweets: número de tuits
    - retweet_count_total: total de retuits
    - retweet_count_mean: promedio de retuits
    - favorite_count_total: total de favoritos
    - favorite_count_mean: promedio de favoritos
    - followers_talking_rate: ratio de seguidores de la cuenta de Twitter del medio que han publicado recientemente un tuit hablando sobre el término
    - user_num_followers_mean: promedio de seguidores de los usuarios que estaban hablando del término
    - user_num_tweets_mean: promedio de tuits publicados por los usuarios que estaban hablando del término
    - user_age_mean: edad promedio en días de los usuarios que estaban hablando del término
    - url_inclusion_rate: ratio de inclusión de URL

3.6.2. Procesos de recogida de datos

El procesamiento y recogida de los datos pertenecientes a los artículos publicados en el medio digital se ha efectuado siguiendo la siguiente estructura:

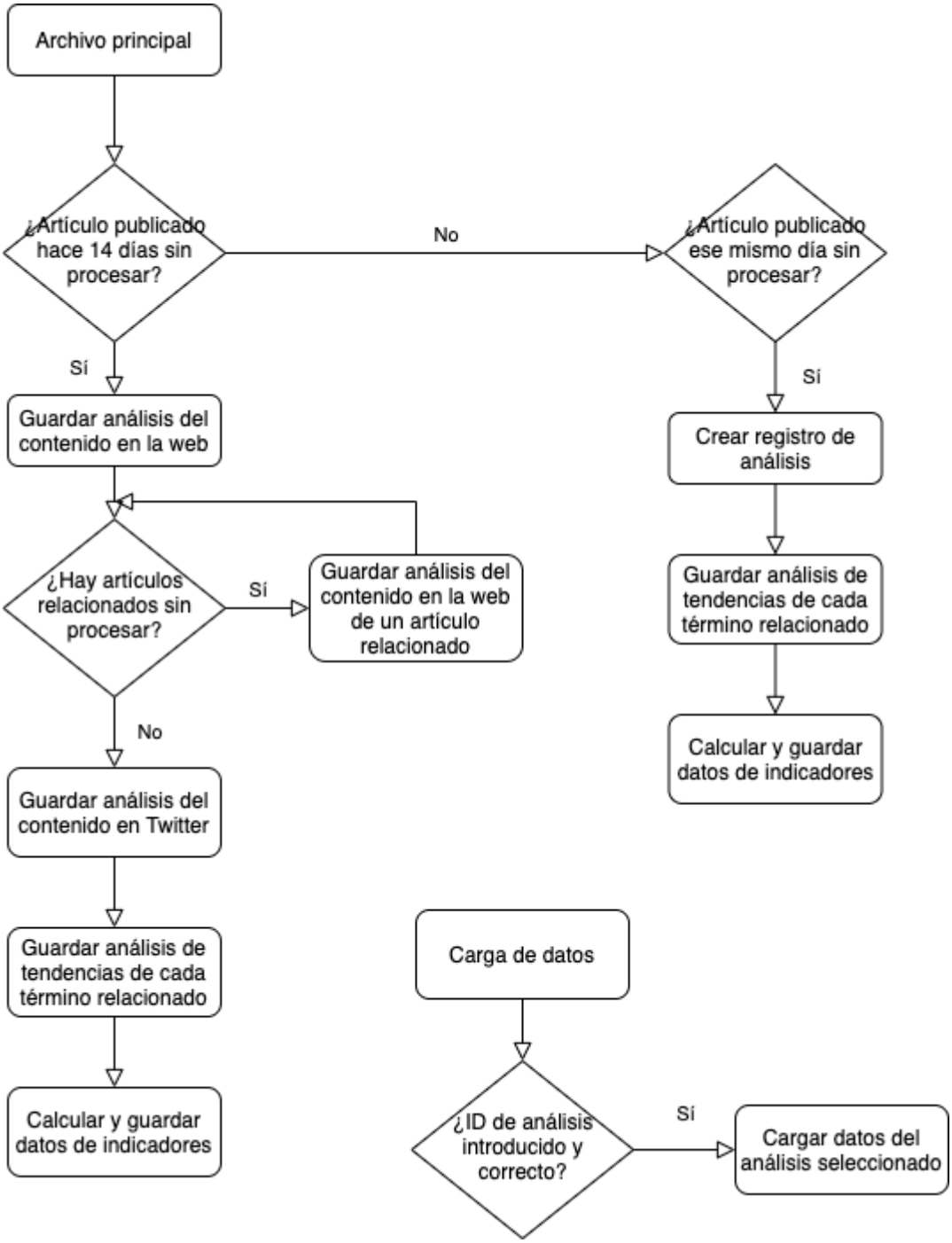


Figura 107. Estructura de procesos de la recogida de datos en la fase 1. Fuente: Elaboración propia

De esta manera, se puede observar los siguientes procesos:

- Ejecución cada 20 minutos desde las 10 de la mañana hasta las 12 de la noche del archivo PHP principal, que evalúa si hay un artículo pendiente de procesamiento. Hay dos tipos de procesamientos: un artículo publicado en ese mismo día, para el cual se deberá obtener los datos de análisis de tendencia iniciales, y un artículo publicado 14 días antes, para el cual se deberá obtener los datos de análisis del

contenido y análisis de tendencia 14 días después. La elección del intervalo de 20 minutos es debido a la cuota de peticiones de búsqueda en Twitter cada 15 minutos, dejando un margen de 5 para el propio procesamiento del archivo, que siempre es menor a este tiempo. La elección de las 10 de la mañana es debido a que prácticamente nunca se publican nuevas noticias antes de esa hora en el medio analizado.

- Si hay un artículo publicado hace 14 días:
    - Se guarda el registro del análisis del contenido en la web de ese artículo en los últimos 14 días, así como todos los artículos que tengan alguna de los términos relacionados con este.
    - Se guardan los datos relacionados con los indicadores del análisis en la cuenta de Twitter del medio.
    - Se guardan los registros del análisis de tendencias de cada término relacionado con dicho artículo, obteniéndose estos de las etiquetas dadas de alta en WordPress, el sistema de contenidos donde el artículo ha sido publicado.
    - Se calculan los datos de los indicadores relacionados con el análisis del contenido en la web, el análisis de contenido en Twitter y el análisis de tendencias 14 días después de la publicación del artículo, guardándose todo ello en el registro del análisis en la base de datos.
- Si hay un artículo nuevo publicado ese mismo día:
    - Se crea el registro del análisis de dicho artículo en la base de datos.
    - Se guardan los registros del análisis de tendencias de cada término relacionado con dicho artículo, obteniéndose estos de las etiquetas dadas de alta en WordPress, el sistema de contenidos donde el artículo ha sido publicado.
    - Se calculan los datos de los indicadores relacionados con el análisis de tendencias en el día de la publicación del artículo y se guardan en el registro del análisis en la base de datos.
- Transversalmente, se pueden cargar los datos de los indicadores del registro de un análisis en concreto y mostrarlos en la pantalla, seleccionado el análisis a mostrar

mediante el campo de un formulario, en el cual se pide al usuario el identificador de este.

Este proceso de recogida de datos ha exigido la programación de 15 archivos, sumando un total de 3.617 líneas de programación en PHP y JSON.

3.7. Acotación y selección de las unidades de tiempo

El proceso de recogida de datos de cada artículo se produce en dos instantes de tiempo: el día de la publicación y catorce días después de la publicación. Tal y como se ha descrito en el apartado 3.6.2, se realizan los siguientes procesos que captan datos de los periodos que se describen a continuación:

- El día de la publicación del artículo se recogen los datos del análisis de tendencia de ese día, perteneciendo a los siete días anteriores.
- Catorce días después, se recogen los datos de la analítica del contenido en la web, en la cuenta de Twitter del medio y los datos del análisis de tendencias, siendo éstos últimos los pertenecientes a los siete días anteriores a este momento.

Por ejemplo, en un artículo publicado el día 23 de septiembre de 2020 se realizarían las siguientes acotaciones de datos:

- Análisis inicial de tendencias del 16 de septiembre de 2020 al 23 de septiembre de 2020.
- Análisis del contenido en la web del 23 de septiembre de 2020 al 6 de octubre de 2020.
- Análisis del contenido en la cuenta de Twitter del medio del 23 de septiembre de 2020 al 6 de octubre de 2020.
- Análisis final de tendencias del 30 de septiembre de 2020 al 6 de octubre de 2020.

La línea de tiempo se puede observa en la Figura 108:

Figura 108. Línea de tiempo de la toma de datos de un artículo publicado el 23 de septiembre de 2020

3.8. Procedimiento para la representación e interpretación de los datos

En el capítulo siguiente se va a realizar el estudio de los datos de entrenamiento y de test del experimento de esta tesis, cuya representación se realizará mediante tablas y gráficos. Se resumirán así los resultados de los análisis realizados en el programa STATGRAPHICS Centurion.

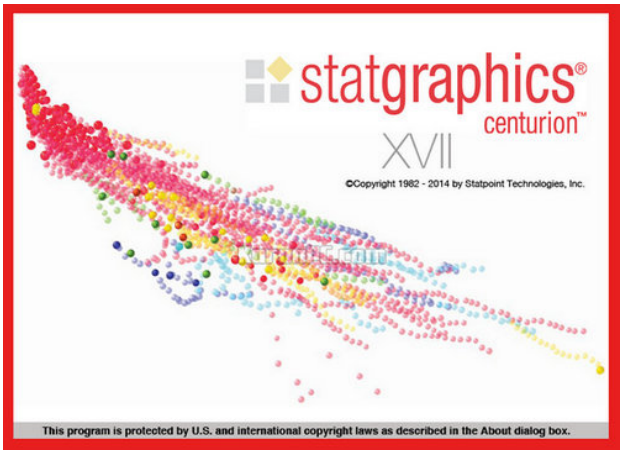


Figura 109. Logo de STATGRAPHICS Centurion XVII

STATGRAPHICS Centurion es una herramienta de análisis de datos con muchos procedimientos analíticos y funciones estadísticas avanzadas, como análisis de la varianza, regresión, control estadístico de procesos, diseño de experimentos, métodos multivariantes y análisis de series temporales y predicción (Anon., 2011). Gracias a este software, se representará la información cuantitativamente con dos objetivos: la representación gráfica de una gran cantidad de información que facilite su comprensión y la descripción y análisis de estos datos.

Para ello, los resultados de la investigación se han mostrado en los siguientes gráficos producidos por STATGRAPHICS Centurion:

- Gráfico de Caja y Bigotes. Ilustra propiedades importantes de una columna de datos numérica, resumiendo una muestra de datos a través de cinco estadísticas: el mínimo, el cuartil inferior, la mediana, el cuartil superior y el máximo. También facilita la detección de la presencia de datos atípicos y extremos. Los valores se ordenan de menor a mayor y se dibuja una caja desde el cuartil inferior al cuartil superior, mostrando un intervalo que incluye el 50% de los datos. Se dibuja una linea vertical que señala la mediana (la mitad de los datos). El símbolo “+” en rojo señala la media muestral. Los bigotes se dibujan desde los extremos de la caja

hasta los valores más grandes y pequeños de los datos, hasta alcanzar 1,5 veces el rango intercuartil (ancho de la caja). Cualquier valor más allá de estos bigotes se señala con un cuadrado, y si está situado a más de 3 veces el rango intercuartil de la caja son llamados valores extremos y se señalan con un "+" en rojo dentro del cuadrado. Todo ello se puede visualizar en el ejemplo siguiente, extraído desde la página oficial de Statgraphics (2005):

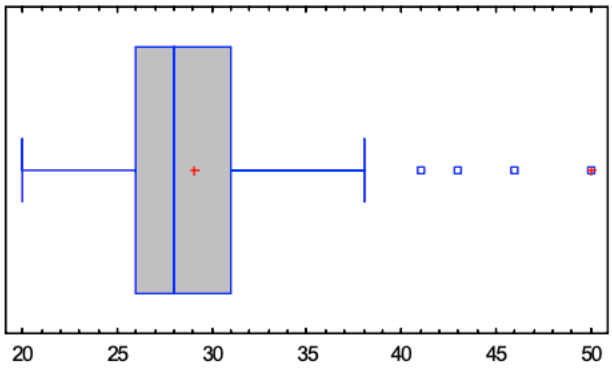


Figura 110. Ejemplo de gráfico de Caja y Bigotes (Anon., 2005)

- Gráfico de Probabilidad Normal. Sirve para juzgar si una muestra de datos proviene o no de una distribución normal, de manera que se pueda comprobar el tipo de alejamiento de la normalidad que tiene la muestra según cómo sus datos se desvían de la línea de referencia normal. Los datos se ordenan de menor a mayor, determinando las estadísticas de orden siendo la j-ésima estadística de orden la j-ésima observación más pequeña de la muestra. A continuación, se puede ver un ejemplo de este tipo de gráfico, extraído desde la página oficial de Statgraphics (2007):

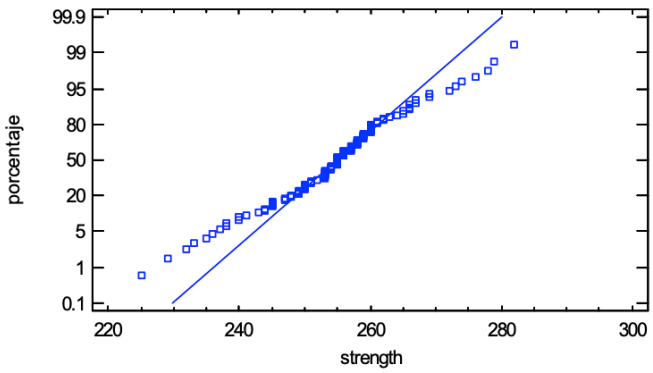


Figura 111. Ejemplo de gráfico de Probabilidad Normal (Anon., 2007)

- Matriz de correlaciones Pearson. El coeficiente de correlación de Pearson es una prueba de medición de la relación estadística entre dos variables continuas. Puede

tomar valores en un rango de +1 a -1, siendo 0 la asociación nula entre las dos variables (Ortega, 2019). A continuación, se puede ver un ejemplo de la matriz de estas correlaciones Pearson entre cada par de variables, siendo por tanto los mismos números a uno y otro lado de la diagonal. El color va desde el azul de -1 al rojo de +1, siendo verdes los valores cercanos al cero:

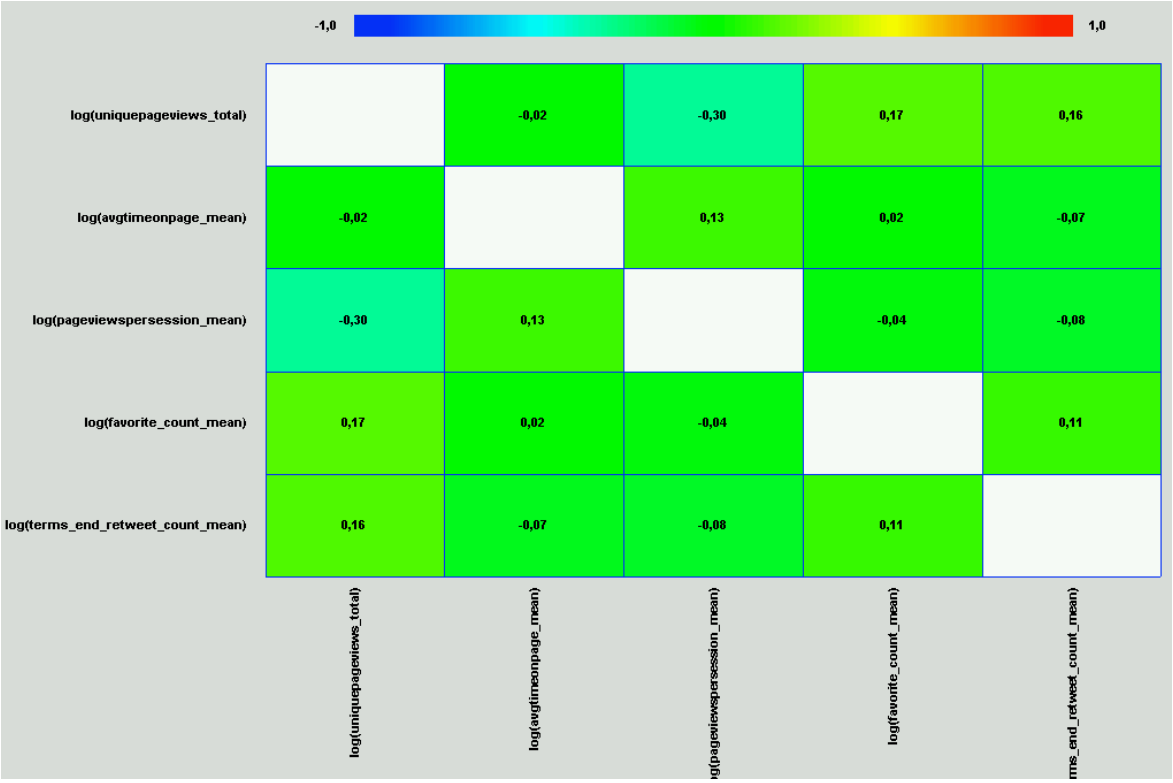


Figura 112. Ejemplo de Matriz de correlaciones Pearson

- Gráfico de Componentes Principales 2D y 3D. Las componentes principales son un conjunto de combinaciones lineales ortogonales que se extrae de las variables cuantitativas, de manera que tengan la máxima varianza. Se usa frecuentemente para reducir la dimensionalidad de un conjunto de variables predictoras, ya que se trata de representar la información con un número menor de variables o componentes principales. Este gráfico se utiliza para visualizar la localización de cada variable en el espacio de 2 o 3 componentes elegidas, de manera que aquellas que se alejen más de las líneas de referencia en 0 hacen la mayor contribución a dichas componentes, tal y como se puede ver en el ejemplo siguiente en 2D, extraído desde la página oficial de Statgraphics (2007):

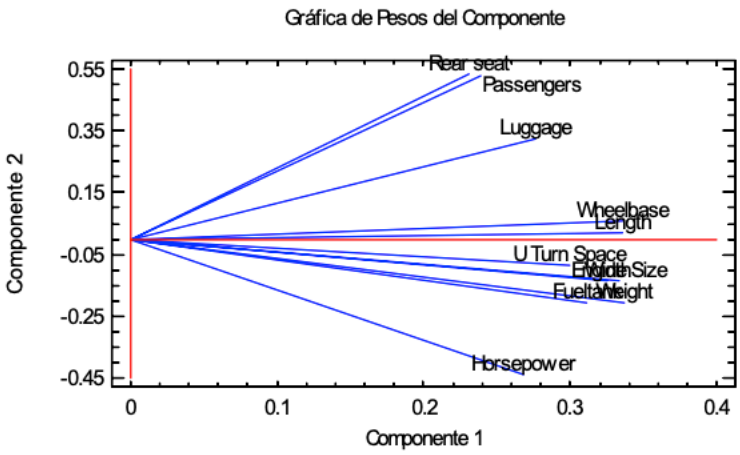


Figura 113. Ejemplo de gráfico de Componentes Principales 2D (Anon., 2007)

- Gráfico Observado contra Predicho. Muestra los valores observados en el eje vertical y los valores predichos por la regresión en el eje horizontal, de manera que, si el modelo se ajusta bien, los puntos se dispersarán aleatoriamente alrededor de la línea diagonal. A continuación, se puede comprobar en un ejemplo que la variabilidad aumenta conforme los valores predichos se hacen mayores, extraído desde la página oficial de Statgraphics (2006):

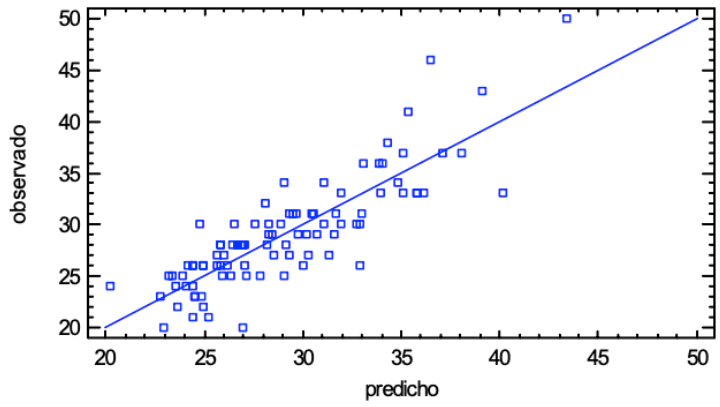


Figura 114. Ejemplo de gráfico Observado contra Predicho (Anon., 2006)

Además, para analizar valores anómalos de páginas vistas, se ha hecho uso de gráficos pertenecientes a la herramienta Google Trends, un servicio gratuito de Google que muestra de esta manera la frecuencia de búsqueda de palabras, frases y temas en un tiempo y una localización determinados. Permite la comparación de hasta cinco términos a la vez, ver la frecuencia de mención en noticias y las regiones geográficas en las que se han buscado más. Estos datos se extraen de una muestra de las búsquedas de Google para calcular el número de búsquedas total diariamente. A continuación, se puede ver un ejemplo de

búsqueda de los términos “google trends”, “google ads” y “adwords” para España en los últimos 12 meses, extraído de la web Arimetrics (2020):

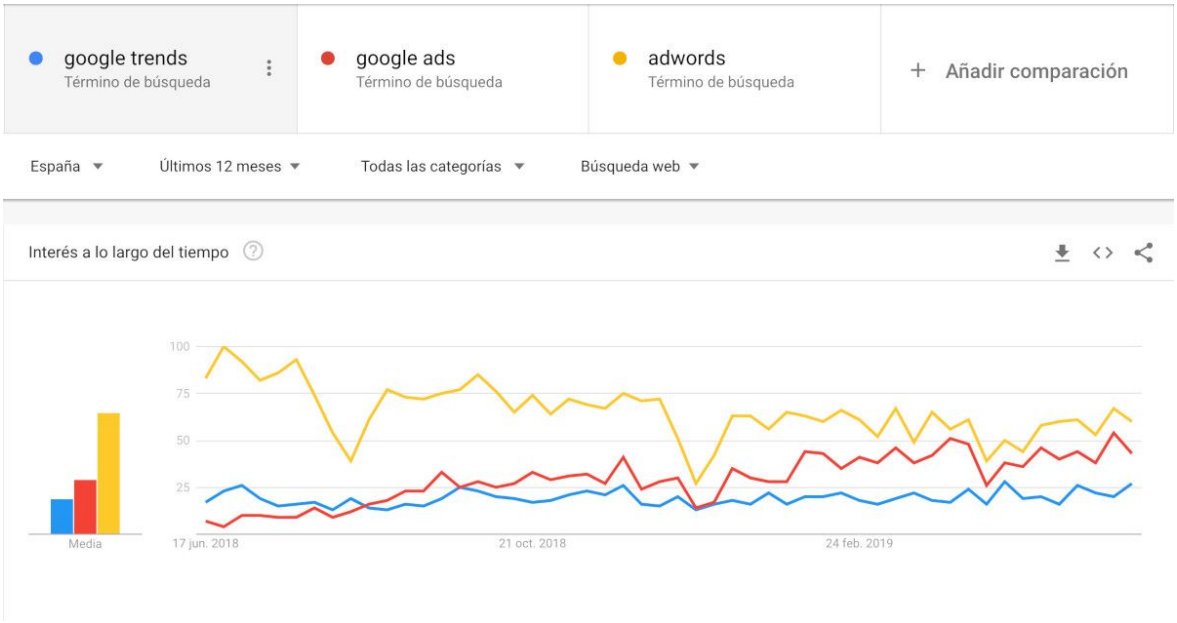


Figura 115. Ejemplo de gráfico de frecuencia de búsquedas en Google Trends (Anon., 2020)

3.9. Limitaciones metodológicas

A continuación, se pueden leer una serie de limitaciones metodológicas que se han producido debido a las características de las API a las que se acude en busca de información o a las características inherentes de la metodología en sí.

La función que accede y devuelve los datos de la API de Google Analytics solo permite elegir fechas comprendidas entre los 17 días anteriores a la fecha en que se ejecuta, por lo que el análisis se comprenderá en los 14 días anteriores, sumando un total de 14 días naturales. Pese a que esta limitación no está especificada en la documentación de la API, se pudo averiguar realizando pruebas con la función oficial de la API y, en cuanto el rango de fechas supera los 17 días, los valores se vuelven 0.

Debido a la anterior limitación, no se dispone de un histórico mediante el cual saber las visitas que recibió un artículo en los días posteriores a su publicación, un indicador que resultaría muy interesante para averiguar cuál fue su impacto en el momento en el que fue publicado. Esto se une a la característica implícita de las noticias de última hora, que solo suelen atraer tráfico en los días posteriores a su publicación, por lo que la mayoría de las noticias publicadas en el pasado presentan un número de páginas vistas igual a cero debido a su poca repercusión en el momento presente.

El apartado de publicidad digital de la metodología propuesta debe adecuarse al formato y sistema que utilice el caso de uso en el que se quiera aplicar. En el caso que nos ocupa, se utiliza Google AdSense con la modalidad de anuncios automáticos, lo cual ha servido para descartar variables como la localización y tamaño de los anuncios, así como su variabilidad según las temáticas principales del sitio web. De esta manera, no se ha podido estudiar las características propias de los anuncios, centrándose solo en su visualización, CTR y estimación de ingresos.

En cuanto a la versión estándar de la API de Twitter, los límites y cuotas supusieron un desafío a la hora de recolectar y medir las tendencias de los términos a estudiar. La función de búsqueda estándar presenta un máximo de 180 peticiones en franjas de 15 minutos, con un máximo de 100 tuits por petición, por lo que impone un máximo de 18.000 peticiones por cada ejecución del algoritmo. Una vez superado este número, el resto de los términos no pueden recibir datos de la API y sus métricas no se rellenan. Para evitarlo, se impuso un límite de 50 peticiones por cada término a estudiar en la misma ejecución, siendo 25 en su formato de cadena normal y 25 en su formato hashtag. El límite de tuits a estudiar por término es, por tanto, de 5.000, ignorando el potencial de este más allá de este número

(Twitter Developers, s.f.). Debido a estas decisiones, si se supera el máximo de 2.500 tuits en su formato de cadena normal o en su formato hashtag, el conjunto devuelto es una muestra aleatoria del conjunto total de tuits que incluyan ese término.

Es importante recordar, además, que también se ignora una parte del potencial temporal ya que la versión estándar de la API de Twitter solo devuelve resultados de una muestra de los últimos 7 días (Twitter Developers, s.f.).

Sería interesante también añadir un apartado de Conversación en los indicadores de Twitter tanto los pertenecientes a la cuenta del medio como en el análisis de tendencias, incorporando indicadores que muestren la conversación que se ha producido gracias a los tuits relacionados con el contenido. Sin embargo, tanto el número de respuestas (reply_count) como el número de citas (quote_count) no están disponibles en la versión estándar de la API de Twitter (Twitter Developers, s.f.).

Precisamente estas métricas son parte de las nuevas características de la Twitter API v2, estando presentes en el objeto Tweet (Twitter Developers, s.f.) junto con otras como la de los clics en los enlaces, otra métrica que también sería interesante tener en cuenta (Twitter Developers, s.f.). Esta nueva versión de la API de Twitter todavía está en fase de desarrollo y con Acceso Temprano disponible. Permite unos límites diferentes y funciones más avanzadas, así como otras características como las URL finales de las acortadas en el objeto JSON, un mejor filtro de Spam, detección de entidades y anotaciones, etc (Twitter Developers, s.f.).

El procesamiento de los términos en sí también presenta una serie de desafíos que escapan a la presente tesis. Uno de ellos es el tratamiento de los términos con nombres muy genéricos o que puedan tener más de un significado, como podría ser el caso de una película titulada *65*. El análisis de la tendencia relacionada con dicho término se complica muchísimo debido a la gran cantidad de ruido que incluye la búsqueda del término, en la cual la mayoría de los tuits devueltos no están realmente relacionados con este.

Otro desafío que destacar es que la selección de los términos relacionados es manual. Esto puede provocar que el redactor olvide incluir determinados términos o que se equivoque al escribirlos. También se puede producir el uso de etiquetas o términos en un artículo que en realidad no están estrictamente relacionados con el contenido de este. Un ejemplo podría ser el tráiler de una película que ha sido publicado por una famosa plataforma de contenidos. Los términos relacionados podrían ser el título de la película e incluso el actor o la actriz protagonista, pero la plataforma de contenidos en sí no forma parte del mensaje principal del contenido, como sí lo sería si el artículo tratara directamente

de esta. De todo ello se puede extrapolar que sería conveniente realizar una selección de los términos relacionados a estudiar ya que deben presentar una relación directa con el mensaje principal del contenido. De este modo se facilita el análisis de las tendencias de este, descartando tendencias que no afectan al contenido de manera directa sino indirecta. Esta selección de términos vinculantes, que en esta ocasión se ha realizado a juicio del autor del artículo, o incluso el estudio de tendencias secundarias de manera ponderada podrían servir para dar un contexto más rico al estudio en investigaciones futuras. La selección de temáticas del estudio actual se explica en el apartado 3.3.

La propia acotación temporal del estudio representa otra limitación metodológica, ya que se circunscribe a unas fechas concretas en las cuales pueden haberse dado días festivos y otras singularidades que hayan podido afectar global o específicamente a los datos.

Este estudio también se enfrenta a la limitación de que hay factores que pueden afectar al éxito o no de un contenido y que, sin embargo, no son medibles. Por ejemplo, hay eventos globales que pueden afectar al uso de Twitter, consumo de información e incluso origen de esta en sectores afectados por dichos eventos. Eventos como las pandemias, los terremotos o los conflictos bélicos pueden influenciar e incluso interrumpir tanto el flujo de información como el transcurso de las tendencias y el uso de redes sociales en general.

Un factor que también limita esta metodología es el hecho de no disponer de acceso directo a las fuentes originales de los contenidos, puesto que así se produce una dependencia directa en el flujo de información de las fuentes que primero lo comuniquen, como es el caso de las noticias de última hora. Si se produjera una interrupción editorial o hubiera equivocaciones en la comunicación original, la metodología se vería afectada bien por tener menos contenido del que alimentarse o, aún peor, realizaría cálculos en base a un contenido erróneo o falso.

Finalmente, es necesario tener en cuenta la caducidad implícita de la metodología al depender de indicadores clave de rendimiento y los límites propios de Twitter y Google Analytics. Los estudios posteriores deberán adecuarse a los mismos para adaptar la metodología a otros casos de uso.

4. Resultados de la investigación

4.1. Datos

4.1.1. Resumen

Se han almacenado los siguientes datos en total:

Tabla 2

Cantidades de datos de las tablas de la base de datos

Tabla de MySQL	Columnas	Filas	Total
tesis_followers	2	15.889	
tesis_hometimeline	8	2.333	
tesis_hometimeline_other	8	1.360	
tesis_posts	16	12.286	
tesis_stats	50	528	
tesis_terms	16	2.558	
Total			325.226

Se han almacenado, por tanto, 325.226 datos, aunque los que se van a analizar a continuación son los pertenecientes a la tabla tesis_stats, puesto que contienen los datos finales de los estudios realizados por cada noticia de última hora publicada en el medio. El resto sirven para procesar y calcular los datos que se almacenan finalmente en la tabla tesis_stats.

En cuanto a los artículos y términos analizados según la sección a la que pertenecen, se puede observar en la Tabla *3*. Se muestra el número de artículos publicados en cada fase, el número de artículos analizados (los publicados y los relacionados con éstos), el número de artículos analizados con tráfico en los 14 días posteriores a la publicación del artículo que suscitó su análisis, el número de artículos analizados con datos de retweets y favoritos en la cuenta de Twitter del medio y los términos o temáticas analizados que están relacionados con los artículos publicados.

Tabla 3

Cantidades de artículos y términos analizados en la fase 1

Sección	Artículos publicados	Artículos analizados	Artículos analizados con tráfico reciente	Artículos analizados con datos en Twitter	Términos analizados
Cine	187	2.973	706	1.005	464
Cómics	16	820	333	299	44
Frikinoticias	4	71	16	22	6
Literatura	4	14	7	9	10
Series	104	2.306	530	908	215
Videojuegos	35	851	181	454	78
Totales	350	7.035	1.773	2.697	817

Tal y como se puede observar en la Tabla *3*, en la Fase 1 se han analizado 7.035 artículos a partir de 350 artículos publicados, y 817 términos relacionados con éstos a través de la funcionalidad de etiquetas de WordPress, el sistema de contenidos utilizado por el medio.

Hay solo tres secciones con más de 30 artículos publicados, cantidad mínima de una muestra para la cual se podrían extraer conclusiones válidas (González Rodríguez, et al., 2013), y son: cine, series y videojuegos. Por tanto, solo se realizará un análisis específico para estas.

Además, se ha propuesto analizar también dos subcategorías de artículos para comprobar la evolución de los datos conforme aumenta la especificación de su selección, ya sea por temática (superhéroes) o por contenido (tráileres).

Tabla 4

Cantidades de artículos y términos analizados de subcategorías en la fase 1

Subcategoría	Artículos publicados	Artículos analizados	Artículos analizados con tráfico reciente	Artículos analizados con datos en Twitter	Términos analizados
Superhéroes	25	833	200	189	66
Tráileres	101	761	261	296	194

Se puede comprobar en la Tabla *4* que solo la subcategoría Tráileres tiene un número de artículos publicado suficiente (igual o mayor a 30) para dar lugar a la cantidad mínima de una muestra para la cual se podrían extraer conclusiones válidas (González Rodríguez, et al., 2013). Por lo tanto, solo se realizará un análisis específico para Tráileres.

Respecto a la fase 2, se han observado las siguientes cifras:

Tabla 5

Cantidades de artículos y términos analizados en la fase 2

Sección	Artículos publicados	Artículos analizados	Artículos analizados con tráfico reciente	Artículos analizados con datos en Twitter	Términos analizados
Cine	80	1.659	380	549	210
Cómics	5	630	251	219	13
Frikinoticias	0	0	0	0	0
Literatura	3	281	40	120	12
Series	74	1.590	378	486	161
Videojuegos	16	1.091	198	461	66
Totales	178	5.251	1.247	1.835	462

Tal y como se puede observar en la Tabla *5*, en la Fase 2 se han analizado 5.251 artículos a partir de 178 artículos publicados, y 462 términos relacionados con éstos a través de la funcionalidad de etiquetas de WordPress.

Tabla 6

Cantidades de artículos y términos analizados de subcategorías en la fase 2

Subcategoría	Artículos publicados	Artículos analizados	Artículos analizados con tráfico reciente	Artículos analizados con datos en Twitter	Términos analizados
Tráileres	57	680	192	265	127

En cuanto a las subcategorías de artículos, se han analizado 680 artículos de la subcategoría por contenido de Tráileres, a partir de 57 artículos publicados, y 127 términos relacionados con éstos a través de la funcionalidad de etiquetas de WordPress.

El conjunto total de los datos se puede descargar desde figshare[42], para hacer posible así la réplica del estudio en el futuro. El conjunto de datos se ha publicado con el título "Dataset from PhD Thesis" en la categoría Social and Community Informatics, con licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) y las palabras clave: *digital journalism*, *machine learning*, *cyberjournalism*, *altmetrics*, *web analytics*, *Twitter analytics*, *trend analysis*, *trend prediction*, *digital content optimization*, *cybermetrics* y *web advertising*.

[42] https://figshare.com/articles/dataset/Dataset_from_the_PhD_Thesis/16553061

4.1.2. Variables

Este estudio parte de una base de 35 valores divididos en las siguientes categorías:

- Analítica del contenido en la web (11 variables)
    - uniquepageviews_total: total de páginas vistas,
    - uniquepageviews_mean: promedio de páginas vistas,
    - entrancerate_mean: promedio de la ratio de entradas,
    - avgtimeonpage_mean: promedio de duración de las visitas,
    - exitrate_mean: promedio de la ratio de salidas,
    - pageviewspersession_mean: promedio de páginas vistas por sesión,
    - adsense_adunitsviewed_total: total de anuncios visualizados,
    - adsense_adunitsviewed_mean: promedio de anuncios visualizados,
    - adsense_viewableimpressionpercent_mean: promedio de la ratio de visualizaciones de los anuncios,
    - adsense_ctr_mean: promedio de la ratio de clics de los anuncios,
    - adsense_ecpm_mean: estimación de ingresos de los anuncios por cada 1000 páginas vistas.
- Analítica del contenido en la cuenta de Twitter del medio (4 variables)
    - retweet_count_total: total de retuits,
    - retweet_count_mean: promedio de retuits,
    - favorite_count_total: total de favoritos,
    - favorite_count_mean: promedio de favoritos.
- Análisis de tendencias el día de la publicación del artículo (10 variables)
    - terms_ini_num_tweets: total de tuits sobre los términos en el día de la publicación,

    - terms_ini_retweet_count_total: total de retuits sobre los términos en el día de la publicación,
    - terms_ini_retweet_count_mean: promedio de retuits sobre los términos en el día de la publicación,
    - terms_ini_favorite_count_total: total de favoritos sobre los términos en el día de la publicación,
    - terms_ini_favorite_count_mean: promedio de favoritos sobre los términos en el día de la publicación,
    - terms_ini_followers_talking_rate: ratio de seguidores de la cuenta de Twitter del medio que han publicado recientemente un tuit hablando sobre los términos en el día de la publicación,
    - terms_ini_user_num_followers_mean: promedio de seguidores de usuarios que han hablado de los términos en el día de la publicación,
    - terms_ini_user_num_tweets_mean: promedio de tuits publicados por los usuarios que han hablado de los términos en el día de la publicación,
    - terms_ini_user_age_mean: edad promedio en días de los usuarios que han hablado de los términos en el día de la publicación,
    - terms_ini_url_inclusion_rate: ratio de inclusión de URL de los tuits que hablan sobre los términos en el día de la publicación.
- Análisis de tendencias 14 días de la publicación del artículo (10 variables)
    - terms_end_num_tweets: total de tuits sobre los términos 14 días después de la publicación,
    - terms_ini_retweet_count_total: total de retuits sobre los términos 14 días después de la publicación,
    - terms_ini_retweet_count_mean: promedio de retuits sobre los términos 14 días después de la publicación,
    - terms_ini_favorite_count_total: total de favoritos sobre los términos 14 días después de la publicación,

- terms_ini_favorite_count_mean: promedio de favoritos sobre los términos 14 días después de la publicación,
- terms_ini_followers_talking_rate: ratio de seguidores de la cuenta de Twitter del medio que han publicado recientemente un tuit hablando sobre los términos 14 días después de la publicación,
- terms_ini_user_num_followers_mean: promedio de seguidores de usuarios que han hablado de los términos 14 días después de la publicación,
- terms_ini_user_num_tweets_mean: promedio de tweets publicados por los usuarios que han hablado de los términos 14 días después de la publicación,
- terms_ini_user_age_mean: edad promedio en días de los usuarios que han hablado de los términos 14 días después de la publicación,
- terms_ini_url_inclusion_rate: ratio de inclusión de URL de los tuits que hablan sobre los términos 14 días después de la publicación.

4.1.2.1. Variables de éxito

Con la siguiente metodología se podría estudiar la posible predicción de las variables cuyos datos se han obtenido, ya que se analiza la correlación, variabilidad y posible cálculo de estas mediante una regresión lineal múltiple en base a las variables que se cuentan en el momento de publicación de un artículo.

Esto significa que se establecerán como variables o escenarios de éxito las que puedan interesar al equipo editorial de un medio, como podría ser detectar qué tendencias suelen tener un mayor número de tuits 14 días después (se mantienen estables o crecen con el tiempo), qué artículos suelen tener un porcentaje menor de salida, un mayor número de páginas vistas únicas o qué artículos suscitan una mejor respuesta en forma de retuits o favoritos en la propia cuenta de Twitter del medio.

En el caso de esta tesis, se han definido como variables de éxito las siguientes:

- Páginas vistas únicas (total): uniquepageviews_total,
- AdSense eCPM (promedio): adsense_ecpm_mean,
- Duración de la visita (promedio): avgtimeonpage_mean,
- Páginas vistas por sesión (promedio): pageviewspersession_mean,

- Número de retuits en la cuenta del medio (promedio): retweet_count_mean,
- Número de favoritos en la cuenta del medio (promedio): favorite_count_mean,
- Número de tuits de la tendencia 14 días después (total): terms_end_num_tweets,
- Número de retuits de la tendencia 14 días después (total): terms_end_retweet_count_total,
- Número de retuits de la tendencia 14 días después (promedio): terms_end_retweet_count_mean.

De esta manera, se seleccionan variables de éxito relacionadas con el tráfico (páginas únicas vistas), la obtención de ingresos (AdSense eCPM), la calidad de las sesiones (duración de la visita y páginas vistas por sesión), el éxito de los contenidos relacionados en la cuenta de Twitter del medio (número de retuits y de favoritos en la cuenta del medio) y la tendencia 14 días después (número de tuits como tamaño y número de retuits como amplitud).

4.1.2.2. Variables de predicción

La toma de decisiones a nivel editorial se produce antes de la publicación del artículo, por lo que solo se cuenta con los datos relacionados con el análisis de tendencias efectuado en el día de la publicación del artículo.

De esta manera, se cuenta inicialmente con las siguientes variables de predicción:

- Número de tuits de la tendencia inicial (total): terms_ini_num_tweets.
- Número de retuits de la tendencia inicial (total): terms_ini_retweet_count_total.
- Número de retuits de la tendencia inicial (promedio): terms_ini_retweet_count_mean.
- Número de favoritos de la tendencia inicial (total): terms_ini_favorite_count_total.
- Número de favoritos de la tendencia inicial (promedio): terms_ini_favorite_count_mean.
- Número de seguidores del medio que hablan sobre la tendencia inicial (total): terms_ini_followers_talking_rate.
- Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio): terms_ini_user_num_followers_mean.

- Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio): terms_ini_user_num_tweets_mean.
- Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio): terms_ini_user_age_mean.
- Ratio de inclusión de URLs en los tuits de la tendencia inicial: terms_ini_url_inclusion_rate.

4.1.2.3. Tabla de variables

Para facilitar la interpretación de la tesis, se puede ver a continuación la tabla de variables inicial al comienzo del estudio, la cual se irá actualizando conforme se avance en los diferentes análisis necesarios hasta obtener las ecuaciones de predicción.

Tabla 7

Lista inicial de variables de predicción y de éxito

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	adsense_ecpm_mean
terms_ini_retweet_count_mean	avgtimeonpage_mean
terms_ini_favorite_count_total	pageviewspersession_mean
terms_ini_favorite_count_mean	retweet_count_total
terms_ini_followers_talking_rate	favorite_count_total
terms_ini_user_num_followers_mean	terms_end_num_tweets
terms_ini_user_num_tweets_mean	terms_end_retweet_count_total
terms_ini_user_age_mean	terms_end_retweet_count_mean
terms_ini_url_inclusion_rate	

4.1.3. Objetivos estadísticos

El primer objetivo estadístico es aplicar el análisis de regresión lineal múltiple mediante el cual producir modelos en forma de ecuaciones lineales que contengan la combinación lineal de variables de predicción que predigan de manera óptima cada variable de éxito (Anon., 2017). El modelo de regresión lineal múltiple permite predecir en base a más de una variable de predicción, lo cual se ha utilizado en estudios basados en datos de analítica web y redes sociales como el estudio del efecto de las sesiones en los visitantes únicos y las páginas vistas únicas (Awichanirost & Phumchusri, 2020); la optimización para buscadores mediante variables como el tamaño de la web, la velocidad de carga, la condición de seguridad y el comportamiento del usuario (Drivas, et al., 2020); el efecto de la publicidad en el tráfico de la web (Moore, 2021); el impacto del marketing de redes sociales en la lealtad a una marca (Eren Erdogmus & Cicek, 2012) o el número de seguidores según los usuarios a los que se sigue, los tuits publicados y los favoritos (Sukunesan, et al., 2021).

Para realizar la regresión múltiple, según Barón López y Téllez Montiel (s.f.) es necesario que se cumplan ciertos requerimientos:

- Linealidad: Cada variable de éxito debe depender linealmente de las variables explicativas, si no, será necesario introducir en el modelo componentes no lineales. En el caso de este estudio, se han incluido en el modelo inicial variables cuya linealidad se intuye. Se demostrará así si están incluidas en las ecuaciones de regresión finales.
- Normalidad y equidistribución de los residuos: Los residuos son las diferencias entre los valores calculados por el modelo y los valores reales en la variable de éxito. Es necesario que los residuos sean pequeños y que, además, los mismos se distribuyan de modo normal y con la misma dispersión para cada combinación de las variables de predicción.
- Número de variables independientes: Incluir al menos 20 observaciones por cada variable independiente o de predicción, condición que se cumple gracias a la población del estudio de 350 artículos publicados.
- Colinealidad: Si dos variables están correlacionadas fuertemente, es posible que ninguna de las dos sea considerada significativa cuando si solo se incluye una de ellas, sí podría serlo. Es recomendable, por tanto, que las variables de predicción no tengan una correlación fuerte entre ellas.

- Datos anómalos: Es necesario poner especial cuidado en identificarlos y descartarlos si procede, ya que a veces se pueden deber a errores en la entrada de datos. Estos datos anómalos se han analizado en el caso de las variables de éxito, incluyendo todos estos salvo los de la variable AdSense eCPM (promedio), por no poder controlarla al venir dada de un precio decidido por Google AdSense.

En suma, se procederá a asegurar la normalidad y equidistribución de los residuos y a evitar la colinealidad, de manera que se pueda aplicar la regresión múltiple a la lista de variables del modelo. Además, una correlación de 0,7 o más entre dos variables sugiere que se debería eliminar una de las dos o combinarlas en una variable compuesta antes de realizar la regresión lineal múltiple (Anon., 2017).

Por otro lado, la homocedasticidad es una propiedad deseable en un modelo de regresión lineal, ya que permite realizar un modelo más fiable. Consiste en que la varianza de los errores sea constante a lo largo del tiempo (López, 2017), así que se prestará también atención a esta característica de las variables del modelo.

El segundo objetivo estadístico es aplicar el análisis de la regresión bionimial negativa o la de Poisson a las variables de éxito que tengan datos de conteo (enteros y sin números negativos). Según Bobbitt (2019), se trata de unos tipos de regresión que solo sirven para variables dependientes que consisten en datos de conteo, que en el caso de esta tesis corresponden a las variables de éxito que suponen un total de algún elemento. De esta manera se tratará de interpretar en qué medida una unidad más o una unidad menos de una variable independiente o de predicción es asociada con un cambio porcentual en el conteo de una variable dependiente o de éxito.

El modelo de Poisson se puede utilizar con variables de respuesta como “me gusta”, comentarios, comparticiones y otras variables de conteo relacionadas con el ámbito de las redes sociales, como hicieron Labrecque et al. (2020) en su estudio del efecto del uso de pronombres en este tipo de variables. Otros estudios que utilizan datos de analític web o redes sociales son el de Salminen et al. (2020) al analizar el impacto en los resultados de los anuncios según datos del usuario, el de Guillory et al. (2020) al estudiar el efecto de las publicaciones de pago en redes socailes en las inscripciones a un programa de salud o el de Rödlund (2020) al analizar el uso de las redes sociales y el nivel de estrés percibido. En el caso de esta tesis, por tanto, se aplicará al número de páginas vistas, tuits y retuits.

La binomial negativa, en cambio, puede captar parte de la varianza que no identifica la regresión de Poisson, ya que para realizar esta, es necesario que no exista una sobre dispersión (o extra-varianza Poisson) en los datos, es decir, una varianza mucho mayor a

la media. Si esto se produce, provoca una subestimación en los errores estándares de los coeficientes y, por tanto, que se interprete una significación estadística de factores que realmente no están asociados con lo que se está analizando. En los casos en los que esto ocurre, y la sobre dispersión suele ser muy frecuente en los fenómenos recurrentes como son la visualización de contenido online, es más apropiado utilizar la binomial negativa que la de Poisson (Navarro, et al., 2001). Algunos ejemplos de estudios de este ámbito son el de Eddy et al. (2021) al analizar el uso de redes sociales por parte de las marcas de deporte para interactuar con sus seguidores, el de Gabarron et al. (2020) al estudiar el comportamiento de los pacientes de diabetes en redes sociales o el de Jackson et al. (2018) al analizar la asociación entre la frecuencia de retuit y el número de hashtags, menciones y enlaces.

No obstante, se comprobará la sobre dispersión para seleccionar una de las dos para así adecuar la metodología de la tesis al caso de estudio. Para ello, según explica Meyer (2018), se realizará la dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utilizará la regresión de Poisson, mientras que, si no es así, se utilizará la biniomial negativa.

Se entiende que una predicción es óptima cuando la suma de las diferencias al cuadrado entre el valor predicho y el valor real de la variable de éxito es mínima, siendo estas diferencias los errores de predicción. El modelo, por tanto, debe minimizar la suma de estos errores cuadrados (Anon., 2017).

4.2. Fase 1. Análisis de los datos de entrenamiento

Se ha efectuado un análisis de los artículos de tipología de noticia de última hora publicados entre el 23 de septiembre y el 22 de noviembre de 2020 en el medio Hello Friki[43], sumando un total de 350 artículos publicados.

Las noticias de última hora se clasifican en las siguientes categorías principales: cine, series, videojuegos, literatura, cómics y frikinoticias. Bajo el paraguas de estas categorías, cada artículo está relacionado con una serie de términos seleccionados por el redactor y publicados en la web en forma de etiquetas, sirviendo estas para el análisis de tendencias.

Se ha seguido el siguiente proceso:

1. Variables de éxito.
    a. Analizar cada variable.
    b. Evaluar la normalidad y equidistribución de los residuos.
    c. Normalizar las variables si es necesario.
    d. Filtrar correlaciones fuertes.
2. Variables de predicción.
    a. Analizar cada variable.
    b. Evaluar la normalidad, equidistribución de los residuos y homocedasticidad.
    c. Normalizar las variables si es necesario.
    d. Filtrar correlaciones fuertes.
    e. Realizar un análisis de componentes principales.
3. Regresión lineal múltiple por cada variable de éxito para extraer la fórmula de predicción.
4. Regresión binomial negativa o de Poisson.
    a. Comprobar cuál de las dos es más adecuado aplicar.
    b. Realizar un filtro de alta correlación entre variables de predicción originales.
    c. Hacer una regresión de Poisson por cada variable de éxito de conteo para extraer la fórmula de predicción.

Con las ecuaciones resultantes de las regresiones se podrá realizar la predicción de los valores en la fase 2 y comprobar hasta qué punto se ha conseguido acertar el valor de las variables de éxito.

[43] https://www.hellofriki.com/

4.2.1. Análisis de todos los artículos

En primer lugar, se van a analizar todos los artículos en un conjunto único, de manera que se puedan extraer conclusiones a nivel general del medio y se pueda comprobar si luego difieren según la sección a la que pertenezca el artículo.

4.2.1.1. Variables de éxito

El objetivo de esta fase es la predicción de los valores de éxito mediante el resto de los indicadores. Por ello, es importante comenzar por analizar estos escenarios por separado para ver tanto sus características como tratar de describir su comportamiento anómalo, si lo tuvieran.

a) Páginas vistas únicas (total)

Este escenario de éxito se ve identificado por la columna uniquepageviews_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 350 valores con un rango de entre 1 y 704.

Presenta un sesgo estandarizado de 44,9048 y una curtosis estandarizada de 177,779. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

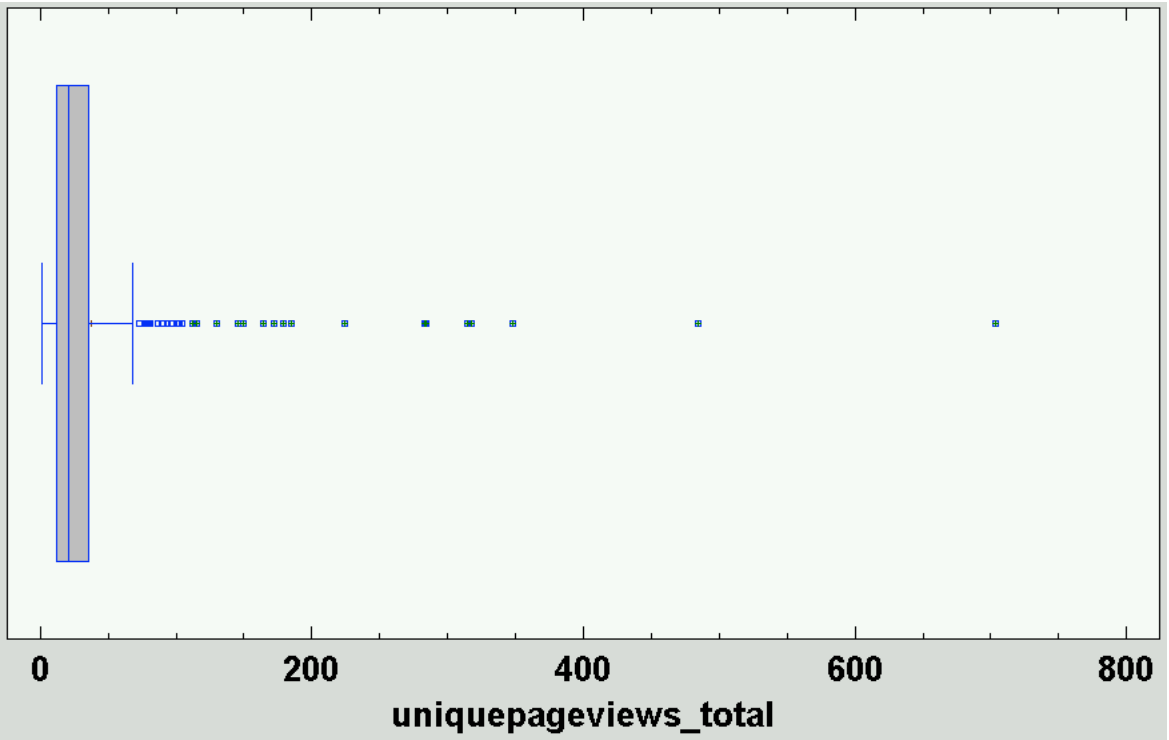


Figura 116. Todos: Gráfico de Caja y Bigotes para el valor uniquepageviews_total

En la Figura 1 se puede comprobar que existen valores anómalos de tipo extremo de 115 páginas vistas o más.

Estos valores indican artículos que han tenido un tráfico mucho mayor de lo que se podría esperar según el resto de los artículos, por lo que podría considerarse su estudio por separado de manera que se trate de explicar y predecir su éxito.

Los valores anómalos corresponden a noticias de alto impacto como son: regresos de sagas muy famosas, retrasos de estrenos de películas muy esperadas (generando una tendencia que consta de muchos términos), cancelaciones de series, novedades de una película muy esperada y novedades editoriales de dos de las editoriales más famosas de cómics.

Si analizamos estas tendencias con más detalle, podemos ver que constan de términos con un uso presente en todo el periodo estudiado en las búsquedas de Google. Utilizando la herramienta Google Trends podemos comprobarlos:

- Un artículo que consta de un término relacionado: HeroQuest publicado el día 24 de septiembre. Como se puede ver en la Figura 117, este término presenta tendencias constantes de búsqueda, pero con un pico en el momento de la publicación.

Figura 117. Análisis del término HeroQuest en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Amazon Prime, Craig Rosenberg y The Boys, publicado el día 25 de septiembre. En el caso del término Craig Rosenberg, no tiene muchas búsquedas y sí presenta un aumento en los días posteriores a la publicación del artículo. Amazon Prime es constante salvo un pico en un momento que no coincide con la publicación, y The Boys presenta una tendencia mayor en la época de emisión de la serie *The Boys*.

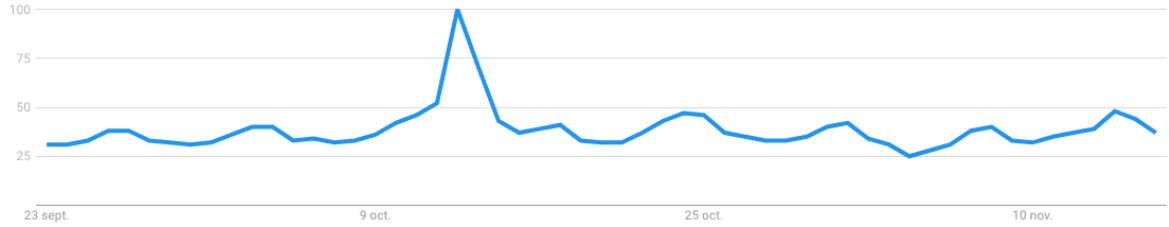


Figura 118. Análisis del término Amazon Prime en Google vía Google Trends en el periodo analizado

Figura 119. Análisis del término Craig Rosenberg en Google vía Google Trends en el periodo analizado

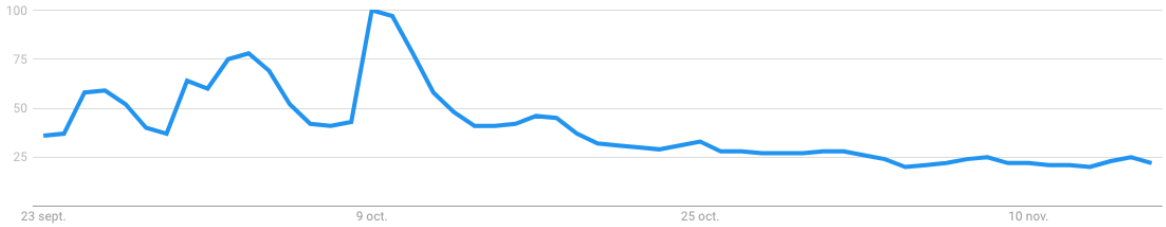


Figura 120. Análisis del término The Boys en Google vía Google Trends en el periodo analizado

- Un artículo que consta de un término relacionado: Raised by Wolves, publicado el día 28 de septiembre. Como se puede ver en las siguientes figuras, este término presenta una tendencia mayor en la época de emisión de la serie *Raised by Wolves*.

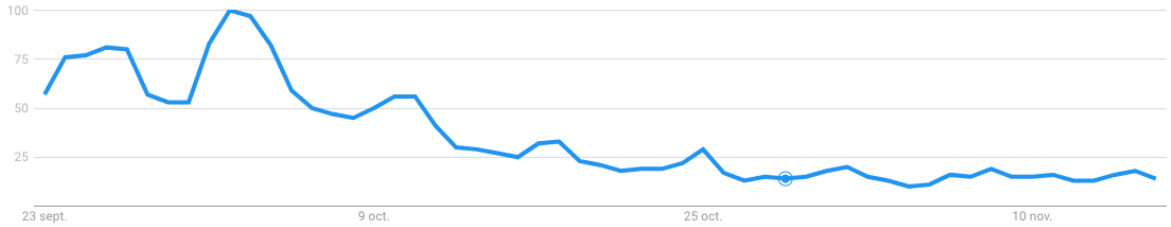


Figura 121. Análisis del término Raised by Wolves en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Animales fantásticos 3 y Eddie Redmayne, publicado el día 29 de septiembre. Eddie Redmayne sí presenta un leve pico el día de la publicación.

Figura 122. Análisis del término Animales fantásticos 3 en Google vía Google Trends en el periodo analizado

Figura 123. Análisis del término Eddie Redmayne en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Conan y Netflix, publicado el día 1 de octubre. Como se puede ver en las siguientes figuras, estos términos presentan tendencias constantes de búsqueda y sin ningún pico en el momento de la publicación.

Figura 124. Análisis del término Conan en Google vía Google Trends en el periodo analizado

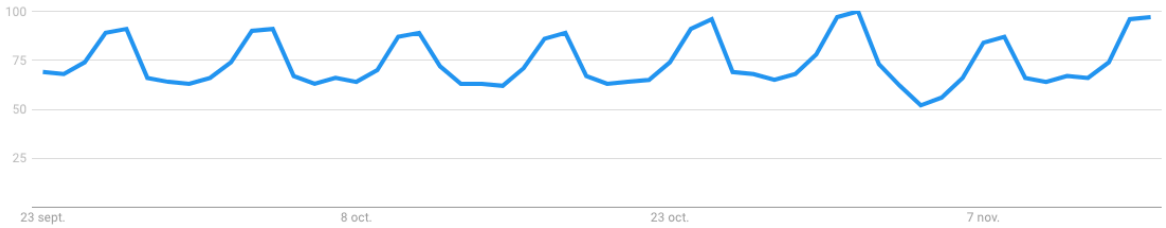


Figura 125. Análisis del término Netflix en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: The CW y The Flash, publicado el día 5 de octubre. Como se puede ver en las siguientes figuras, estos términos presentan tendencias constantes de búsqueda y sin ningún pico en el momento de la publicación.

Figura 126. Análisis del término the CW en Google vía Google Trends en el periodo analizado

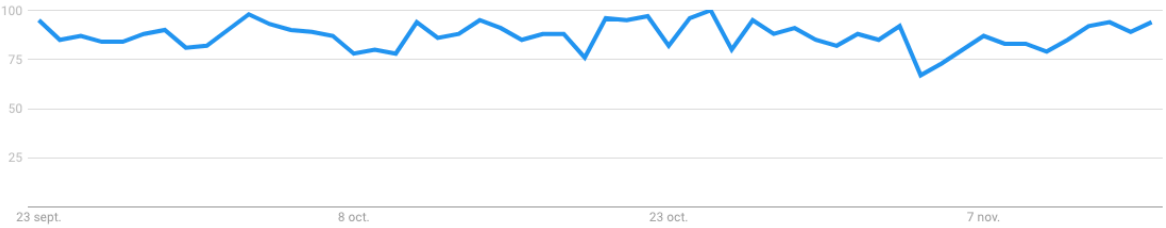


Figura 127. Análisis del término the Flash en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Black Adam, Dune, Matrix 4, Minecraft, Shazam! Fury of the Gods, The Batman, The Flash y Warner Bros., publicado el día 6 de octubre. El de The Flash está también presente en el anterior artículo mencionado. Términos como Dune y Matrix 4 sí que presentan un pico en el momento de la publicación. Shazam! Fury of the Gods no tiene datos de búsqueda suficientes.

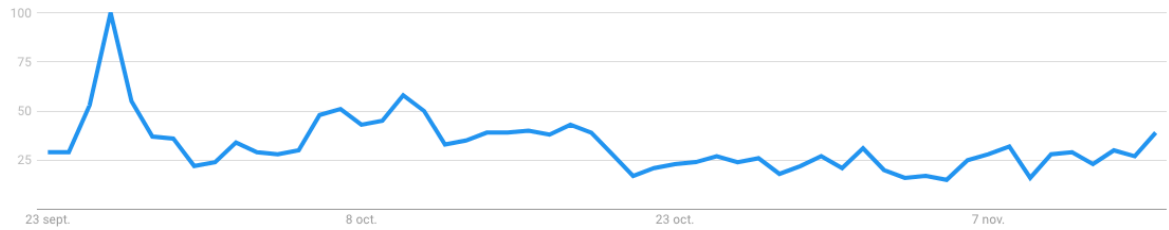


Figura 128. Análisis del término Black Adam en Google vía Google Trends en el periodo analizado

Figura 129. Análisis del término Dune en Google vía Google Trends en el periodo analizado

Figura 130. Análisis del término Matrix 4 en Google vía Google Trends en el periodo analizado

Figura 131. Análisis del término Minecraft en Google vía Google Trends en el periodo analizado

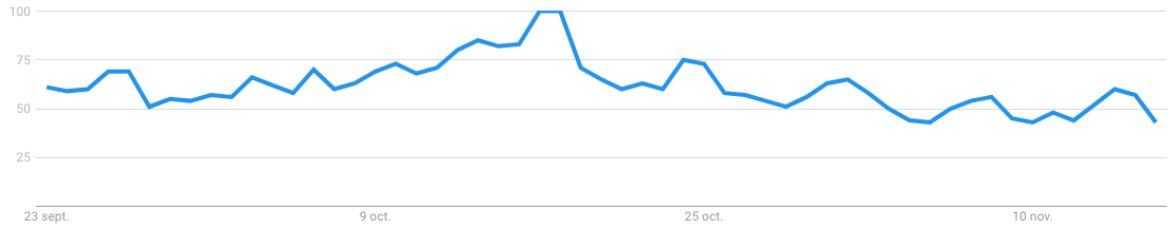


Figura 132. Análisis del término The Batman en Google vía Google Trends en el periodo analizado

Figura 133. Análisis del término Warner Bros. en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Carly Mensch, Glow, Liz Flahive y Netflix, publicado el día 6 de octubre. Netflix ya se ha mostrado en uno de los artículos anteriores. Carly Mensch y Liz Flahive tienen muy pocas búsquedas y sí muestra una subida en el momento de la publicación.

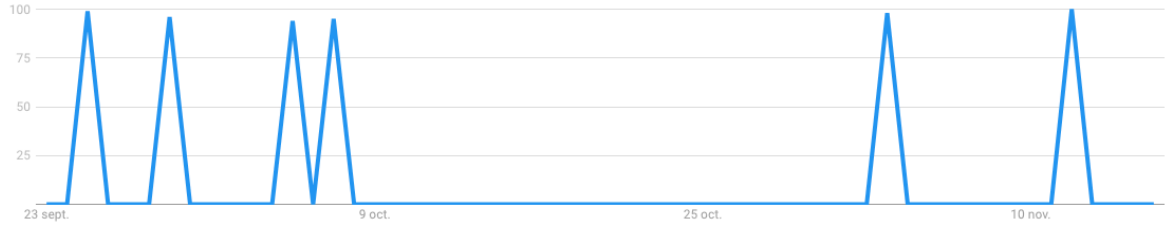


Figura 134. Análisis del término Carly Mensch en Google vía Google Trends en el periodo analizado

Figura 135. Análisis del término Glow en Google vía Google Trends en el periodo analizado

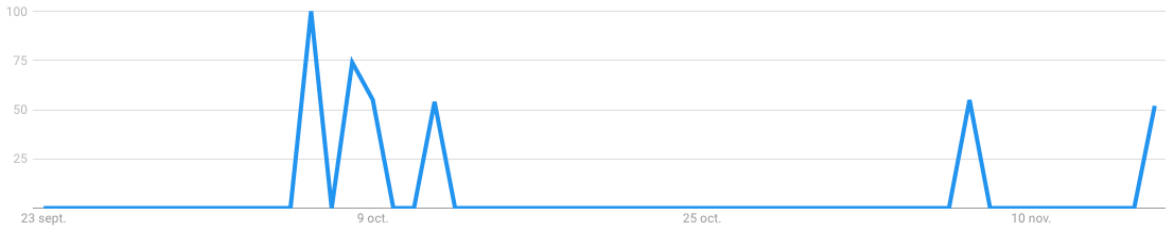


Figura 136. Análisis del término Liz Flahive en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Ecc Ediciones y noviembre 2020, publicado el día 7 de octubre.

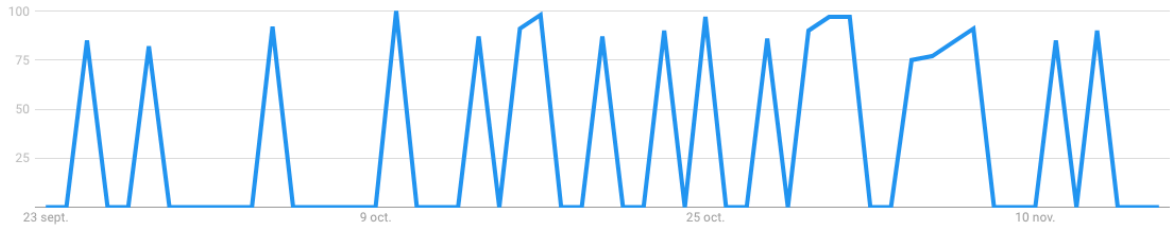


Figura 137. Análisis del término Ecc Ediciones en Google vía Google Trends en el periodo analizado

Figura 138. Análisis del término noviembre 2020 en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Benedict Cumberbatch, Doctor Strange, Marvel y The Amazing Spider-Man 3, publicado el día 12 de octubre.

Figura 139. Análisis del término Benedict Cumberbatch en Google vía Google Trends en el periodo analizado

Figura 140. Análisis del término Doctor Strange en Google vía Google Trends en el periodo analizado

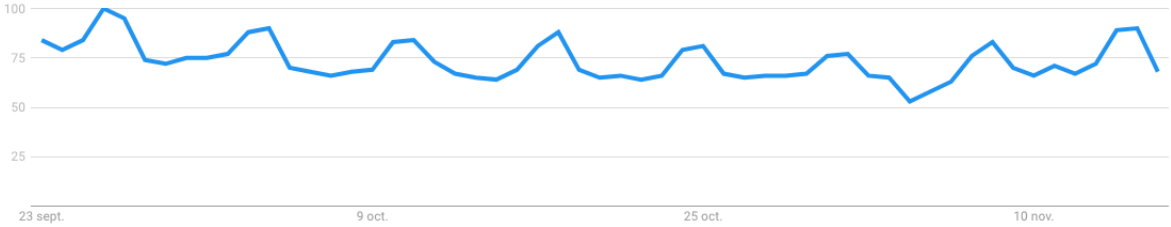


Figura 141. Análisis del término Marvel en Google vía Google Trends en el periodo analizado

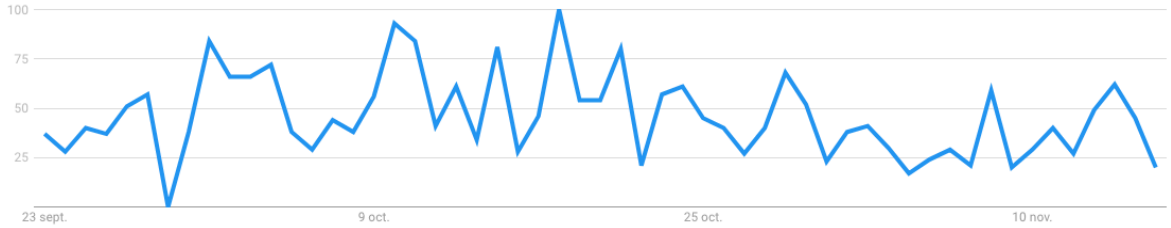


Figura 142. Análisis del término The Amazing Spider-Man 3 en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Nacon, PC, Revolution, Windows y Xbox, publicado el día 14 de octubre.

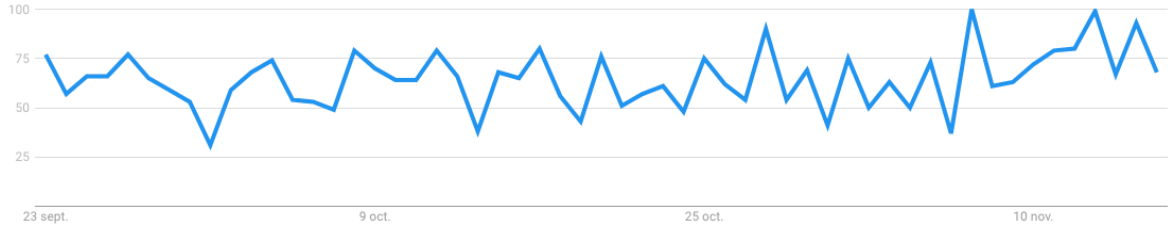


Figura 143. Análisis del término Nacon en Google vía Google Trends en el periodo analizado

Figura 144. Análisis del término PC en Google vía Google Trends en el periodo analizado

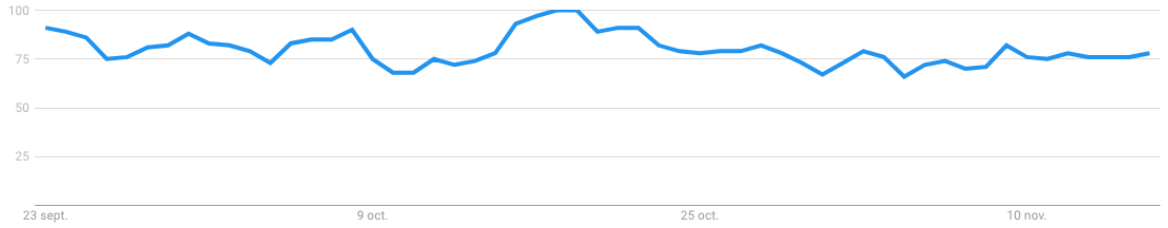


Figura 145. Análisis del término Revolution en Google vía Google Trends en el periodo analizado

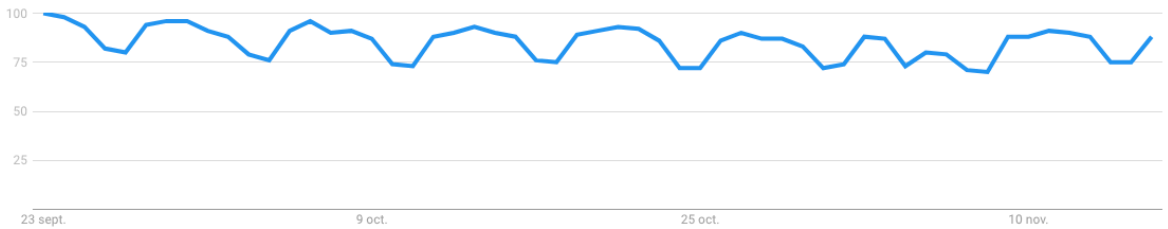


Figura 146. Análisis del término Windows en Google vía Google Trends en el periodo analizado

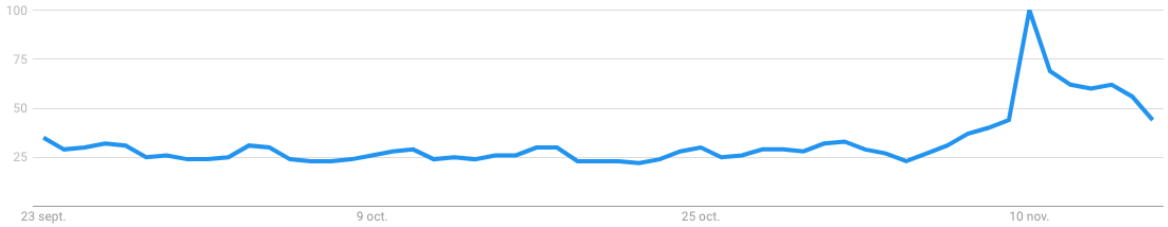


Figura 147. Análisis del término Xbox en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Buenos días tristeza, noviembre 2020 y Planeta Cómic, publicado el día 1 de noviembre. El término noviembre 2020 ya se ha mostrado en uno de los artículos anteriores. El término Planeta Cómic presenta un pico dos días después del día de publicación del artículo.

Figura 148. Análisis del término Buenos días tristeza en Google vía Google Trends en el periodo analizado

Figura 149. Análisis del término Planeta Cómic en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Ecc Ediciones y Tortugas Ninja, publicado el día 2 de noviembre. El término Ecc Ediciones ya se ha mostrado en uno de los artículos anteriores.

Figura 150. Análisis del término Tortugas Ninja en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Jurassic World Evolution: Complete Edition y Nintendo Switch, publicado el día 4 de noviembre. Jurassic World Evolution: Complete Edition tiene pocas búsquedas y sí presenta un pico en los días cercanos a la publicación del artículo.

Figura 151. Análisis del término Jurassic World Evolution: Complete Edition en Google vía Google Trends en el periodo analizado

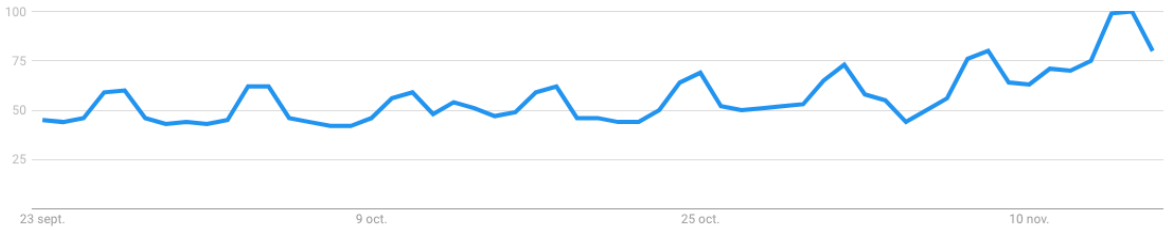


Figura 152. Análisis del término Nintendo Switch en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: Sobrenatural y Supernatural, publicado el día 5 de noviembre. El término Supernatural sí que presenta un pico el día de la publicación del artículo.

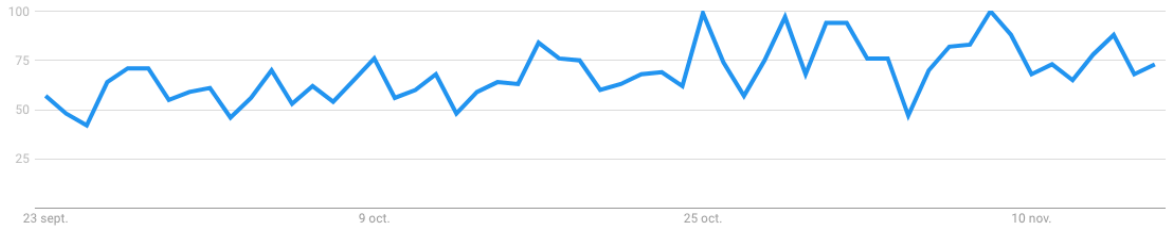


Figura 153. Análisis del término Sobrenatural en Google vía Google Trends en el periodo analizado

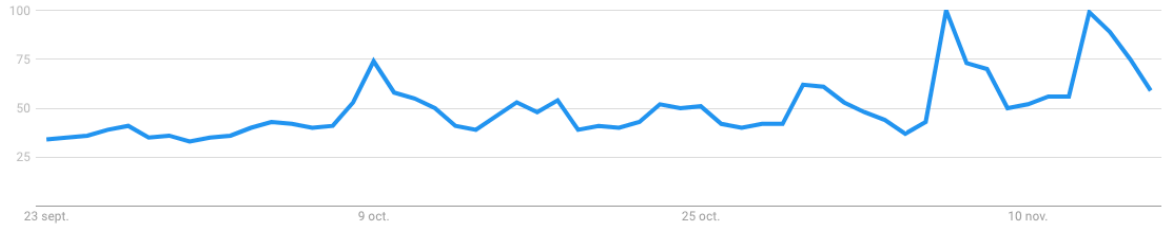


Figura 154. Análisis del término Supernatural en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: diciembre 2020, Planeta Cómic y Saga, publicado el día 10 de noviembre. Planeta Cómic ya ha sido analizado en un artículo anterior.

Figura 155. Análisis del término diciembre 2020 en Google vía Google Trends en el periodo analizado

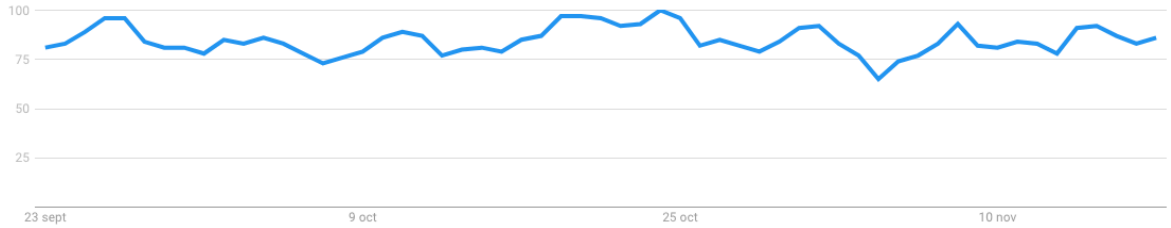


Figura 156. Análisis del término Saga en Google vía Google Trends en el periodo analizado

- Un artículo que consta de los términos relacionados: The CW y Wonder Girl, publicado el día 17 de noviembre. The CW ya ha sido analizado en un artículo anterior.

Figura 157. Análisis del término Wonder Girl en Google vía Google Trends en el periodo analizado

La mayoría de los términos presentan una cantidad de búsquedas constante o a intervalos constantes, salvo otros cuyo significado implican un periodo temporal concreto (noviembre 2020 y diciembre 2020). Sí que se repiten en varios de ellos los términos The Flash, Ecc Ediciones, noviembre 2020, Planeta Cómic y The CW.

No se ha apreciado una presencia de picos abruptos el día de la publicación de los artículos que obtuvieron datos de tráfico anómalos, salvo en el caso de los artículos que presentan un número menor de búsquedas. Es probable que, al deberse a términos que normalmente no son buscados, cuando se produce una noticia de última hora relacionado con ellos, su tendencia aumente en forma de pico.

Todo lo anterior, junto con el conocimiento de la línea editorial del caso de estudio, indica que los picos probablemente se deben a tendencias de potencia alta y estable en el tiempo, es decir, conceptos que son famosos y a cuyas novedades los usuarios siguen y prestan especial atención.

b) AdSense eCPM (promedio)

Este escenario de éxito se identifica con la columna adsense_ecpm_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 2,65.

Presenta un sesgo estandarizado de 77,9369 y una curtosis estandarizada de 487,459. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

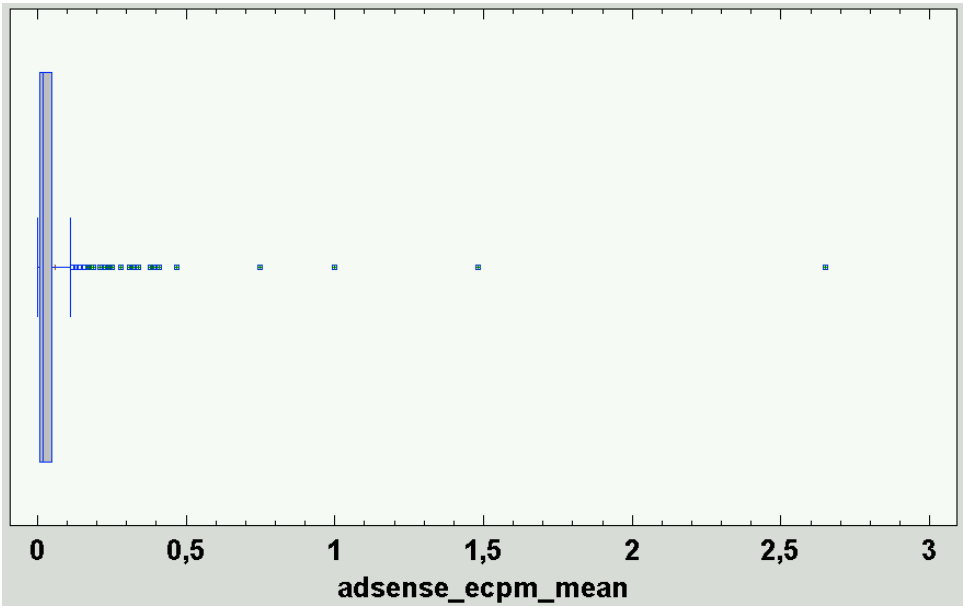


Figura 158. Todos: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean

En la Figura 158 se puede observar que existen valores anómalos de tipo extremo de 0,17 o más.

Estos valores indican que hay determinados contenidos cuyo ingreso generado por los anuncios es mayor, por lo que sería interesante estudiar qué anuncios se han mostrado cuando se han consumido dichos contenidos. Puesto que no es una información que facilite Google AdSense y escapa al control de este estudio, se propone como línea de investigación futura. En esta, se propondría averiguar la relación entre determinadas temáticas o tipos de contenido y el valor de sus anuncios, de manera que se pueda alterar

la estrategia de contenidos para maximizar los ingresos teniendo en cuenta, en esta ocasión, el propio contenido de los anuncios y su CPC.

c) Duración de la visita (promedio)

Este escenario de éxito se identifica con la columna avgtimeonpage_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 884.

Presenta un sesgo estandarizado de 16,6469 y una curtosis estandarizada de 24,4743. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

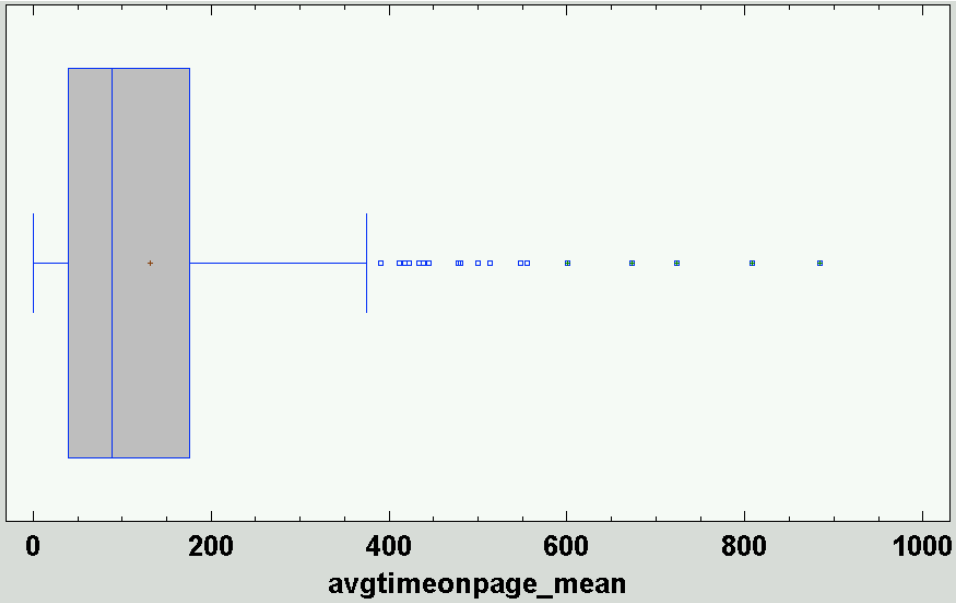


Figura 159. Todos: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean

En la Figura 159 se puede observar que existen valores anómalos de tipo extremo de 600,32 o más.

Los cinco artículos con valores anómalos presentan las siguientes características:

- Un artículo tiene 114 palabras, y no incluye ningún elemento incrustado.
- Un artículo tiene 143 palabras e incluye un vídeo incrustado.
- Un artículo tiene 117 palabras e incluye un vídeo incrustado.
- Un artículo tiene 123 palabras e incluye un vídeo incrustado.
- Un artículo tiene 323 palabras e incluye un tuit incrustado con una imagen y un texto.

- Un artículo tiene 234 palabras, y no incluye ningún elemento incrustado.
- Un artículo tiene 200 palabras, y no incluye ningún elemento incrustado.
- Un artículo tiene 146 palabras, y no incluye ningún elemento incrustado.

La mitad de los artículos anteriormente mencionados incluyen un elemento incrustado que puede haber influenciado en la duración de la visita. Esta podría haber aumentado ya que incluye también el consumo del contenido incrustado. Sería interesante analizar qué tipos de contenido influencian en esta variable y en qué medida, por lo que se propone como línea de investigación futura.

d) Páginas vistas por sesión (promedio)

Este escenario de éxito se identifica con la columna pageviewspersession_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0,39 y 9.

Presenta un sesgo estandarizado de 32,5773 y una curtosis estandarizada de 120,754 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

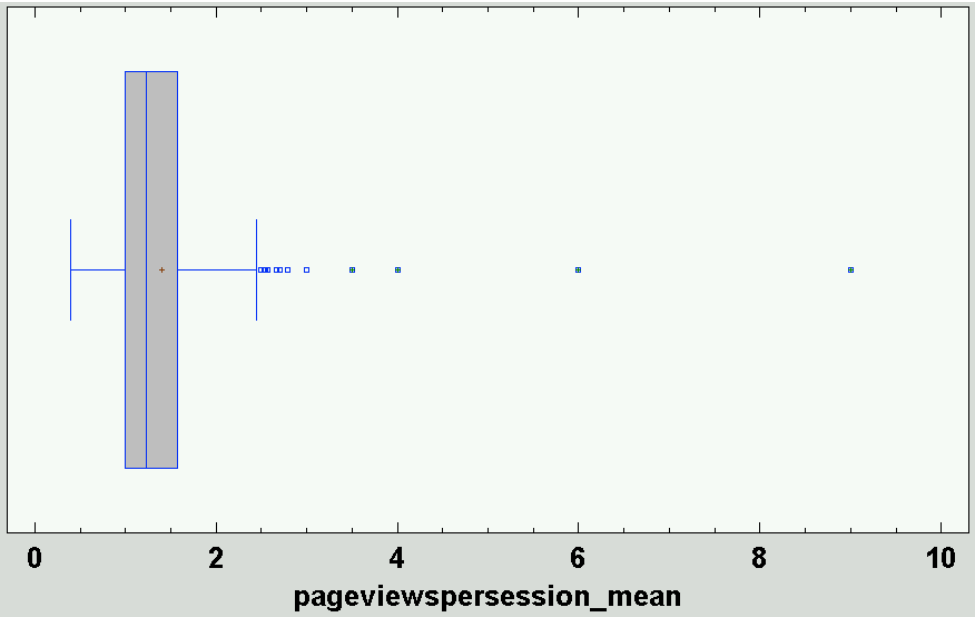


Figura 160. Todos: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean

En la Figura 160 se puede observar que existen valores anómalos de tipo extremo de 3,5 o más.

Los artículos con valores anómalos no presentan enlaces internos en su contenido, por lo que los usuarios han debido clicar en enlaces que se incluyen en el pie del artículo, la cabecera o la barra lateral, por lo que podría ser interesante realizar un estudio que analice los distintos elementos internos y externos al contenido que puedan afectar a esta variable, lo cual se propone como línea de investigación futura.

e) Nº de retuits en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 4.

Presenta un sesgo estandarizado de 3,28716 y una curtosis estandarizada de 16,0458 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

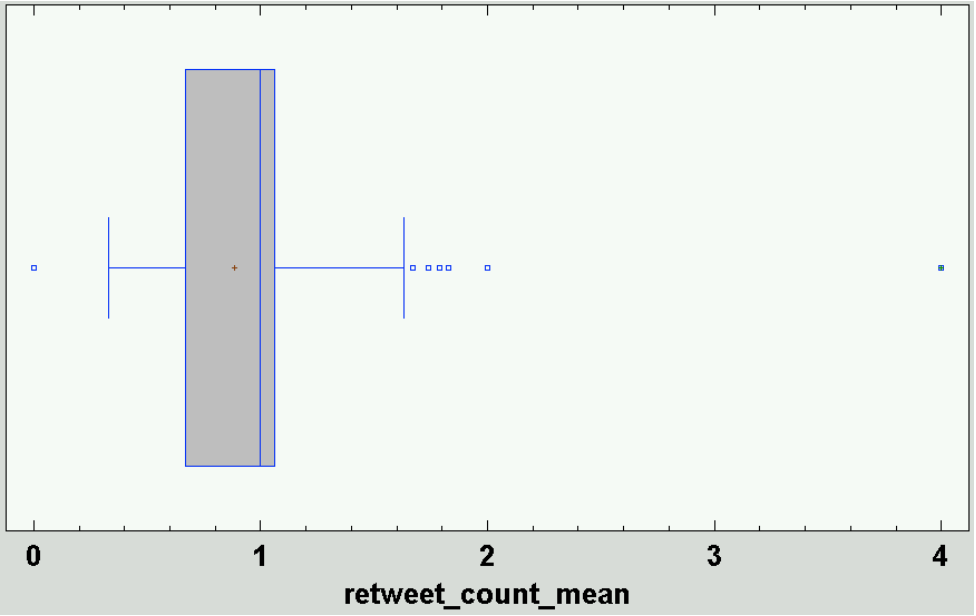


Figura 161. Todos: Gráfico de Caja y Bigotes para el valor retweet_count_mean

En la Figura 161 solo se observa un valor anómalo de tipo extremo con 4 retuits de promedio.

f) Nº de favoritos en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 11.

Presenta un sesgo estandarizado de 22,7574 y una curtosis estandarizada de 94,2744 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

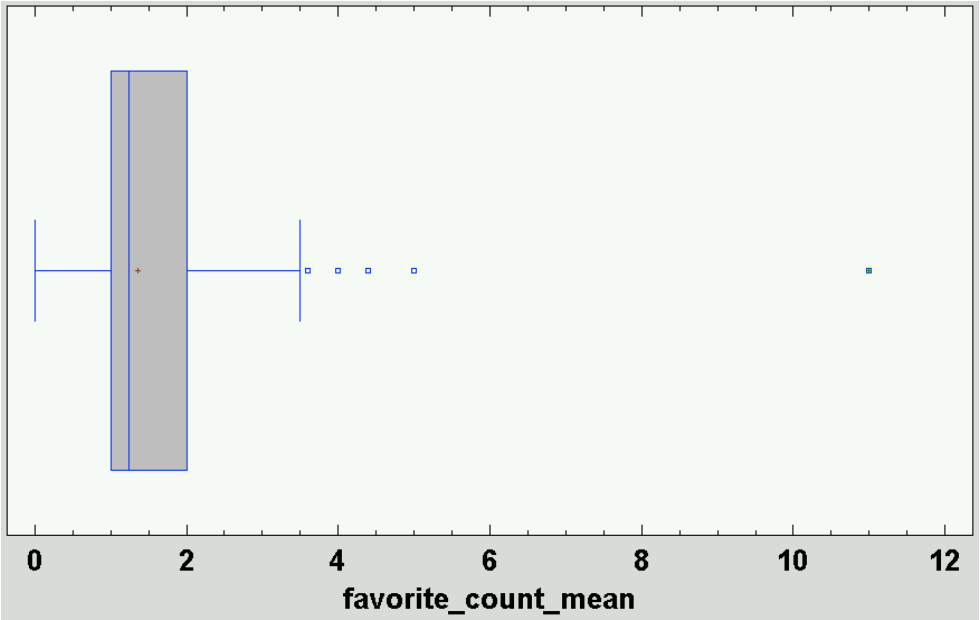


Figura 162. Todos: Gráfico de Caja y Bigotes para el valor favorite_count_mean

En la Figura 162 solo se observa un valor anómalo de tipo extremo con 11 favoritos de promedio.

g) Nº de tuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 350 valores con un rango de entre 0 y 15038.

Presenta un sesgo estandarizado de 12,6836 y una curtosis estandarizada de 14,5726 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

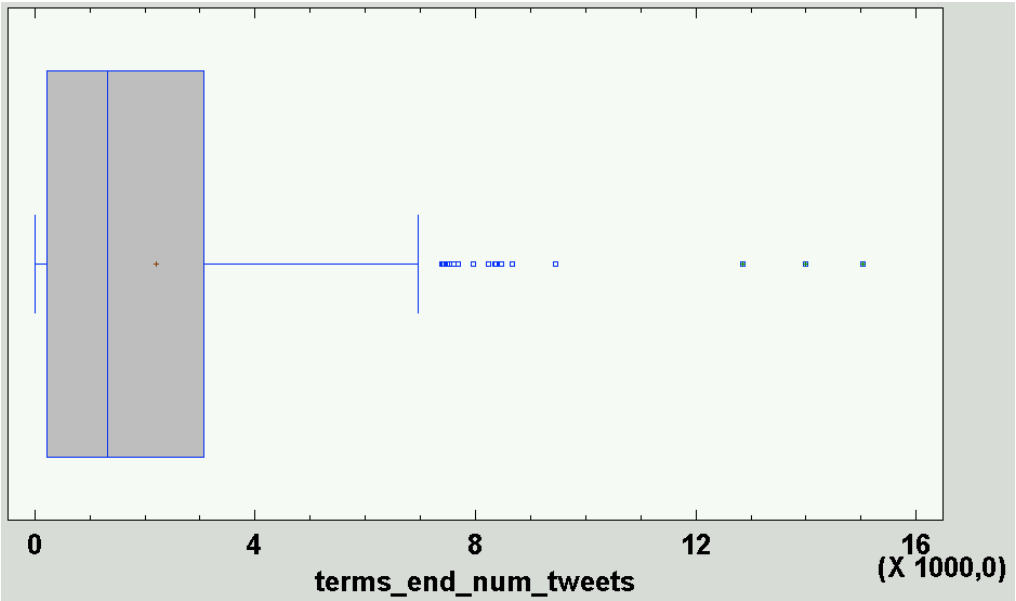


Figura 163. Todos: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets

En la Figura 163 se puede observar que existen valores anómalos de tipo extremo de 12851 o más.

Las tres tendencias con un valor anómalo presentaron valores en terms_ini_num_tweets, qué número de tuits tenía la tendencia en el momento de la publicación del artículo, parecidos a los de terms_end_num_tweets (12.399 y 13.893, 15.188 y 15.038, 15.305 y 12.851). Esto indica que las tendencias con valores anómalos en número de tuits 14 días después también lo tenían en el primer día de su análisis.

h) Nº de retuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 49.599.850.

Presenta un sesgo estandarizado de 91,0248 y una curtosis estandarizada de 647,927. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

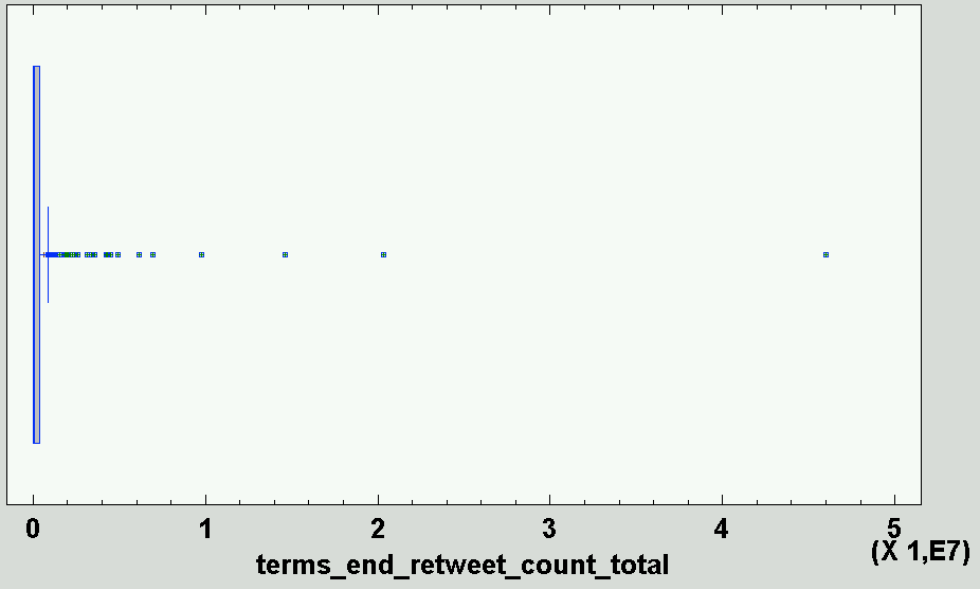


Figura 164. Todos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total

En la Figura 164 se puede observar que existen valores anómalos de tipo extremo de 1.531.660 o más.

Los datos anómalos se repiten también en el valor terms_ini_retweet_count_total, por lo que en ocasiones se debe tanto a tendencias que partían de un número inicial muy alto de tuits como a tendencias que crecieron con mucha intensidad en esos 14 días. Sería muy interesante investigar qué características tienen estas tendencias y a qué se debe ese crecimiento, por lo que se propone como una línea de investigación futura.

i) Nº de retuits de la tendencia 14 días después (promedio)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 3038,46.

Presenta un sesgo estandarizado de 58,5662 y una curtosis estandarizada de 264,538 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

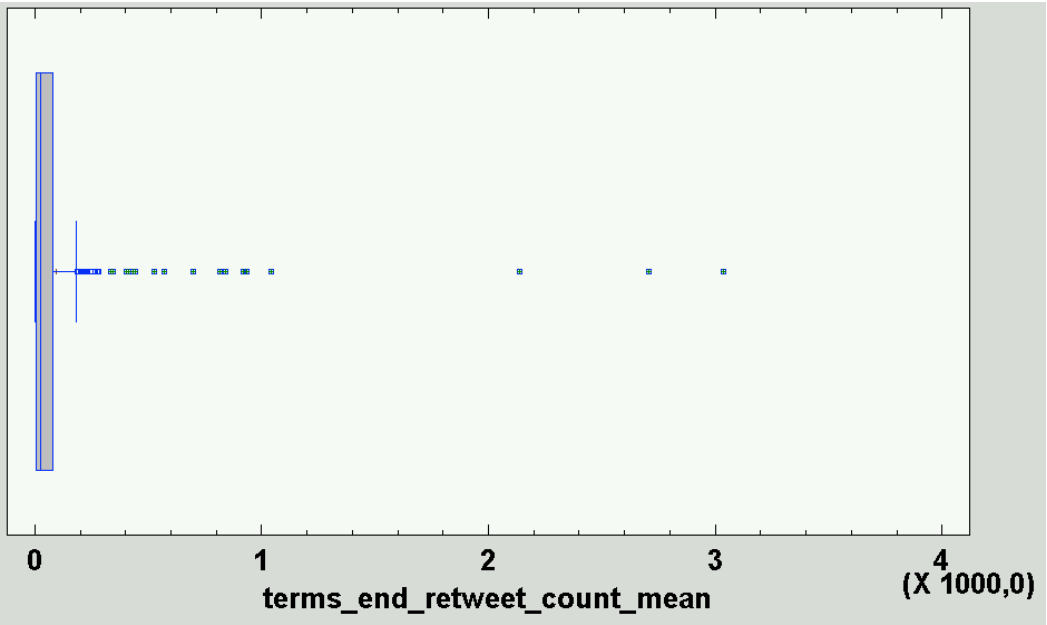


Figura 165. Todos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean

En la Figura 165 se puede observar que existen valores anómalos de tipo extremo de 335,99 o más.

La mayoría de las tendencias con valores anómalos presentaban un número mucho menor al inicio, pero en algunos casos no es así. Se puede interpretar que algunas tendencias aumentaron su nivel de amplitud 14 días después, mientras que otras perdieron fuerza quizá debido a que dejó de ser una novedad. Sería muy interesante investigar qué características tienen estas tendencias y a qué se debe ese crecimiento, por lo que se propone como una línea de investigación futura.

j) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de las variables de éxito. Se pueden observar los siguientes datos de estas:

Tabla 8

Todos: Resumen estadístico de las variables de éxito

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada

uniquepageviews_total	3.976,5	44,9048	177,779
adsense_ecpm_mean	0,0334349	77,9369	487,459
avgtimeonpage_mean	17.817,8	16,6469	24,4743
pageviewspersession_mean	0,609536	32,5773	120,754
retweet_count_mean	0,242637	3,28716	16,0458
favorite_count_mean	0,99587	22,7574	94,2744
terms_end_num_tweets	6.055.660	12,6836	14,5726
terms_end_retweet_count_total	8.646.350.000.000	91,0248	647,927
terms_end_retweet_count_mean	75.631,2	58,4662	264,538

Se puede observar en la Tabla 8 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 9

Todos: Resumen estadístico de las variables de éxito con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada

log(uniquepageviews_total)	0,89564	3,27687	3,8903
log(adsense_ecpm_mean)	1,04616	7,30132	5,41936
log(avgtimeonpage_mean)	1,00948	-2,07503	-0,552283
log(pageviewspersession_mean)	0,181997	3,86272	6,82577
log(retweet_count_mean)	0,136974	-4,36041	5,49816
log(favorite_count_mean)	0,242111	-0,274068	3,71378
log(terms_end_num_tweets)	3,95784	-8,21734	2,04657
log(terms_end_retweet_count_total)	13,0717	-5,61783	-0,594809
log(terms_end_retweet_count_mean)	3,80495	-3,55266	0,336454

Todas las variables mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. Sin embargo, mantienen valores mucho menores, próximos al rango de -2 a +2 y con una dispersión muy parecida, por lo que, debido a la naturaleza de los datos, en este estudio se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 10

Todos: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(adsense_ecpm_mean)	0,00000616002

log(pageviewspersession_mean)	0,0484551
log(retweet_count_mean)	0
log(terms_end_num_tweets)	0,00000393255
log(terms_end_retweet_count_total)	0,000189517

En la Tabla 10 se puede ver que la única variable que se acerca muchísimo al valor-p necesario (mayor o igual que 0,05) para confirmar que sigue una distribución normal es log(pageviewspersession_mean), por lo que dicha variable también es tenida en cuenta en el modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 11

Todos: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(avgtimeonpage_mean)
terms_ini_retweet_count_mean	log(pageviewspersession_mean)
terms_ini_favorite_count_total	log(favorite_count_mean)
terms_ini_favorite_count_mean	log(terms_end_retweet_count_mean)
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	

La lista de variables de éxito queda, por tanto, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

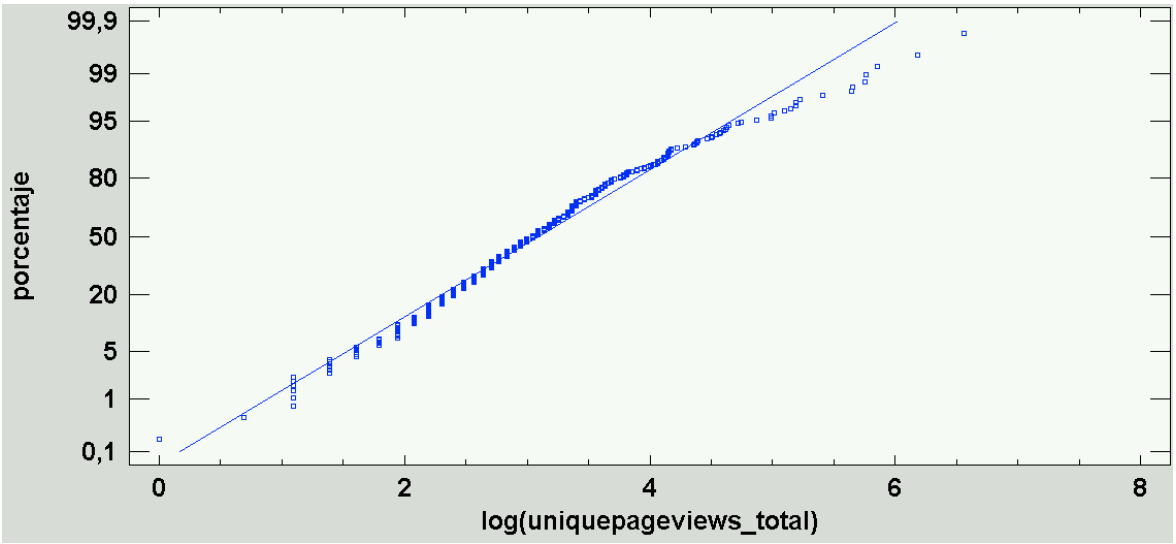


Figura 166. Todos: Gráfico de probabilidad normal de la variable log(uniquepageviews_total)

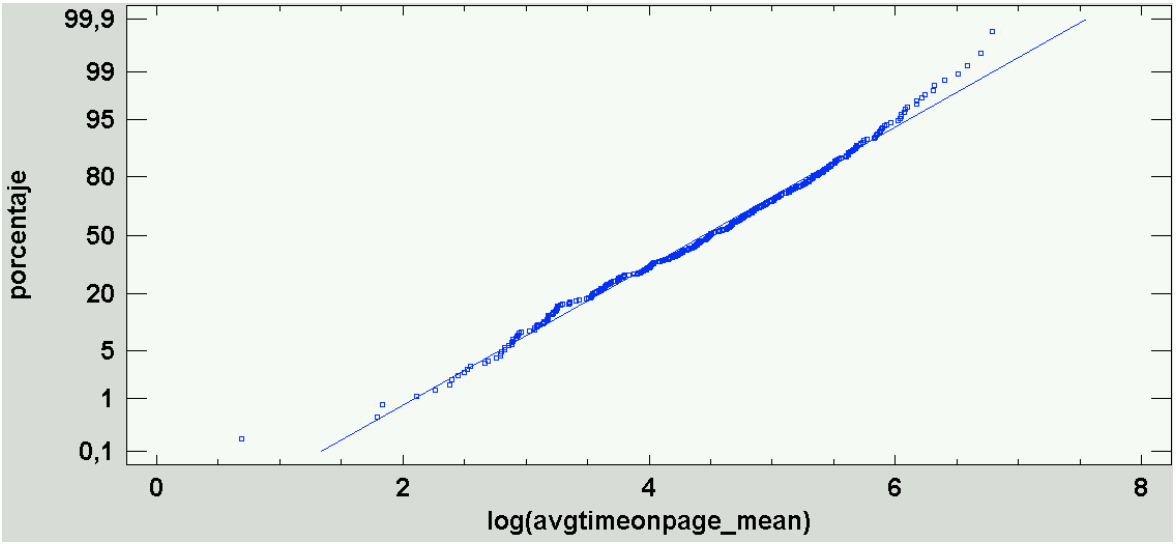


Figura 167. Todos: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean)

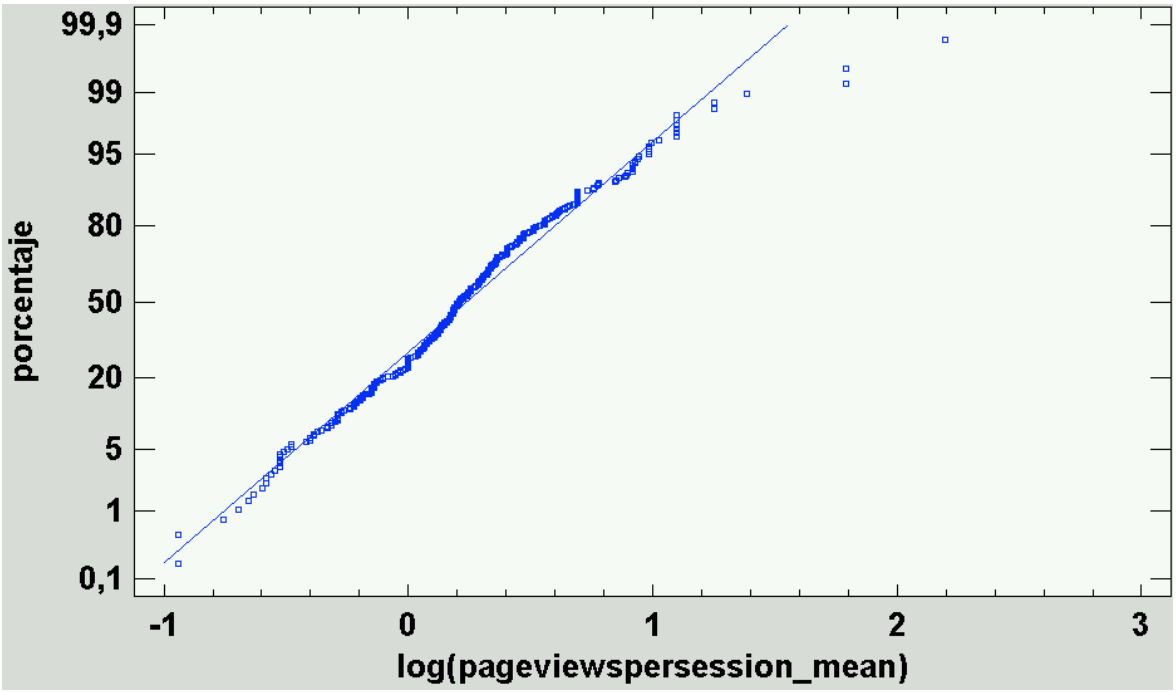


Figura 168. Todos: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean)

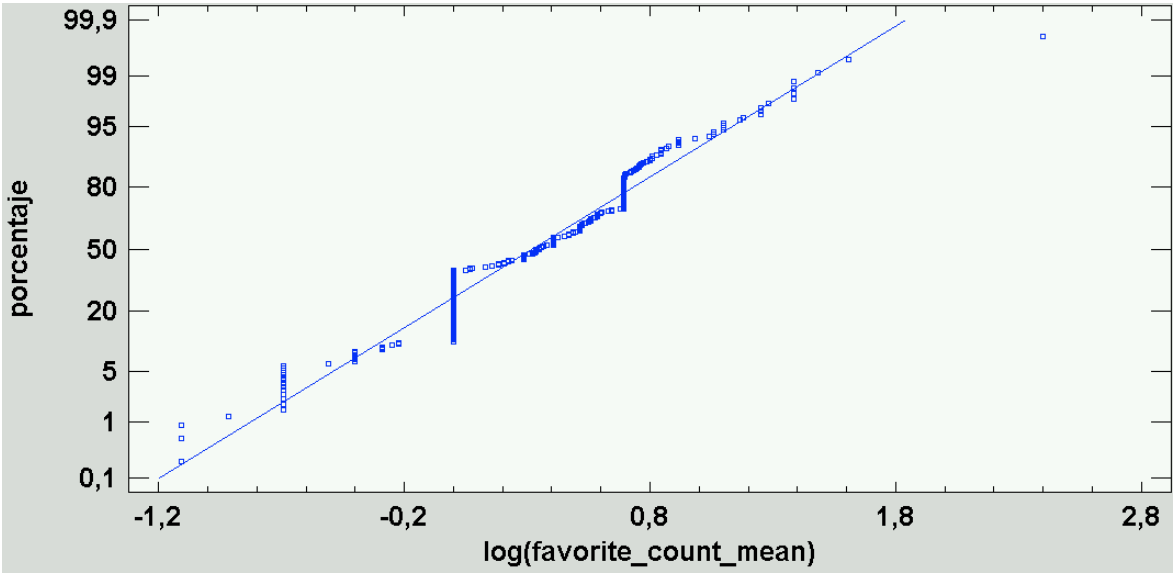


Figura 169. Todos: Gráfico de probabilidad normal de la variable log(favorite_count_mean)

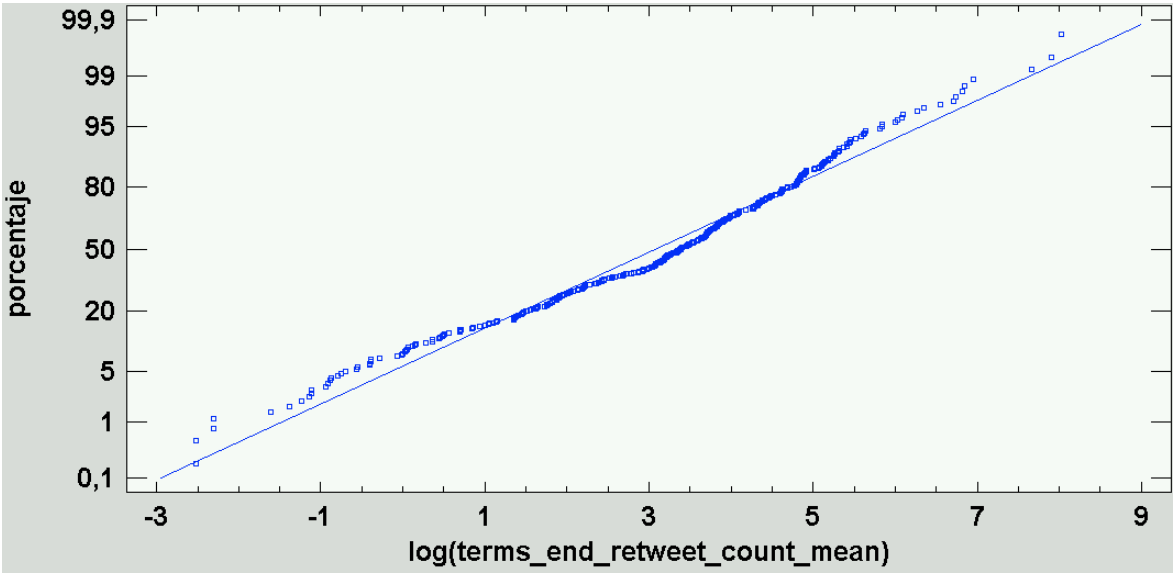


Figura 170. Todos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

k) Filtro de alta correlación

Puesto que se han propuesto más de una variable de éxito, es conveniente evaluar si están asociadas mediante una fuerte correlación. Si fuera ese el caso, el análisis de este estudio se simplificaría ya que las conclusiones de una variable servirían para las que estén fuertemente correlacionadas con ella, lo que permitiría dedicar menos recursos a la predicción y toma de decisiones.

Para ello, es necesario realizar un análisis multivariado de las variables de éxito con su correspondiente transformación logarítmica, cuya matriz de correlaciones Pearson se puede observar en la Figura 171:

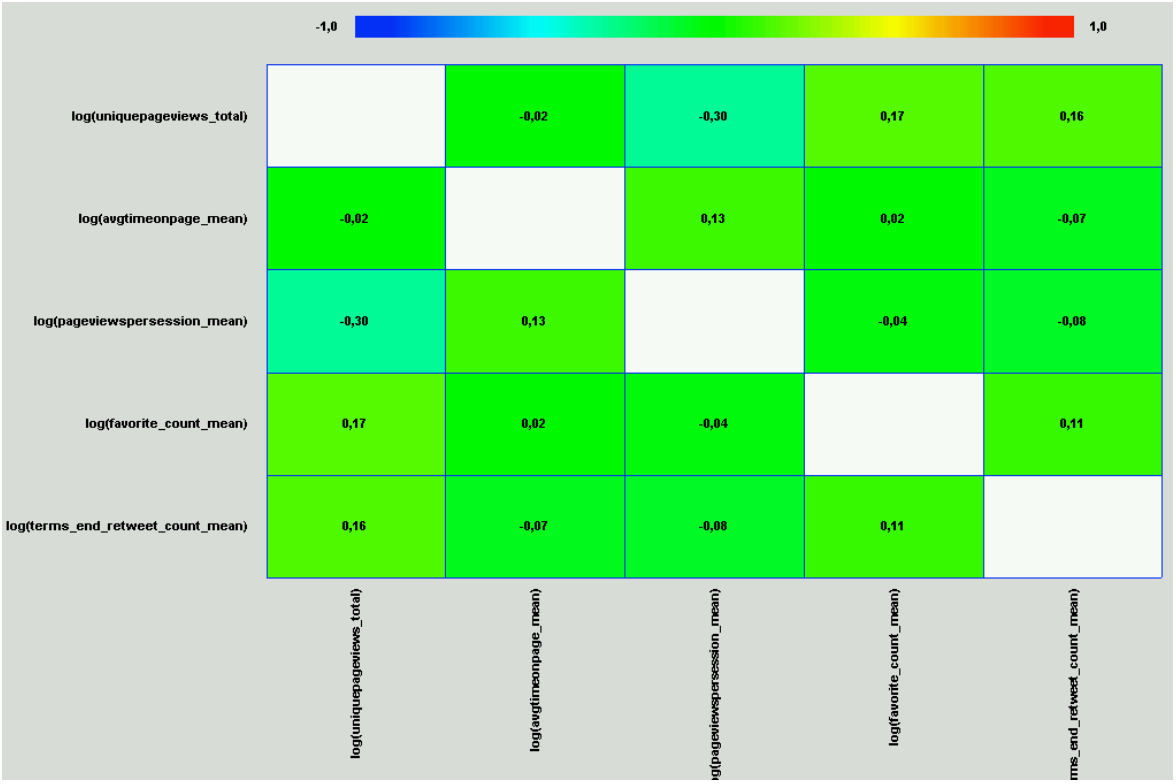


Figura 171. Todos: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente

La Figura 171 indica que no hay ninguna correlación fuerte (igual o mayor a 0,7) entre las variables de éxito, por lo que resulta interesante analizar todas por separado.

La tabla de variables quedaría como sigue:

Tabla 12

Todos: Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(avgtimeonpage_mean)
terms_ini_retweet_count_mean	log(pageviewspersession_mean)
terms_ini_favorite_count_total	log(favorite_count_mean)
terms_ini_favorite_count_mean	log(terms_end_retweet_count_mean)
terms_ini_followers_talking_rate	

terms_ini_user_num_followers_mean

terms_ini_user_num_tweets_mean

terms_ini_user_age_mean

terms_ini_url_inclusion_rate

La lista de variables de éxito queda, finalmente, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

4.2.1.2. Variables de predicción

Se parte de un conjunto de datos con diez características, por lo que es conveniente tratar de reducir la dimensión del conjunto, restableciendo la varianza sin modificar la información relevante de los datos en sí. Esto posibilitará que se reduzca el tiempo y el coste de la computación y facilita la visualización y el análisis de los datos. Además, es una condición necesaria para aplicar la regresión lineal múltiple (Anon., 2017).

a) Número de tuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 350 valores con un rango de entre 0 y 15.305.

Presenta un sesgo estandarizado de 11,3552 y una curtosis estandarizada de 10,2359. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

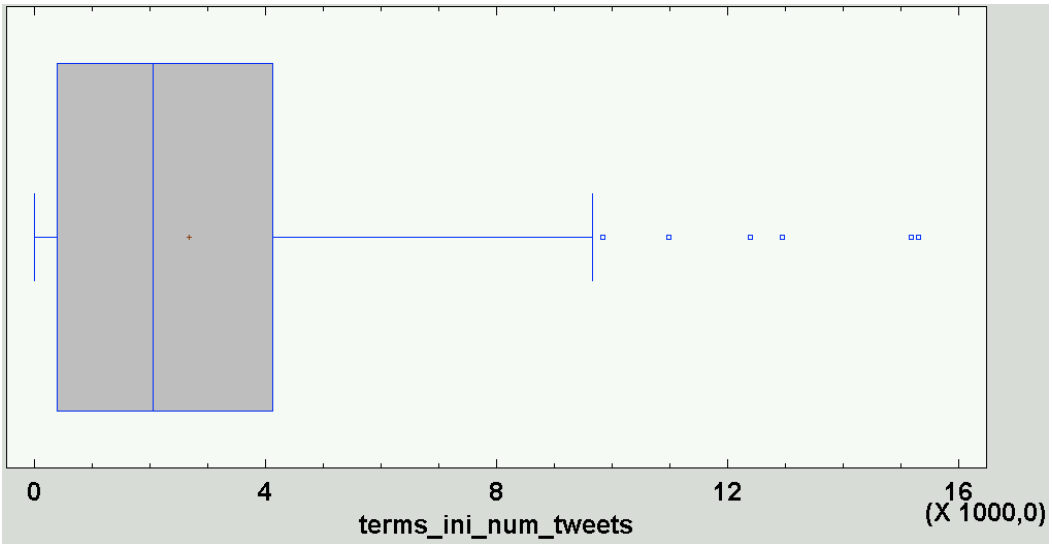


Figura 172. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets

En la Figura 172 se puede observar que no existen valores anómalos de tipo extremo.

b) Número de retuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 350 valores con un rango de entre 0 y 23.578.400.

Presenta un sesgo estandarizado de 48,1547 y una curtosis estandarizada de 189,277. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

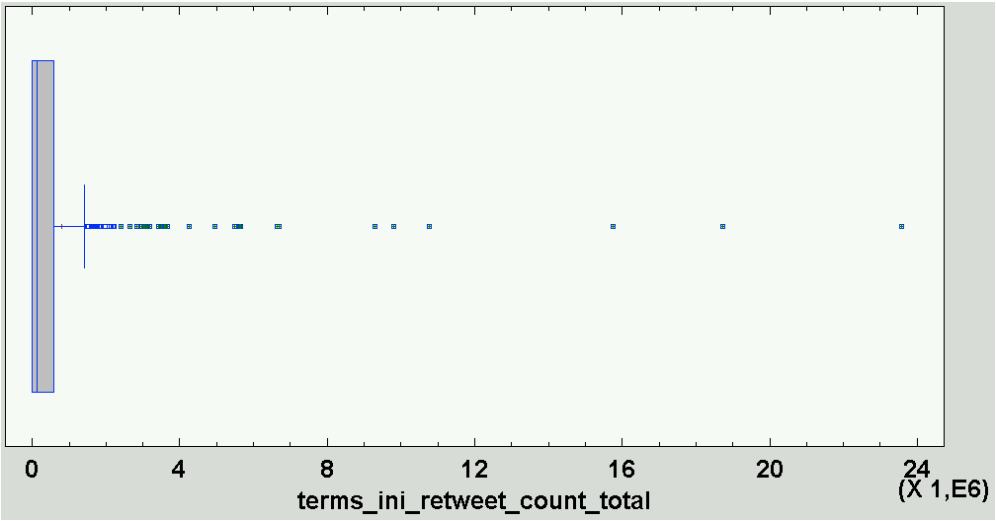


Figura 173. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total

En la Figura 173 se puede observar que existen valores anómalos de tipo extremo de 2.402.670 o más.

c) Número de retuits de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 6.260,75.

Presenta un sesgo estandarizado de 82,5457 y una curtosis estandarizada de 583,577. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

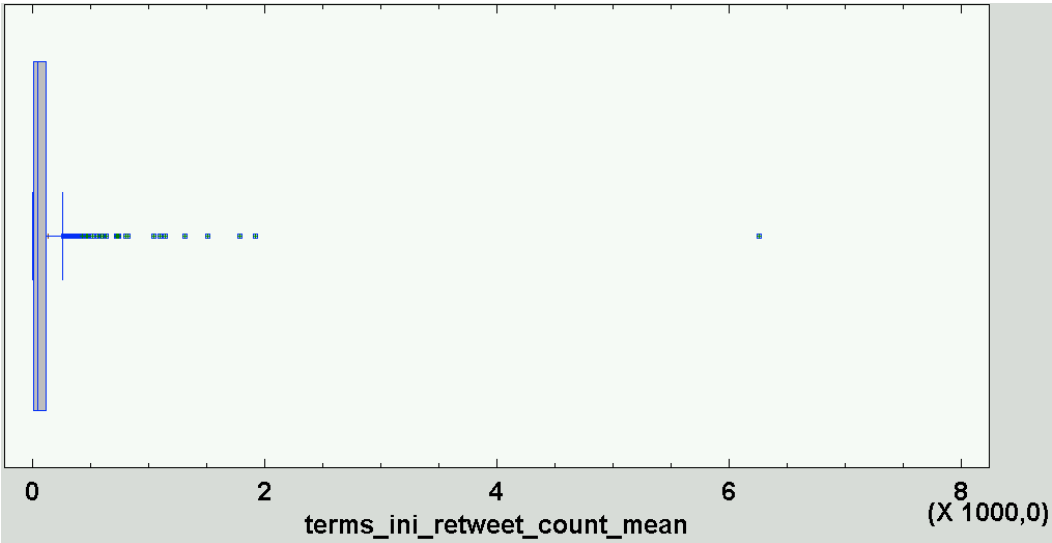


Figura 174. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean

En la Figura 174 se puede observar que existen valores anómalos de tipo extremo de 428,2 o más.

d) Número de favoritos de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 350 valores con un rango de entre 0 y 56.430.

Presenta un sesgo estandarizado de 14,0781 y una curtosis estandarizada de 14,2477. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

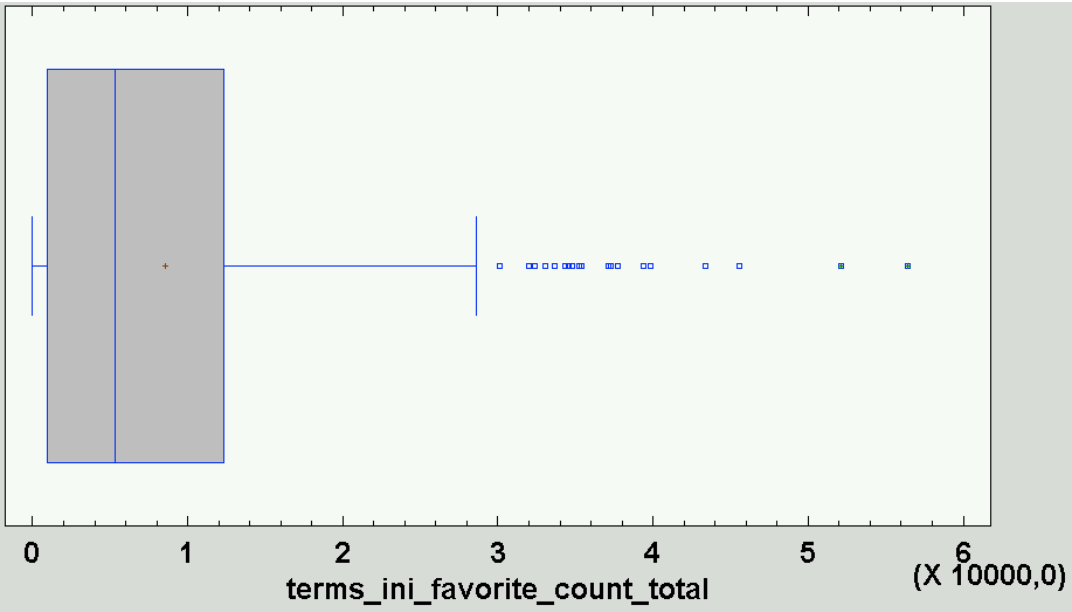


Figura 175. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total

En la Figura 175 se puede observar que existen valores anómalos de tipo extremo de 52.085 o más.

e) Número de favoritos de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 12,88.

Presenta un sesgo estandarizado de 11,8249 y una curtosis estandarizada de 16,937. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

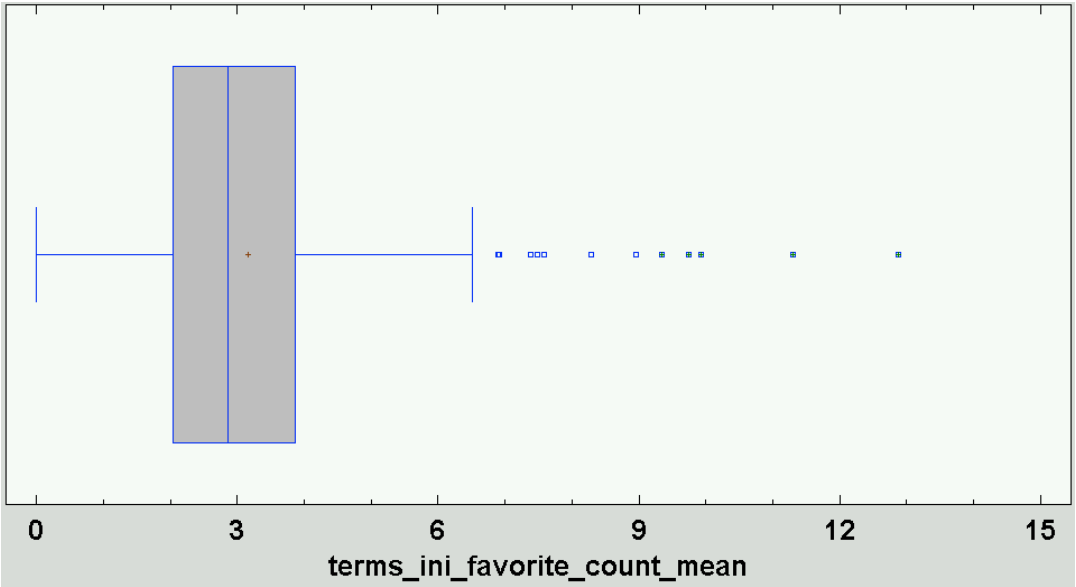


Figura 176. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean

En la Figura 176 se puede observar que existen valores anómalos de tipo extremo de 9,35 o más.

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_followers_talking_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 0,5.

Presenta un sesgo estandarizado de 40,1993 y una curtosis estandarizada de 155,251. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

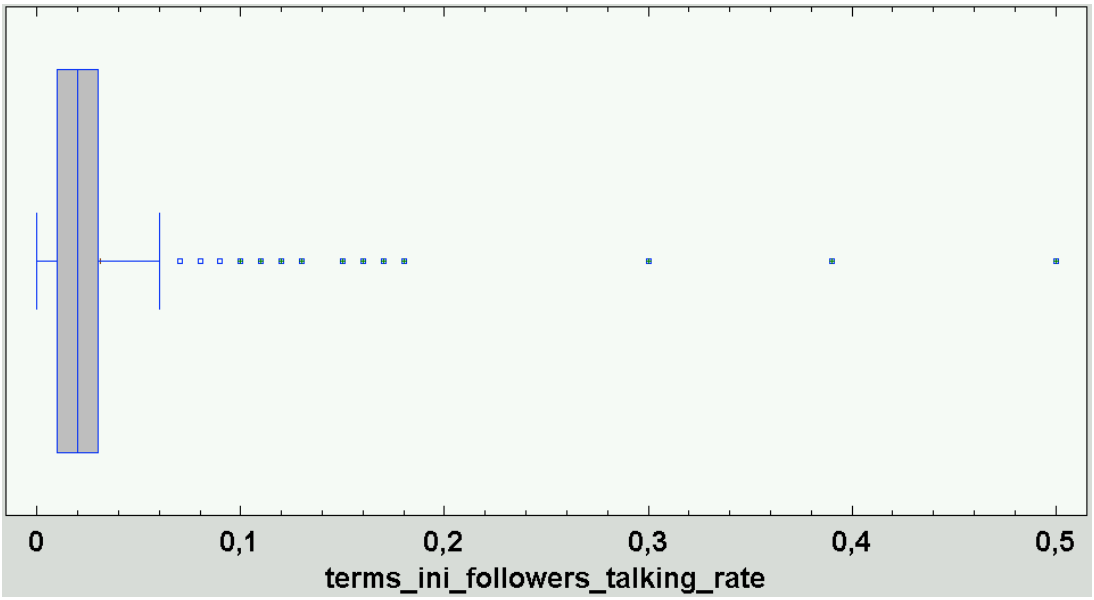


Figura 177. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate

En la Figura 177 se puede observar que existen valores anómalos de tipo extremo de 0,1 o más.

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_followers_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 373.831.

Presenta un sesgo estandarizado de 44,0968 y una curtosis estandarizada de 172,677. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

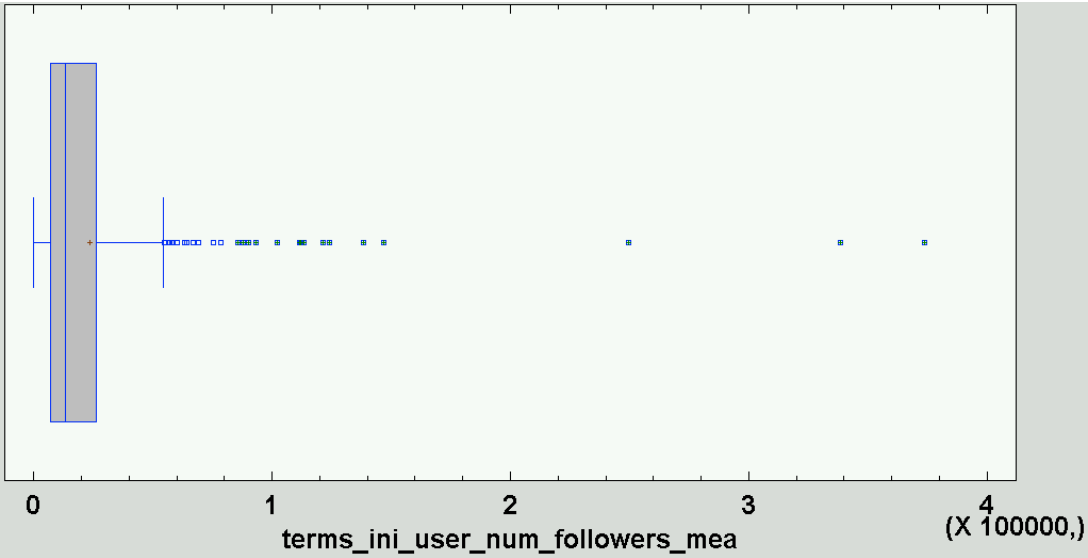


Figura 178. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean

En la Figura 178 se puede observar que existen valores anómalos de tipo extremo de 85721,9 o más.

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_tweets_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 457.769.

Presenta un sesgo estandarizado de 46,4443 y una curtosis estandarizada de 212,782. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

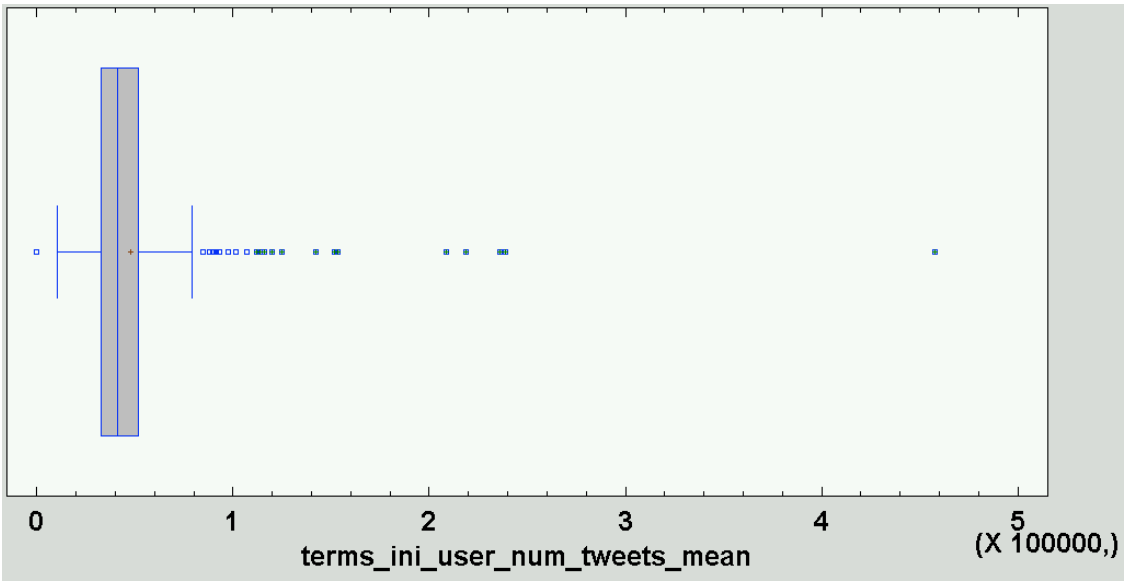


Figura 179. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean

En la Figura 179 se puede observar que existen valores anómalos de tipo extremo de 112.112 o más.

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_age_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 9.264.

Presenta un sesgo estandarizado de 29,9068 y una curtosis estandarizada de 183,824. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

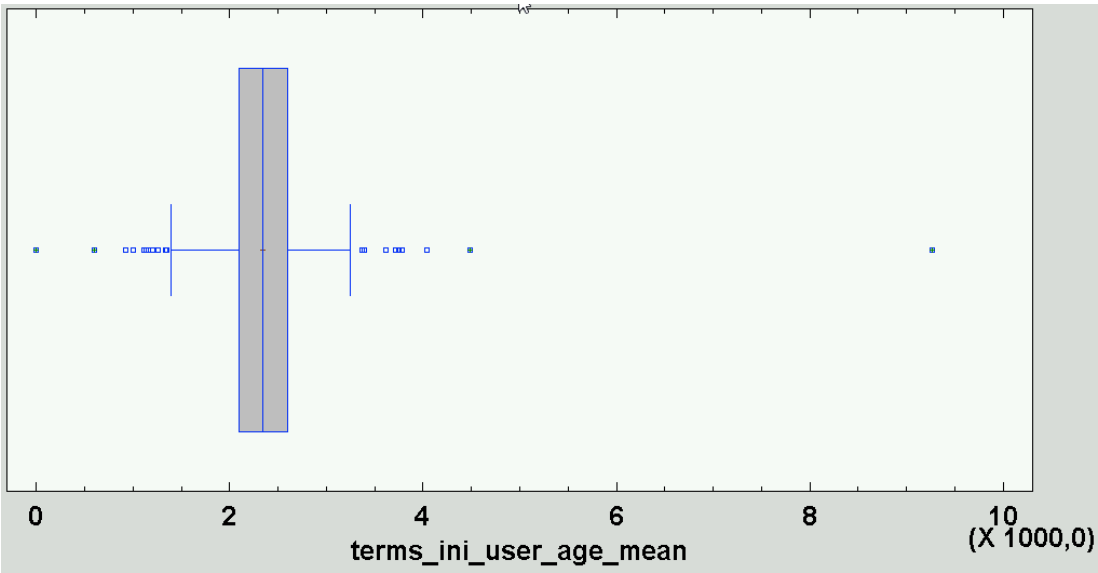


Figura 180. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean

En la Figura 180 se puede observar que existen valores anómalos de tipo extremo de 600,58 o menos, y de 4.487,3 o más.

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial

Esta variable de predicción se identifica con la columna terms_ini_url_inclusion_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 350 valores con un rango de entre 0 y 1.

Presenta un sesgo estandarizado de 5,3621 y una curtosis estandarizada de 3,42831. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

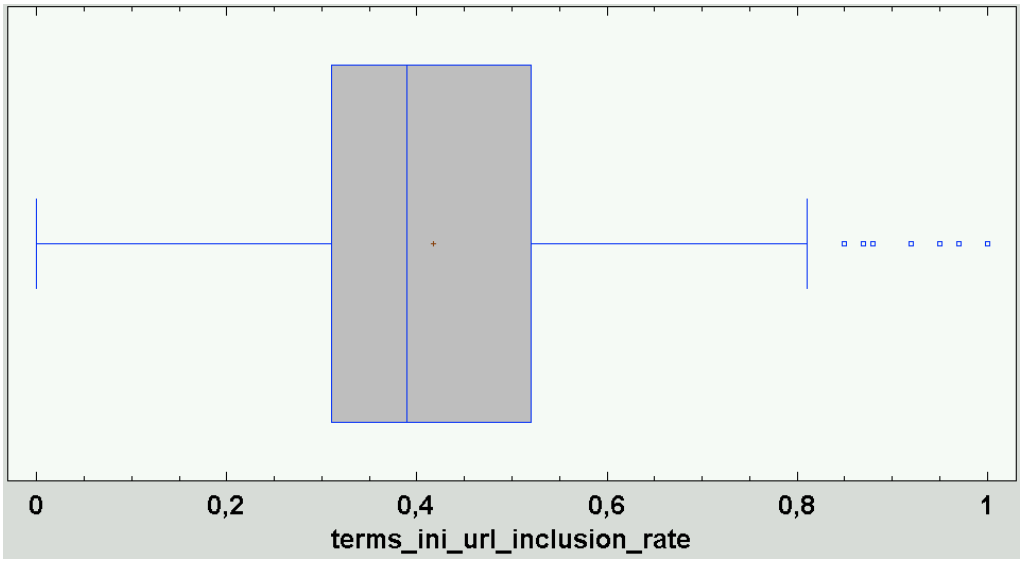


Figura 181. Todos: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate

En la Figura 181 se puede observar que no existen valores anómalos de tipo extremo.

k) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de todas las variables de predicción y éxito. De esta manera se obtienen todos los datos de las variables del modelo en un mismo análisis. Se pueden observar los siguientes datos de estas:

Tabla 13

Todos: Resumen estadístico de las variables de predicción

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
terms_ini_num_tweets	7.629.880	11,3552	10,2359
terms_ini_retweet_count_total	5.001.530.000.000	48,1547	189,277
terms_ini_retweet_count_mean	163.757	82,5457	583,577

terms_ini_favorite_count_total	99.136.600	14,0781	14,2477
terms_ini_favorite_count_mean	3,14681	11,8249	16,937
terms_ini_followers_talking_rate	0,00220133	40,1993	155,251
terms_ini_user_num_followers_mean	1.311.500.000	44,0968	172,677
terms_ini_user_num_tweets_mean	1.276.700.000	46,4443	212,782
terms_ini_user_age_mean	373.209	29,9068	183,824
terms_ini_url_inclusion_rate	0,0297356	5,3621	3,42831

Se puede observar en la Tabla 13 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple, es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 14

Todos: Resumen estadístico de las variables de predicción con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(terms_ini_num_tweets)	3,01397	-8,0618	2,27116

log(terms_ini_retweet_count_total)	11,1083	-5,73547	-0,187504
log(terms_ini_retweet_count_mean)	3,43933	-2,35353	-0,52743
log(terms_ini_favorite_count_total)	3,57898	-9,18412	5,09551
log(terms_ini_favorite_count_mean)	0,34247	-7,38854	12,2243
log(terms_ini_followers_talking_rate)	0,706743	5,63064	0,398568
log(terms_ini_user_num_followers_mean)	1,05598	0,0316429	2,18601
log(terms_ini_user_num_tweets_mean)	0,214282	6,22322	14,3022
log(terms_ini_user_age_mean)	0,0513791	-4,71801	33,3606
log(terms_ini_url_inclusion_rate)	0,186727	-4,67502	3,48789

Todas las variables mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. En este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

También se puede apreciar que los valores de varianza son bastante parecidos salvo en el caso de log(terms_ini_retweet_count_total), por lo que se cumple la condición de homocedasticidad menos en esa variable. log(terms_ini_retweet_count_total) se elimina del modelo actual para que dicho requisito se cumpla.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 15

Todos: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(terms_ini_num_tweets)	0,000000104216
log(terms_ini_favorite_count_total)	0,0000304315
log(terms_ini_favorite_count_mean)	0,0568527
log(terms_ini_followers_talking_rate)	0
log(terms_ini_user_num_tweets_mean)	0,00480817
log(terms_ini_user_age_mean)	0,000139889

En la Tabla 15 se puede ver que la única variable que supera el valor-p necesario (mayor o igual que 0,05) para confirmar que sigue una distribución normal es log(terms_ini_favorite_count_mean), por lo que dicha variable también es tenida en cuenta en el modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 16

Todos: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_url_inclusion_rate)	log(favorite_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todas ellas provenientes de la tendencia el día de la publicación del artículo.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

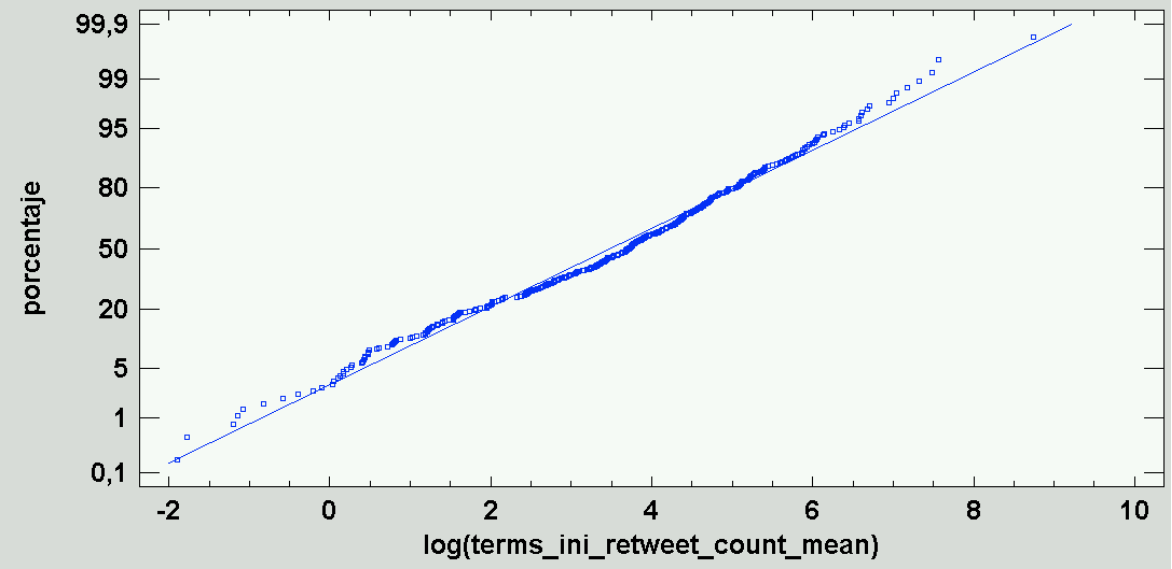


Figura 182. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean)

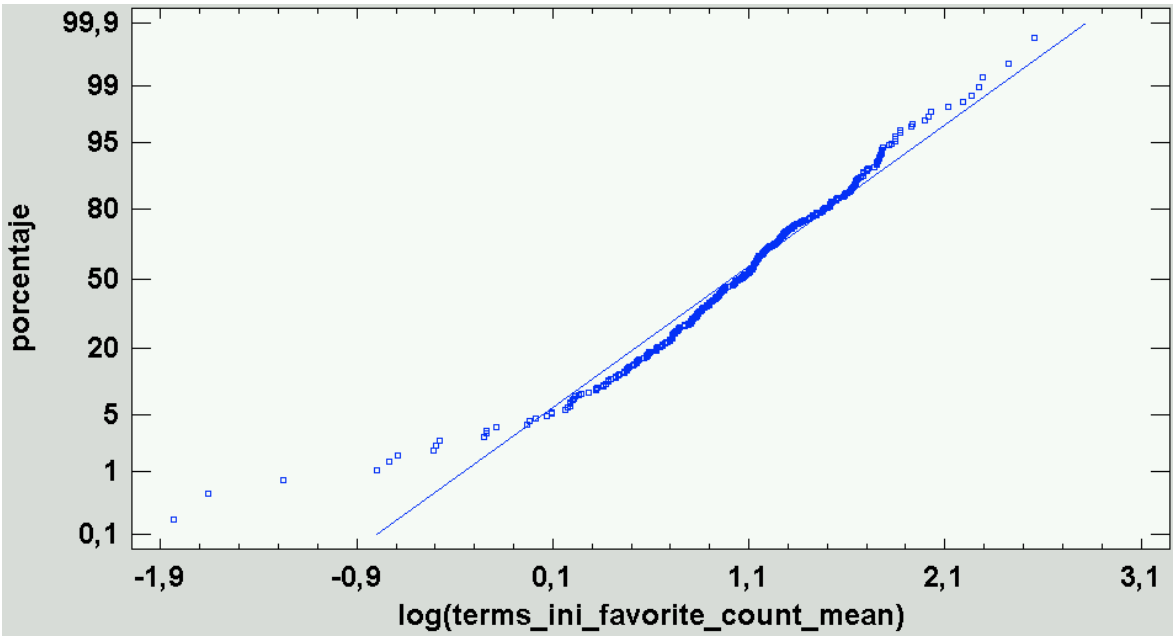


Figura 183. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean)

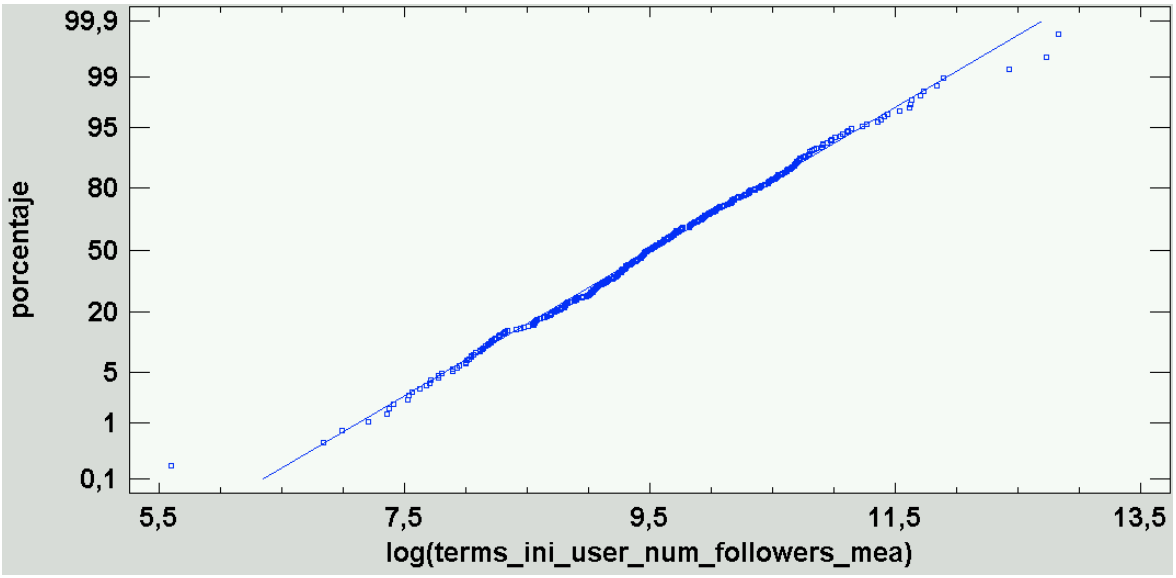


Figura 184. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean)

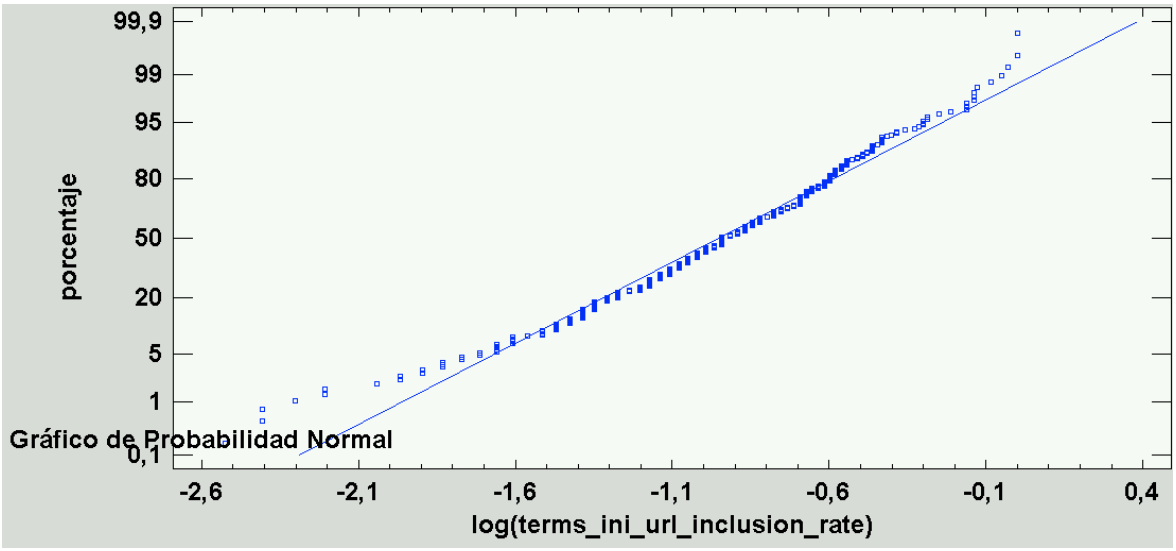


Figura 185. Todos: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

l) Filtro de alta correlación (colinealidad)

Antes de proceder, es conveniente analizar la correlación que existe entre las variables con las que contamos para el modelo. Puesto que estas han sido normalizadas en el apartado anterior, dicho análisis de correlación Pearson se realizará con su transformación logarítmica.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

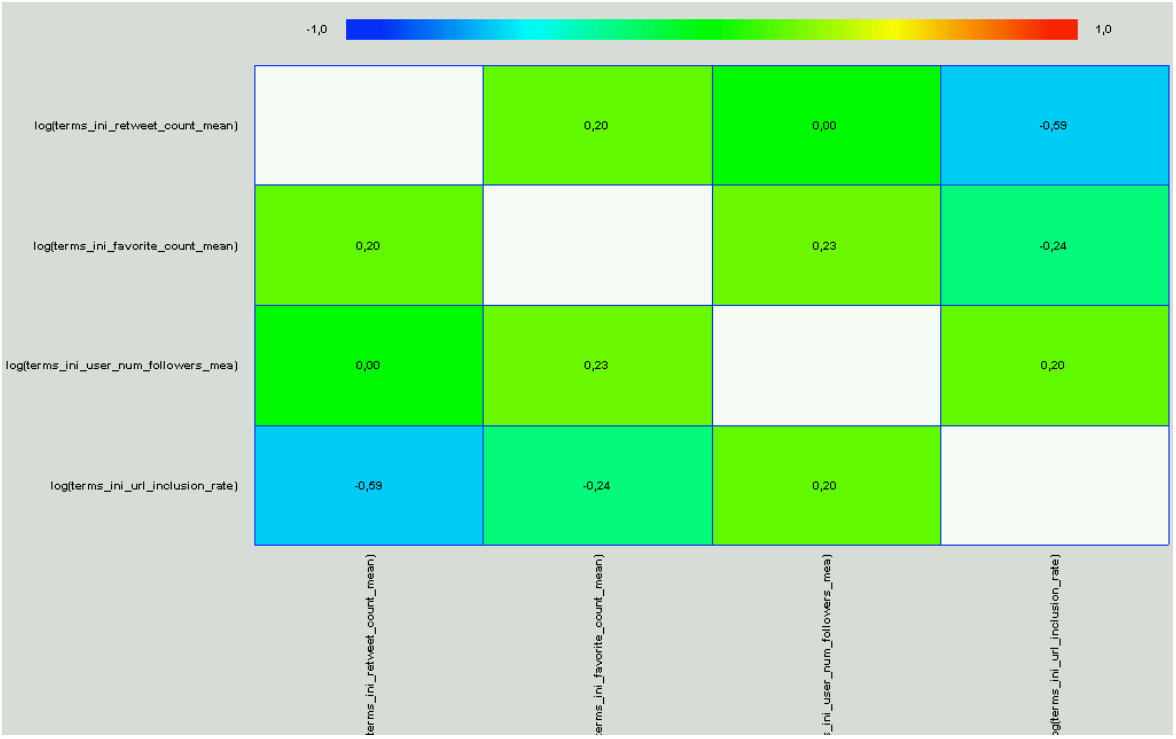


Figura 186. Todos: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente

En la Figura 186 no se aprecia ninguna correlación fuerte (0,7 o más) entre las variables de predicción, por lo que la tabla de variables quedaría como sigue:

Tabla 17

Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_url_inclusion_rate)	log(favorite_count_mean)
	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, una vez más, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, el promedio de seguidores

de los usuarios que participan y la ratio de inclusión de URL en los tuits, todas ellas provenientes de la tendencia el día de la publicación del artículo.

m) Análisis de componentes principales (ACP)

A continuación, se aplica el análisis de componentes principales (ACP, o PCA en inglés), una técnica que sirve par describir un conjunto de datos según nuevas variables no correlacionadas llamadas componentes (Dunteman, 1989).

El objetivo es representar los datos de la mejor manera posible a través de mínimos cuadrados, construyendo una transformación lineal según un nuevo sistema de coordenadas para los datos originales. Es decir, se plantea representar la variabilidad de los datos con el menor número de componentes o fórmulas posible, las cuales son combinaciones lineales de las variables originales (Dunteman, 1989).

Si las variables originales están muy correlacionadas entre sí, la mayor parte de la variabilidad se podrá expresar en pocas componentes. Si están totalmente incorrelacionadas, el número de componentes será igual al de las variables y este análisis carecerá de interés.

Las componentes se ordenan según la varianza original siendo la primer componente la que tenga la varianza de mayor tamaño. Cuanto mayor sea su varianza, mayor será la información que aporta esa componente (Amat Rodrigo, 2017).

Al construir la matriz de coeficientes de correlación, es posible una base de vectores propios, cuya transformación lineal es necesaria para mejora la simplicidad e interpretación que permita tratar de reducir la dimensionalidad de los datos (Dunteman, 1989). Esta reducción se efectuaría seleccionando las componentes principales que más aportan a la varianza e ignorando el resto. Esta selección se produce ordenando las componentes de mayor a menor aportación a la explicación de la variabilidad, y seleccionando tantas como sean necesarias hasta alcanzar un valor propio mayor o igual a 1.

De esta manera, el método ACP condensa la información de múltiples características en unas pocas, ya que se pretende explicar aproximadamente la información con menos valores que los originales.

Realizando el análisis de componentes principales se ha obtenido un total de dos componentes, explicando así el 73,602% de los datos con un valor propio de 1,72175. Las dos componentes tienen la Tabla 18 de pesos, siendo cada peso un valor de entre -1 y 1.

Tabla 18

Todos: Tabla de pesos de las componentes

	Componente	*Componente*
	1	2
log(terms_ini_retweet_count_mean)	0,643898	-0,042444
log(terms_ini_favorite_count_mean)	0,374089	0,575032
log(terms_ini_user_num_followers_mea)	-0,0640924	0,79053
log(terms_ini_url_inclusion_rate)	-0,664338	0,206396

De esta manera, por ejemplo, la primer componente principal tiene la fórmula siguiente, en donde los valores de las variables se han estandarizado restándoles su promedio y dividiéndolos entre su desviación estándar:

0,643898 * log(terms_ini_retweet_count_mean) + 0,374089 * log(terms_ini_favorite_count_mean) –0,0640924 * log(terms_ini_user_num_followers_mean – 0,664338*log(terms_ini_url_inclusion_rate)

Se puede representar la mayor parte de esa variabilidad con solo dos compontes principales, de los cuales:

- Componente 1: está en buena medida explicada por el número medio de retuits y la no inclusión de una URL. Se podría interpretar que consta de aquellas tendencias con un gran nivel de participación conversacional.
- Componente 2: está en buena medida explicada por el número medio de favoritos y usuarios con un gran número de seguidores. Se puede comprender como aquellas tendencias con recepción positiva de usuarios con muchos seguidores.

También se observa que la única variable que aporta positivamente a las dos componentes principales es el número medio de favoritos. El resto sí que aportan un valor negativo en alguna de ellas, aunque solo la inclusión de una URL en los tuits obtiene un valor negativo considerable, siendo además en la primera componente.

La relación entre las variables y las dos componentes principales se puede ver en la siguiente gráfica, ya que las variables se muestran en dos dimensiones formadas por estas componentes:

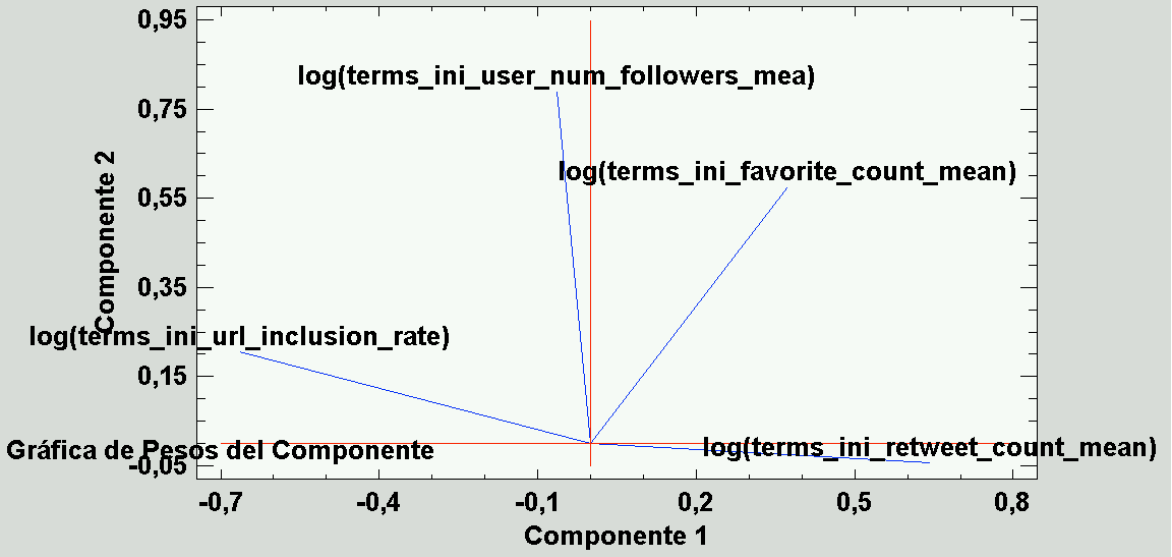


Figura 187. Todos: Gráfica de pesos de cada componente principal

En la Tabla 18 se puede comprobar que todas las variables tienen una presencia significativa en alguna de las componentes principales, por lo que no se puede eliminar ninguna de las variables originales mediante esta técnica.

4.2.1.3. Regresión lineal múltiple

Para estudiar la posible relación entre las variables independientes de predicción de que disponemos y cada variable dependiente de éxito o, dicho de otro modo, para tratar de predecir el cálculo de estas, vamos a realizar un modelo de regresión múltiple.

De esta manera, se pretende realizar la identificación de las variables explicativas para cada variable de éxito, es decir, aquellas variables que influyan en la respuesta de dicha variable, y en qué medida lo hacen mediante sus coeficientes. La descripción de su relación desvelará la fórmula del modelo ajustado de regresión lineal que permitirá la predicción de cada variable de éxito, en el caso de que el modelo sea significativo.

El modelo es una expansión de la regresión simple lineal y tiene la siguiente estructura:

$$Y_{pred} = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

En la fórmula, Y_{pred} es la variable de éxito, a es la constante, las Xs son las variables de predicción y las bs son sus correspondientes coeficientes o pesos.

Para realizar las regresiones múltiples, se cuenta con la Tabla *19* de variables resultante de todos los análisis anteriores:

Tabla 19

Todos: Lista final de variables de predicción y de éxito para la regresión lineal múltiple

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_url_inclusion_rate)	log(favorite_count_mean)
	log(terms_end_retweet_count_mean)

Las variables de predicción responden a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos aplicados a la tendencia el día de la publicación del artículo.

La lista de variables de éxito está formada por la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

a) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión múltiple con la variable dependiente log(uniquepageviews_total). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 20

Todos: Valor-P de las variables de la regresión múltiple de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	3,79977	0

log(terms_ini_retweet_count_mean)	0,0574119	0,0945
log(terms_ini_favorite_count_mean)	0,0395773	0,6725
log(terms_ini_user_num_followers_mean)	-0,103491	0,0586
log(terms_ini_url_inclusion_rate)	-0,051989	0,7382
Modelo		0,0352

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 21

Todos: Valor-P de las variables de la regresión múltiple simplificada de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	3,84642	0
log(terms_ini_retweet_count_mean)	0,0669324	0,0149
log(terms_ini_user_num_followers_mean)	-0,102428	0,0434
Modelo		0,0068

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(3,84642 + 0,0669324 * log(terms_ini_retweet_count_mean) - 0,102428 * log(terms_ini_user_num_followers_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

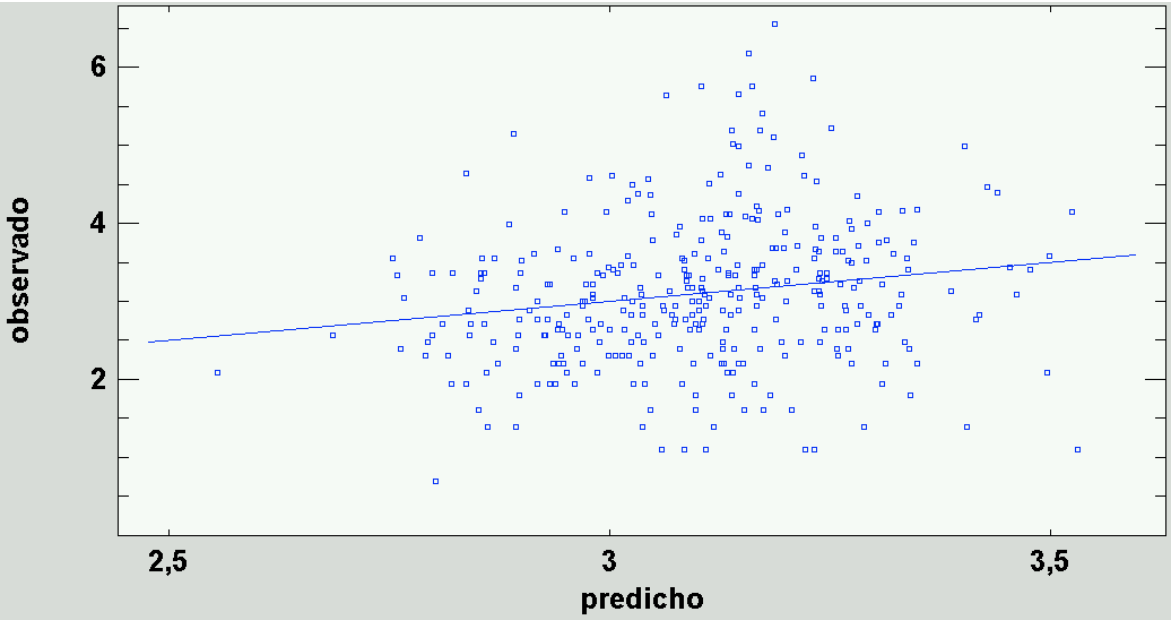


Figura 188. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 2,87745% de la variabilidad de log(uniquepageviews_total), mientras que el R-Cuadrado ajustado indica un 2,30948%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

b) Duración de la visita (promedio)

Para tratar de predecir el valor de la duración de la visita (promedio) es necesario realizar la regresión múltiple con la variable dependiente log(avgtimeonpage_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 22

Todos: Valor-P de las variables de la regresión múltiple de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	4,81494	0
log(terms_ini_retweet_count_mean)	0,00577653	0,8742

log(terms_ini_favorite_count_mean)	0,0167378	0,8658
log(terms_ini_user_num_followers_mean)	-0,0736711	0,2065
log(terms_ini_url_inclusion_rate)	-0,290134	0,0799
Modelo		0,0598

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 23

Todos: Valor-P de las variables de la regresión múltiple simplificada de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	4,15622	0
log(terms_ini_url_inclusion_rate)	-0,294819	0,0185
Modelo		0,0185

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

avgtimeonpage_mean = exp(4,15622 - 0,294819 * log(terms_ini_url_inclusion_rate))

Para realizar el cálculo de avgtimeonpage_mean será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

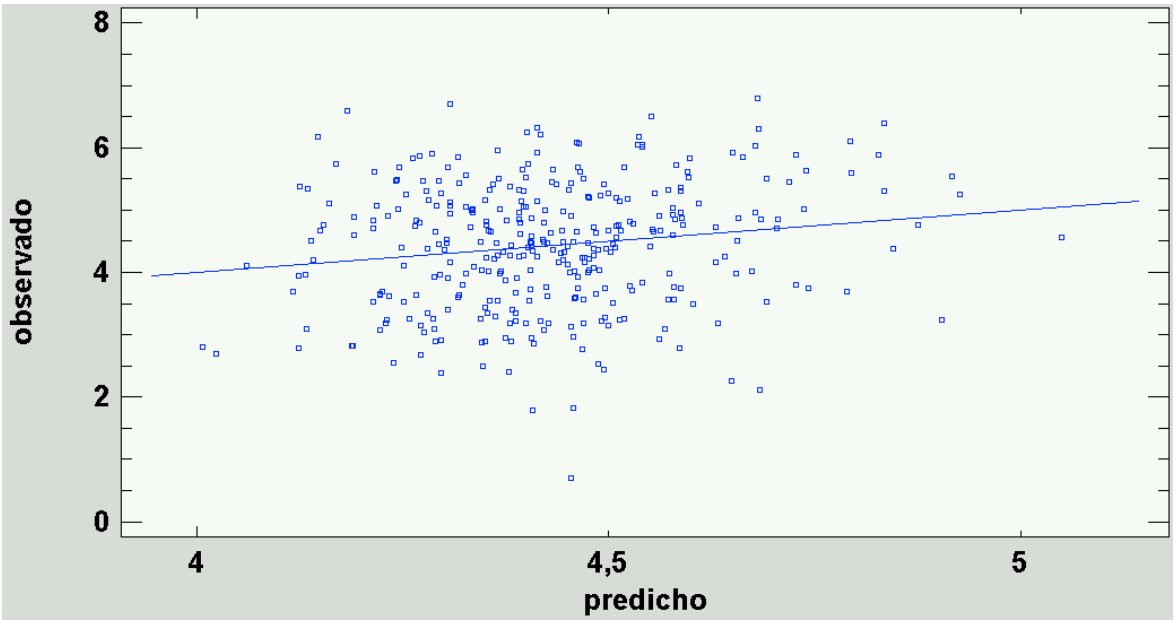


Figura 189. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 1,61182% de la variabilidad de log(avgtimeonpage_mean), mientras que el R-Cuadrado ajustado indica un 1,32414%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Páginas vistas por sesión (promedio)

Para tratar de predecir el valor de las páginas vistas por sesión (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(pageviewspersession_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 24

Todos: Valor-P de las variables de la regresión múltiple de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	0,63497	0,0094
log(terms_ini_retweet_count_mean)	-0,00440915	0,7747
log(terms_ini_favorite_count_mean)	-0,0912531	0,0307

log(terms_ini_user_num_followers_mean)	-0,0301402	0,2197
log(terms_ini_url_inclusion_rate)	0,00276459	0,9685
Modelo		0,0588

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 25

Todos: Valor-P de las variables de la regresión múltiple simplificada de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	0,34813	0
log(terms_ini_favorite_count_mean)	-0,108948	0,0052
Modelo		0,0052

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

pageviewspersession_mean = exp(0,34813 - 0,108948 * log(terms_ini_favorite_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

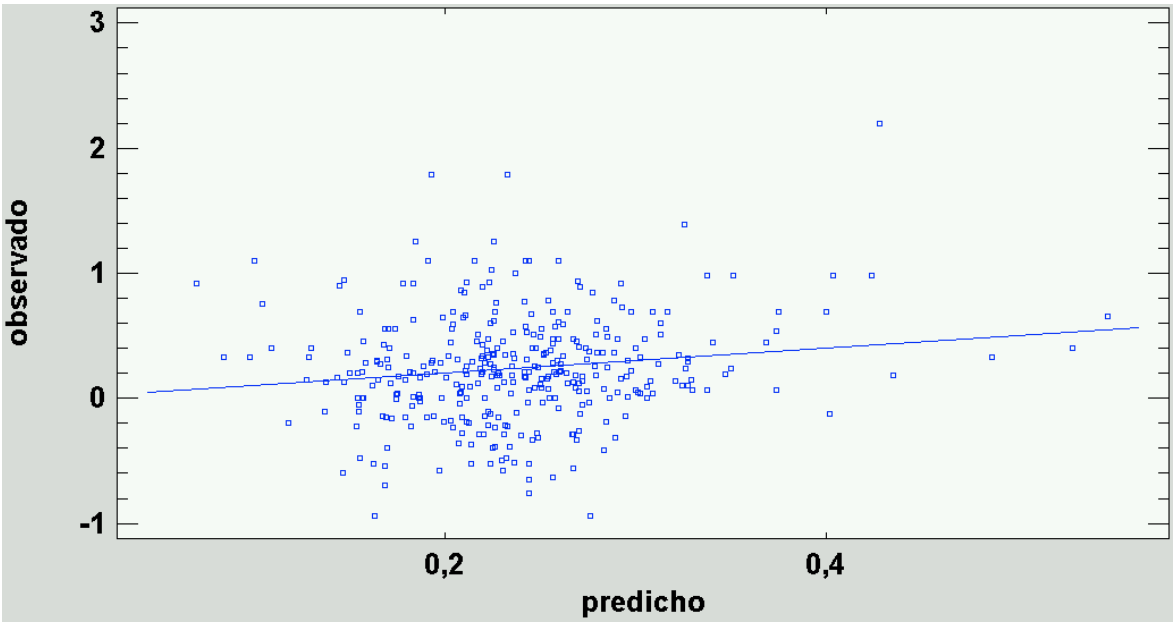


Figura 190. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 2,23659% de la variabilidad de log(pageviewspersession_mean), mientras que el R-Cuadrado ajustado indica un 1,95322%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

d) Nº de favoritos en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de favoritos en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(favorite_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 26

Todos: Valor-P de las variables de la regresión múltiple de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	0,450465	0,1374
log(terms_ini_retweet_count_mean)	0,0194036	0,3189
log(terms_ini_favorite_count_mean)	-0,00741444	0,8884

log(terms_ini_user_num_followers_mean)	-0,0296624	0,3301
log(terms_ini_url_inclusion_rate)	-0,0939052	0,2792
Modelo		0,1264

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 27

Todos: Valor-P de las variables de la regresión múltiple simplificada de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	0,177589	0,0108
log(terms_ini_url_inclusion_rate)	-0,146747	0,0258
Modelo		0,0258

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

favorite_count_mean = exp(0,177589 - 0,146747 * log(terms_ini_url_inclusion_rate))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

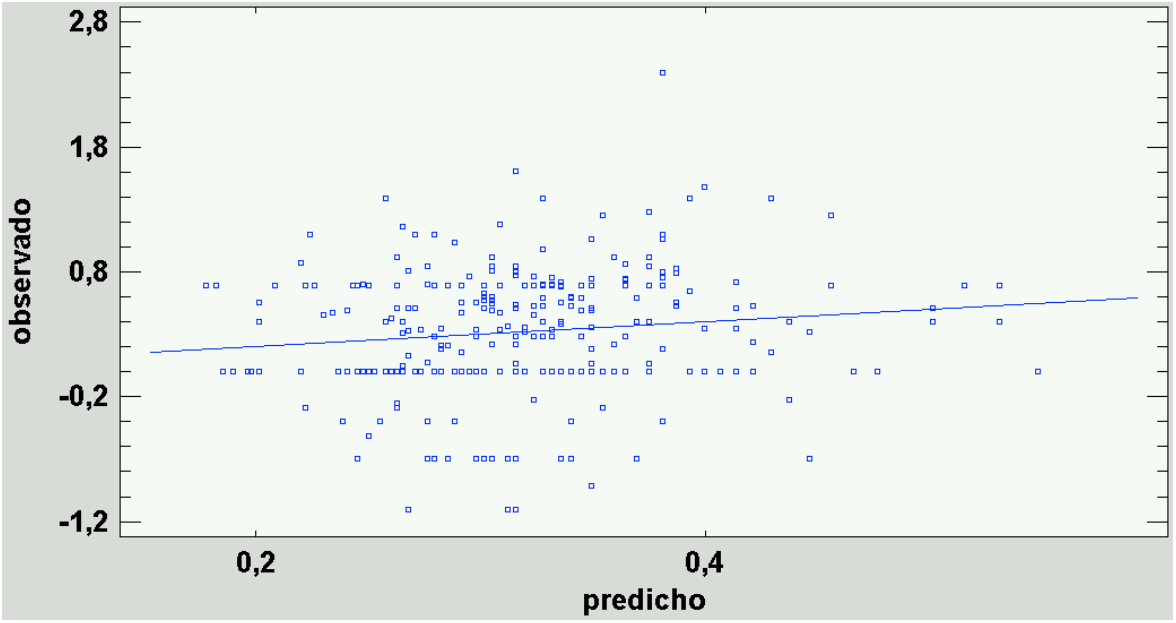


Figura 191. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 1,63475% de la variabilidad de log(favorite_count_mean), mientras que el R-Cuadrado ajustado indica un 1,30904%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Nº de retuits de la tendencia 14 días después (promedio)

Para tratar de predecir el valor de el número de retuits de la tendencia 14 días después (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 28

Todos: Valor-P de las variables de la regresión múltiple de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	0,0726939	0,9381
log(terms_ini_retweet_count_mean)	0,646824	0
log(terms_ini_favorite_count_mean)	-0,200717	0,2294

log(terms_ini_user_num_followers_mean)	0,0985493	0,2959
log(terms_ini_url_inclusion_rate)	0,0202266	0,9416
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 29

Todos: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	0,832979	0
log(terms_ini_retweet_count_mean)	0,633544	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_mean = exp(0,832979 + 0,633544 * log(terms_ini_retweet_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

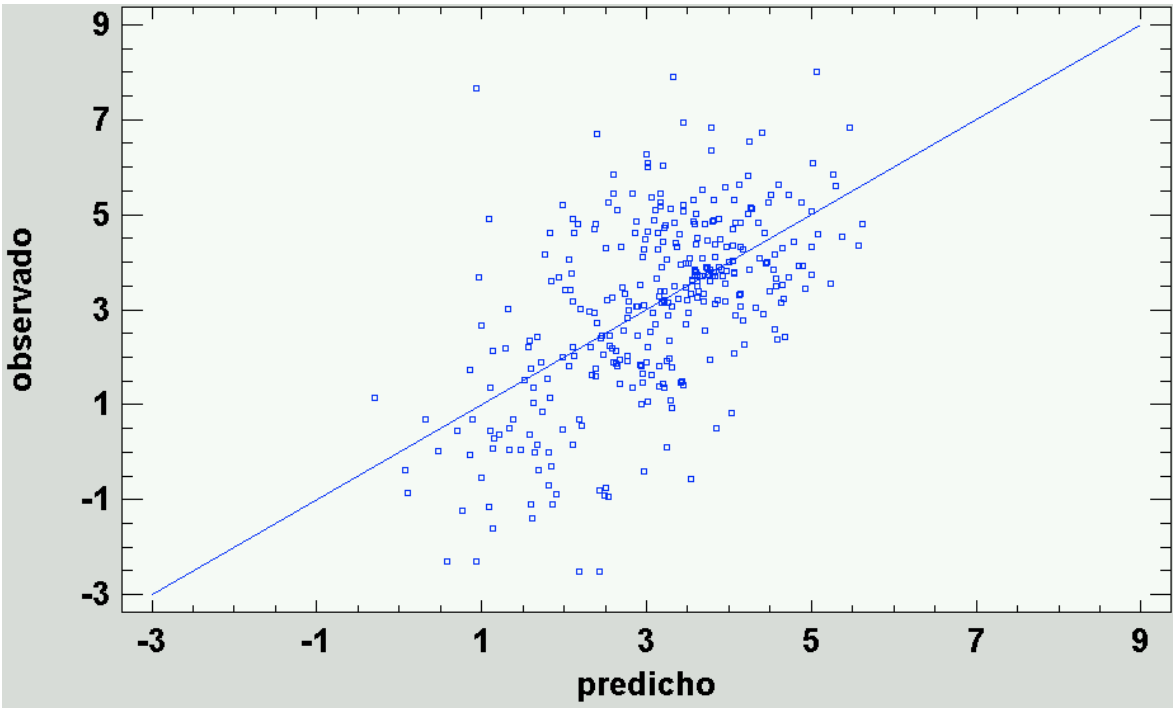


Figura 192. Todos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 33,3444% de la variabilidad de log(terms_end_retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 33,1424%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen más la línea que en las regresiones lineales múltiples anteriores.

f) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones lineales múltiples de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 30

Todos: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones lineales múltiples

Variables de predicción	Variables de éxito
terms_ini_retweet_count_mean	uniquepageviews_total
terms_ini_favorite_count_mean	avgtimeonpage_mean
terms_ini_user_num_followers_mean	pageviewspersession_mean
terms_ini_url_inclusion_rate	favorite_count_mean
	terms_end_retweet_count_mean

Se puede observar en la Tabla *30* que todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que todas son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- El promedio de retuits y el promedio de seguidores de los usuarios que participan en una tendencia explican parte de los datos de páginas vistas únicas.
- La inclusión de URL en los tuits explica parte de los datos de la duración de la visita.
- El promedio de favoritos explica parte de los datos del promedio de páginas vistas por sesión.
- La inclusión de URL en los tuits explica parte de los datos del promedio de favoritos en la cuenta del medio.
- El promedio de retuits explica en parte el promedio de retuits 14 días después.

4.2.1.4. Regresión binomial negativa o de Poisson

A continuación, se tratará de predecir todas las variables de éxito que sean de conteo (enteros y sin números negativos) a partir de todas las variables de predicción que sean independientes entre sí según la regresión binomial negativa o la regresión de Poisson.

Se trata de unos tipos de regresión que solo sirven para variables dependientes que consisten en datos de conteo, que en este caso corresponden a las variables de éxito que

suponen un total de algún elemento. De esta manera se tratará de interpretar en qué medida una unidad más o una unidad menos de una variable independiente o de predicción es asociada con un cambio porcentual en el conteo de una variable dependiente o de éxito (Bobbitt, 2019).

Las variables de éxito que tienen datos de conteo son:

- Páginas vistas únicas (total): uniquepageviews_total,
- Número de tuits de la tendencia 14 días después (total): terms_end_num_tweets,
- Número de retuits de la tendencia 14 días después (total): terms_end_retweet_count_total.

a) Filtro de alta correlación (colinealidad)

Las variables que aporten información para tratar de realizar la regresión deben ser independientes, motivo por el cual es necesario hacer un filtro de alta correlación de manera que se asegure que todas aportan información diferente.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

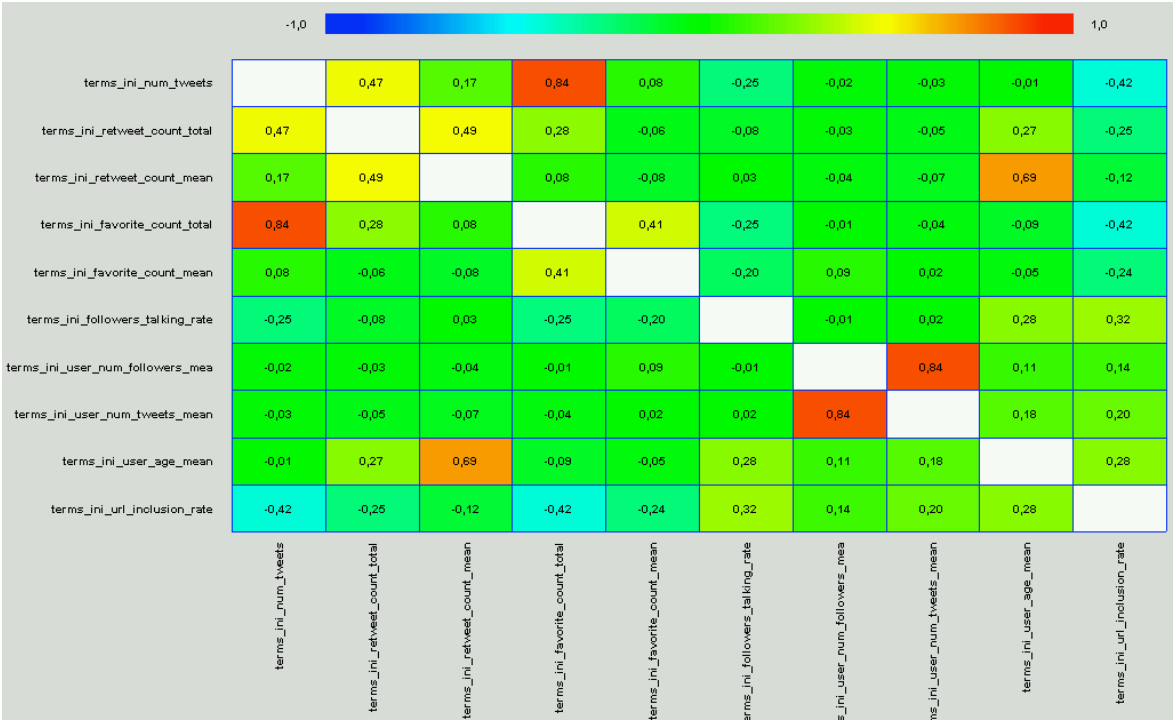


Figura 193. Todos: Matriz de correlaciones Pearson entre las variables de predicción

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_ini_num_tweets y terms_ini_favorite_count_total tienen un coeficiente de correlación de 0,8433 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- tems_ini_user_num_followers_mean y terms_ini_user_num_tweets_mean tienen un coeficiente de correlación de 0,8401 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige terms_ini_num_tweets y tems_ini_user_num_followers_mean por tener un sesgo y una curtosis estandarizados menores, como se puede comprobar en el apartado 4.2.1.2.

La tabla de variables quedaría como sigue:

Tabla 31

Todos: Lista de variables de predicción y de éxito para la regresión binomial negativa o de Poisson tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	terms_end_num_tweets
terms_ini_retweet_count_mean	terms_end_retweet_count_total
terms_ini_favorite_count_mean	
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a: el número total de tuits, el número total de retuits, el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que

participan, el promedio de edad en días de la cuenta de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

b) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión con la variable dependiente uniquepageviews_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable uniquepageviews_total, el Chi-cuadrado calculado es 261.384 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 32

Todos: Valor-P de las variables de la regresión binomial negativa de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,36014	0
terms_ini_num_tweets	0,000174773	0
terms_ini_retweet_count_total	-0,0000000539834	1
terms_ini_retweet_count_mean	-0,00329312	0,001
terms_ini_favorite_count_mean	-0,0482153	1
terms_ini_followers_talking_rate	1,87514	0
terms_ini_user_num_followers_mean	-0,00000519185	1
terms_ini_user_age_mean	0,0000100375	1
terms_ini_url_inclusion_rate	-0,0323201	1

Modelo	1

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 33

Todos: Valor-P de las variables de la regresión binomial negativa simplificada de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,22714	0
terms_ini_num_tweets	0,000169576	0
terms_ini_retweet_count_mean	-0,0000724996	0
terms_ini_followers_talking_rate	2,05712	0
terms_ini_user_num_followers_mean	-0,00000577671	0
terms_ini_url_inclusion_rate	0,052177	0
Modelo		1

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable uniquepageviews_total no puede ser predicha mediante regresión binomial negativa con las variables de que se dispone en el modelo.

c) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_num_tweets, el Chi-cuadrado calculado es 901.761.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 34

Todos: Valor-P de las variables de la regresión binomial negativa de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	18,3412	0
terms_ini_num_tweets	0,000234872	0
terms_ini_retweet_count_total	0,0000000205706	1
terms_ini_retweet_count_mean	-0,000716686	1
terms_ini_favorite_count_mean	-0,0511978	0,4585
terms_ini_followers_talking_rate	0,554972	1
terms_ini_user_num_followers_mean	-0,00000512676	1
terms_ini_user_age_mean	0,00013857	0,7095
terms_ini_url_inclusion_rate	-0,220061	0,5525
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 35

Todos: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	18,2233	0
terms_ini_num_tweets	0,00023897	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(18,2233 + 0,00023897 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

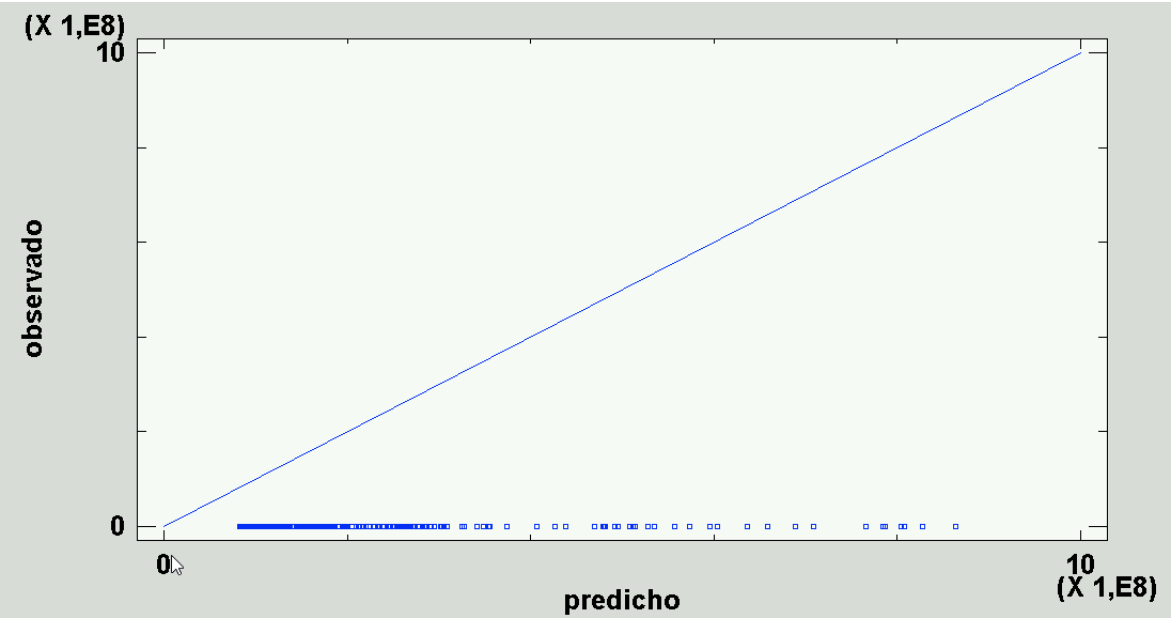


Figura 194. Todos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 8,57777% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 7,99404%. Este

número se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

d) Nº de retuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de retuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_retweet_count_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir la regresión de binomial negativa y la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_retweet_count_total, el Chi-cuadrado calculado es 2.367.740.000.000.000 con un valor-P cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 36

Todos: Valor-P de las variables de la regresión de binomial negativa de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	49,5159	0
terms_ini_num_tweets	0,00020689	0
terms_ini_retweet_count_total	-0,000000272783	0
terms_ini_retweet_count_mean	0,00244261	0
terms_ini_favorite_count_mean	-0,0645126	0
terms_ini_followers_talking_rate	-52,1232	0
terms_ini_user_num_followers_mean	-0,000133018	1
terms_ini_user_age_mean	-0,00189127	1
terms_ini_url_inclusion_rate	2,53398	0

Modelo	1

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 37

Todos: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	36,4392	0
terms_ini_num_tweets	0,000519484	0
terms_ini_retweet_count_total	-0,000000394099	0
terms_ini_retweet_count_mean	0,00471131	0
terms_ini_followers_talking_rate	-35,8743	0
terms_ini_user_age_mean	-0,00274665	0,0119
terms_ini_url_inclusion_rate	9,38011	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_total = exp(36,4392 + 0,000519484 * terms_ini_num_tweets - 0,000000394099 * terms_ini_retweet_count_total + 0,00471131 *

terms_ini_retweet_count_mean - 35,8743 * terms_ini_followers_talking_rate - 0,00274665 * terms_ini_user_age_mean + 9,38011 * terms_ini_url_inclusion_rate)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

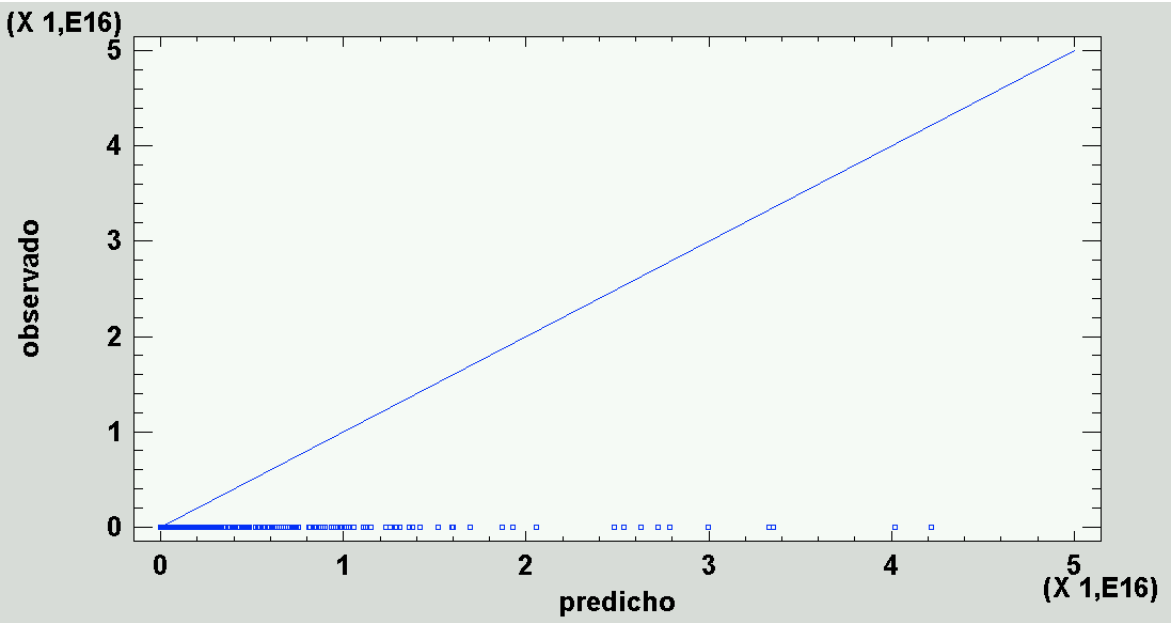


Figura 195. Todos: Gráfico de la relación entre la función de predicción según la regresión de binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 11,1255% de la variabilidad de terms_end_retweet_count_total, mientras que el R-Cuadrado ajustado indica un 9,22799%. Este número se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

Con la intención de tratar de mejorar las predicciones, se procede a realizar el análisis siguiendo el mismo procedimiento, pero limitándose a cada una de las secciones principales del medio a estudiar. Como solo tres de ellas tienen 30 elementos o más publicados en el periodo analizado, se hará para estas tres a continuación.

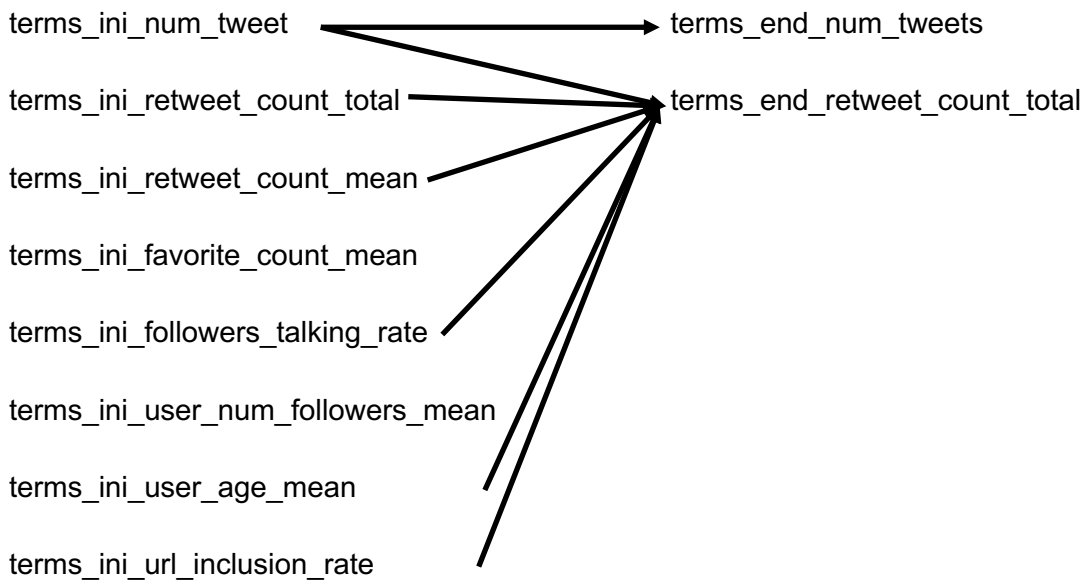
e) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones binomiales negativas de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 38

Todos: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones binomiales negativas

Variables de predicción	Variables de éxito
terms_ini_num_tweet	terms_end_num_tweets
terms_ini_retweet_count_total	terms_end_retweet_count_total
terms_ini_retweet_count_mean	
terms_ini_favorite_count_mean	
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	



Se puede observar en la Tabla *38* que no todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que no todas son necesarias para las variables de éxito elegidas y que se pueden estudiar. Es el caso del promedio de favoritos y la ratio de seguidores de los usuarios que participan en la tendencia.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits explica parte de los datos del número de tuits 14 días después.
- El número de tuits, el número de retuits, el promedio de retuits, la ratio de seguidores de la cuenta del medio que participan, el promedio de la edad en días de los usuarios que participan y la ratio de inclusión de URL en los tuits explican en parte el número de retuits 14 días después.

4.2.2. Análisis de subconjuntos de artículos

Se ha incluido en Anexos los análisis de los subconjuntos de artículos, pudiendo leerse el de Cine en el anexo 6.1.1, el de Series en el anexo 6.1.2, el de Videojuegos en el anexo 6.1.3 y el de Tráileres en el anexo 6.1.4.

4.2.3. Resumen de todos los análisis

Debido a la gran cantidad de análisis que se han realizado para cada categoría, resulta conveniente resumir las predicciones que se han obtenido y qué conclusiones se obtienen de estas.

A continuación, se puede ver la tabla resumen de R-cuadrado de los análisis realizados de tipo regresión lineal múltiple:

Tabla 39

Resumen de R-cuadrado de cada análisis realizado de regresión lineal múltiple por cada variable de éxito (ya sea con transformación logarítmica o no)

Variable	Todos (350)	Cine (187)	Series (104)	Videojuegos (35)	Tráileres (101)
uniquepageviews_total	2,87745%	3,26236%	6,38706%	46,0113%	6,04245%
adsense_ecpm_mean					8,49787%
avgtimeonpage_mean	1,61182%	4,02429%	9,31683%		10,9437%
pageviewspersession_mean	2,23659%	5,86202%	4,09697%		18,3504%
retweet_count_mean					11,7912%
favorite_count_mean	1,63475%	3,89283%	11,6459%		18,2644%
terms_end_num_tweets			72,2708%	51,3521%	
terms_end_retweet_count_total					

terms_end_retweet _count_mean	33,3444%	33,3411%	43,4481%	36,7106%	45,2254%

En la Tabla 39 se puede ver que las únicas variables de éxito que se han podido predecir en todos los casos son: uniquepageviews_total y terms_end_retweet_count_mean. La variable terms_end_retweet_count_total no ha podido predecirse en ningún caso.

En todos los casos, el R-cuadrado se ha visto beneficiado por una selección más restrictiva de la población a estudiar, ya sea por temática o por contenido (tráileres), salvo la excepción de la variable terms_end_retweet_count_mean entre Todos y Cine, en la cual es prácticamente la misma.

En general el R-cuadrado se ha visto influenciado en gran medida por las variables de predicción disponibles en el modelo. Si alguna de ellas que podría explicar una parte de los datos se ha visto eliminada en el proceso de normalización, el R-cuadrado se ha visto perjudicado por ello. Es algo que ha afectado sobre todo a la población de todos los artículos, ya que había más variables que no seguían una distribución casi normal y ha habido mucha menos información de predicción para los modelos de regresión lineal múltiple.

El R-cuadrado también se ha visto influenciado por las variables que se han eliminado en el proceso de filtrado debido a que compartían un alto coeficiente de correlación. En esta tesis se ha optado por eliminar las variables con mayores sesgo y curtosis estandarizados que tuvieran una fuerte correlación con otra (0,7 o mayor). Otra posible solución sería combinar linearmente dichas variables de manera que no se pierda la información que no está correlacionada al eliminar una de ellas (Anon., 2013). Sin embargo, el problema de ello es que, al ser variables combinadas, dificulta la identificación de la influencia de cada una por separado en las predicciones. En futuros estudios se podría sacrificar esa claridad de identificación e investigar cuál es la repercusión de cada opción, para así comprobar si merece la pena combinarlas. En el caso de esta tesis, la identificación de la repercusión de cada variable es muy importante de cara a extraer conclusiones posteriores.

Salvo en el caso de la variable uniquepageviews_total, la selección de la población por contenido (tráileres) ha dado el mejor resultado en el R-cuadrado de las predicciones. Ha permitido, además, el mayor número de predicciones, siendo capaz de predecir siete de las nueve variables de éxito.

Po otro lado, también se han realizado regresiones binomiales negativas para tratar de predecir las variables de éxito de tipo conteo. A continuación, se puede ver la tabla resumen de R-cuadrado de los análisis realizados de este tipo:

Tabla 40

Resumen de R-cuadrado de cada análisis realizado de regresión binomial negativa por cada variable de éxito

Variable	Todos (350)	Cine (187)	Series (104)	Videojuegos (35)	Tráileres (101)
uniquepageviews_total		6,54364%	9,2758%	60,1601%	3,83227%
terms_end_num_tweets	8,57777%	9,05712%	6,99566%	10,4097%	9,66905%
terms_end_retweet_count_total	11,1255%	20,0829%	3,30127%	6,38892%	9,81024%

Si se compara la tabla de R-cuadrado de las predicciones según la regresión binomial negativa con la de las predicciones según una regresión lineal múltiple, la binomial negativa consigue un mayor porcentaje para las páginas vistas únicas (uniquepageviews_total) y, por tanto, una mayor capacidad de predicción. En los casos en los que el número de tuits 14 días después (terms_end_num_tweets) se puede predecir con la regresión lineal múltiple, esta tiene un mayor R-cuadrado. Por otro lado, se cuenta con la ventaja de poder predecir el número de retuits 14 días después (terms_end_retweet_count_total) en todos los casos, ya que no fue posible en ninguno de ellos con la regresión lineal múltiple.

Se probarán ambas en la siguiente fase de la investigación, de manera que se obtenga su precisión calculando el margen de error de predecir las variables con cada fórmula obtenida en los análisis previos.

4.3. Fase 2. Análisis de los datos de test

Se ha efectuado un análisis de los artículos de tipología de noticia de última hora publicados entre el 23 de noviembre y el 22 de diciembre de 2020 en el medio Hello Friki, sumando un total de 178 artículos publicados, que sirven como conjunto de datos de test.

En este apartado, el objetivo es validar las ecuaciones de predicción obtenidas en las regresiones del apartado 4.2. Para ello, se utilizará la raíz del error medio cuadrático (*Root Mean Square Error*, RMSE), un método de validación de resultados experimentales muy común en el análisis de regresión. La raíz del error medio cuadrático es la desviación estándar de los residuos o errores de predicción, siendo estos la medida de distancia entre los datos y la línea de la regresión, definida por la fórmula. Por tanto, la raíz del error medio cuadrático mide hasta qué punto los residuos están dispersos o, dicho de otro modo, la concentración o cercanía de los datos a la línea de mejor ajuste, definida por la fórmula (Glen, 2020). Cuanto menor sea el valor de RMSE, mejor será el ajuste de la predicción a los datos. RMSE es una medida apropiada de la precisión con la que el modelo predice los datos, y el criterio más importante si el objetivo es la predicción en sí. (Grace-Martin, 2008)

La fórmula para calcular la raíz del error medio cuadrático es, según Barnston (1992), la siguiente:

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

Donde:

f = valores predichos

o = valores observados

N = tamaño de la muestra

z = dato

En suma, la fórmula anterior calcula el cuadrado de los residuos (diferencias entre el valor predicho y el observado), calcula su promedio y elimina el cuadrado del resultado mediante la raíz cuadrada.

Con ello, se podrá comprobar hasta qué punto se ha errado en la predicción según las ecuaciones disponibles y, por tanto, hasta qué punto tiene repercusión su predicción en la toma de decisiones de la estrategia editorial del medio. Se comparará el RMSE de la fase

1, para la cual se obtuvieron las ecuaciones de predicción, y el RMSE de la fase 2 con un conjunto de datos nuevo (test). Si ambos son parecidos, se confirmará la validez de la predicción.

Por otro lado, para comprobar también la precisión, se obtendrá asimismo el *Scatter Index* (SI), que consiste en normalizar el RMSE por el valor medio de los datos observados. Según Campos y Soares (2016), cuanto menor sea el valor tanto en el RMSE como especialmente en el *Scatter Index* (ya que este se calcula de manera relativa a cada variable), más precisa habrá sido la predicción:

$$\mathrm{SI} = \frac{\mathrm{RMSE}}{\bar{S}}$$

Donde:

- S = valores observados

4.3.1. Validación de la predicción de todos los artículos

A continuación, se incluyen tablas con el R-Cuadrado de la predicción de cada tipo de regresión, calculada en el apartado 4.2.1, la raíz del error medio cuadrático (RMSE) y el *Scatter Index* (SI) resultantes de comparar los valores predichos con los observados en todos los artículos publicados tanto en la fase 1 como en la fase 2.

Tabla 41

Todos: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Lineal Múltiple

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	2,87745%	64,53935	306,93455	4,04999
avgtimeonpage_mean	1,61182%	140,42575	105,15186	1,06722
pageviewspersession_mean	2,23659%	0,78205	0,82038	0,60404
favorite_count_mean	1,63475%	0,98869	1,01762	0,9928
terms_end_retweet_count_mean	33,3444%	276,87733	152,15648	1,88999

Tal y como se puede comprobar en la Tabla 41, las páginas vistas totales (uniquepageviews_total) han presentado un RMSE significativamente mayor en la fase 2, lo cual ha provocado un SI muy grande de más de 4. El RMSE de las páginas vistas por sesión (pageviewspersession_mean) y el número medio de favoritos en la cuenta del medio (favorite_count_mean) tienen valores parecidos y con un SI menor al resto de variables. La duración promedio de la visita (avgtimeonpage_mean) y el promedio de retuits de la tendencia (terms_end_retweet_count_mean) han disminuido significativamente, pero presentando un SI mayor que 1.

Tabla 42

Todos: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Binomial Negativa

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)

terms_end_num_tweets	8,57777%	362.755.973,43718	695.023.087,10875	264.745,769270242
terms_end_retweet_count_total	11,1255%	248.505.507.112.460.000	576.790.006.189.960.000	727.878.085.240,5603

En la Tabla 42 se observan valores más altos en el RMSE de la fase 2, provocando SI extremadamente altos.

4.3.2. Validación de la predicción de los artículos de la categoría Cine

A continuación, se incluyen tablas con el R-Cuadrado de la predicción de cada tipo de regresión, calculada en el apartado 6.1.1, y la raíz del error medio cuadrático (RMSE) y el *Scatter Index* (SI) resultantes de comparar los valores predichos con los observados en los artículos de la categoría Cine publicados tanto en la fase 1 como en la fase 2.

Tabla 43

Cine: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Lineal Múltiple

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	3,26236%	38,54189	96,06163	1,26753
avgtimeonpage_mean	4,02429%	133,5426	88,78624	0,90112
pageviewspersession_mean	5,86202%	0,84527	0,93788	0,69056
favorite_count_mean	3,89283%	1,09269	1,00286	0,9784
terms_end_retweet_count_mean	33,3411%	273,66438	88,78487	1,10283

En la Tabla 43 se puede observar que los SI son, mayoritariamente, menores que en las predicciones generales del apartado 4.3.1, lo cual demuestra que, salvo en el caso del promedio de páginas vistas por sesión (pageviewspersession_mean) por una diferencia de 0,087, la predicción para Cine es mejor que la general.

Tabla 44

Cine: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Binomial Negativa

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)

uniquepageviews_total	6,54364%	28,56437	83,04752	1,620439415
terms_end_num_tweets	9,05712%	295.970.621,85039	708.283.433,57701	374.592,104493180
terms_end_retweet_count_total	20,0829%	844.712.918.285.210.000	20.903.657.318.052.000.000	91.323.252.404.661,1708

Según se aprecia en la Tabla 44, las páginas vistas (uniquepageviews_total) tienen un SI más cercano a 1, mientras que las otras dos variables tienen un SI extremadamente mayor, debido a un aumento muy significativo del RMSE en la fase 2 con respecto a la fase 1 que ha provocado un gran incremento en el SI.

4.3.3. Validación de la predicción de los artículos de la categoría Series

A continuación, se incluyen tablas con el R-Cuadrado de la predicción de cada tipo de regresión, calculada en el apartado 6.1.2, y la raíz del error medio cuadrático (RMSE) y el *Scatter Index* (SI) resultantes de comparar los valores predichos con los observados en los artículos de la categoría Series publicados tanto en la fase 1 como en la fase 2.

Tabla 45

Series R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Lineal Múltiple

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	6,38706%	96,47029	462,09336	6,09730
avgtimeonpage_mean	9,31683%	138,42724	132,93239	1,34917
pageviewspersession_mean	4,09697%	0,84274	0,85498	0,62952
favorite_count_mean	11,6459%	0,92892	1,14198	1,11413
terms_end_num_tweets	72,2708%	1.215,61175	1.533,45615	0,58412
terms_end_retweet_count_mean	43,4481%	172,03596	177,3212	2,20257

En la Tabla 45 se comprueba un SI mayor en todas las variables, salvo en el caso del número de tuits de la tendencia (terms_end_num_tweets), que no estaba disponible en el análisis general (apartado 4.3.1). Esto demuestra que la predicción, en el caso de la sección de Series, es peor que la realizada por todos los artículos, pese a que el RMSE de la fase 2 se ha mantenido parecido al de la fase 1 salvo en el caso de las páginas vistas (uniquepageviews_total) y el número de tuits (terms_end_num_tweets).

Tabla 46

Series: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Binomial Negativa

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)

uniquepageviews_total	9,2758%	85,2248	458,44148	4,705224621
terms_end_num_tweets	6,99566%	266.171.237,84021	622.105.770,6356	243.293,89240533
terms_end_retweet_count_total	3,30127%	165.064.452.012.530	162.510.085.984.470	401.970.209,0824

Según se aprecia en la Tabla 46, todas las variables tienen un valor SI alto. Las páginas vistas (uniquepageviews_total) tienen un SI menos alto, mientras que las otras dos variables tienen un SI extremadamente alto, debido a un aumento muy significativo del RMSE en la fase 2 con respecto a la fase 1 que ha provocado un gran incremento en el SI.

4.3.4. Validación de la predicción de los artículos de la categoría Videojuegos

A continuación, se incluyen tablas con el R-Cuadrado de la predicción de cada tipo de regresión, calculada en el apartado 6.1.3, y la raíz del error medio cuadrático (RMSE) y el *Scatter Index* (SI) resultantes de comparar los valores predichos con los observados en los artículos de la categoría Videojuegos publicados tanto en la fase 1 como en la fase 2.

Tabla 47

Videojuegos: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Lineal Múltiple

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	46,0113%	23,97768	65,16239	0,85982
terms_end_num_tweets	51,3521%	7.447,80119	11.116,93479	2,10369
terms_end_retweet_count_mean	36,7106%	510,66842	266,24808	3,30717

En la Tabla 47 se observa que las páginas vistas (uniquepageviews_total) han presentado un SI mucho menor que en el caso general (apartado 4.3.1), con una diferencia de 3,19. El número de tuits (terms_end_num_tweets) también es apropiado en este caso, ya que no está disponible en el caso general, mientras que el promedio de retuits de la tendencia (terms_end_retweet_count_mean) ha empeorado en un 1,42, siendo más conveniente usar la predicción general en ese caso.

Tabla 48

Videojuegos: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Binomial Negativa

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	60,1601%	19,05703	61,29738	0,904758376

terms_end_num_tweets	10,4097%	714.288.540,3053	748.080.768,02186	141.561,314792669
terms_end_retweet_count_total	6,38892%	1.441.045.348.618.300.000	16.346.613.290.769.000	3.287.446.887,0522

Según se aprecia en la Tabla 48, las páginas vistas (uniquepageviews_total) tienen un SI menor que 1, mientras que las otras dos variables tienen un SI extremadamente mayor, debido a un aumento muy significativo del RMSE en la fase 2 con respecto a la fase 1 que ha provocado un gran incremento en el SI.

4.3.5. Validación de la predicción de los artículos sobre Tráileres

A continuación, se incluyen tablas con el R-Cuadrado de la predicción de cada tipo de regresión, calculada en el apartado 6.1.4, y la raíz del error medio cuadrático (RMSE) y el *Scatter Index* (SI) resultantes de comparar los valores predichos con los observados en los artículos sobre Tráileres publicados tanto en la fase 1 como en la fase 2.

Tabla 49

Tráileres: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Lineal Múltiple

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	6,04245%	70,07766	57,60066	0,76003
adsense_ecpm_mean	8,49787%	0,05946	0,14134	2,80787
avgtimeonpage_mean	10,9437%	150,51153	150,44009	1,52686
pageviewspersession_mean	18,3504%	1,00766	1,2001	0,88363
retweet_count_mean	11,7912%	0,55238	0,55513	0,88614
favorite_count_mean	18,2644%	1,20543	1,00843	0,98383
terms_end_retweet_count_mean	45,2254%	154,90103	133,77985	1,66173

En la Tabla 49 hay variables presentes que no estaban en el caso general (apartado 4.3.1), como son el promedio de eCPM de Google AdSense (adsense_ecpm_mean) y el promedio de retuits en la cuenta del medio (retweet_count_mean). Las páginas vistas (uniquepageviews_total) tienen un SI mucho menor, al igual que el promedio de favoritos en la cuenta del medio (favorite_count_mean) y el promedio de retuits de la tendencia (terms_end_retweet_count_mean), aunque ambos en menor grado. La duración de la visita (avgtimeonpage_mean) y las páginas vistas por sesión (pageviewspersession_mean), en cambio, tienen valores más altos que en el caso general.

En cuanto a las demás secciones, la selección temática de Tráileres ha presentado un menor SI en páginas vistas en todos los casos. La duración de la visita (avgtimeonpage_mean) y las páginas vistas por sesión (pageviewspersession_mean) han empeorado con respecto a los casos en los que estaba disponible. El promedio de favoritos en la cuenta del medio (favorite_count_mean) es ligeramente menor que en Series (0). El promedio de retuits de la tendencia (terms_end_retweet_count_mean) es menor que en Series (0) y Videojuegos (0), pero mayor que en Cine (0). En el caso del eCPM de Google AdSense (adsense_ecpm_mean) y el promedio de retuits en la cuenta del medio (retweet_count_mean), este es el único apartado en el que están disponibles.

Tabla 50

Tráileres: R-Cuadrado, la raíz del error medio cuadrático y el Scatter Index de cada variable de éxito según la Regresión Binomial Negativa

Variable	R-Cuadrado	RMSE (fase 1)	RMSE (fase 2)	SI (fase 2)
uniquepageviews_total	3,83227%	64,69814	56,71271	1,343009751
terms_end_num_tweets	9,66905%	191.349.583,95872	613.426.846,31294	427.349,762767226
terms_end_retweet_count_total	9,81024%	12.956.614.341.885.000	24.375.900.093.443.000	27.418.837.373,8368

Según se aprecia en la Tabla 50, las páginas vistas (uniquepageviews_total) tienen un SI más cercano a 1, mientras que las otras dos variables tienen un SI extremadamente mayor, debido a un aumento muy significativo del RMSE en la fase 2 con respecto a la fase 1 que ha provocado un gran incremento en el SI.

Si se compara el SI de uniquepageviews_total con el obtenido en cada sección, tiene un SI menor salvo en el caso de Videojuegos.

4.4. Selección de ecuaciones de predicción

Para llevar a cabo la selección de las ecuaciones de predicción a utilizar para cada variable y cada sección o selección temática, se utilizará como baremo el menor SI, como se puede apreciar en la Tabla 51, en la que se incluyen tanto los *Scatter Index* de las Regresiones Lineales Múltiples (RLM) como de las Regresiones Binomiales Negativas (RBN):

Tabla 51

Ecuaciones de predicción seleccionadas para cada variable según el criterio de un menor SI

Variable	SI - RLM (general)	SI - RLM (Cine)	SI - RLM (Series)	SI - RLM (Videojuegos)	SI - RLM (Tráileres)	SI - RBN (general)	SI - RBN (Cine)	SI - RBN (Series)	SI - RBN (Videojuegos)	SI - RBN (Tráileres)
uniquepageviews_total	4,04999	1,26753	6,09730	0,85982	0,76003		1,62043941	4,7052246	0,90475837	1,343009751
adsense_ecpm_mean					2,80787					
avgtimeonpage_mean	1,06722	0,90112	1,34917		1,52686					
pageviewspersession_mean	0,60404	0,69056	0,62952		0,88363					
retweet_count_mean					0,88614					
favorite_count_mean	0,9928	0,9784	1,11413		0,98383					
terms_end_num_tweets			0,58412	2,10369		264.745,769270242	374.592,104493180	243.293,89240533	141.561,314792669	427.349,762767226
terms_end_retweet_count_total						727.878.085.240,5603	91.323.252.404.661,170	401.970.209,0824	3.287.446.887,0522	27.418.837.373,8368
terms_end_retweet_count_mean	1,88999	1,10283	2,20257	3,30717	1,66173					

Para seleccionar la fórmula de predicción de una variable de éxito de interés, hay que elegir la que menor SI tenga entre sus diferentes posibilidades. Para facilitar la comprensión, se incluyen a continuación varios ejemplos ilustrativos:

- Si se desea predecir el promedio de tiempo de visita (avgtimeonpage_mean) de una noticia de Cine, se acudiría a la fórmula de predicción de RLM de Cine, con un SI de 0,90112 menor que el SI general de 1,06722.

- Si se pretende predecir el promedio de favoritos en la cuenta del medio (favorite_count_mean) de una noticia de tráiler de Series, se acudiría a la fórmula de predicción de RLM de Tráileres, con un SI de 0,98383 menor que el SI general de 0,9928 y el SI de Series de 1,11413.

- Si se escoge predecir las páginas vistas únicas (uniquepageviews_total) de una noticia de Videojuegos, se acudiría a la fórmula de predicción de RLM de Videojuegos, con un SI de 0,85982 menor que el SI (RLM) general de 4,04999 y el SI (RBN) de Videojuegos de 0,904758376.

- Si se escoge predecir las páginas vistas (uniquepageviews_total) de una noticia de tráiler de Cine, se acudiría a la fórmula de predicción de RLM de Tráileres, con un SI de 0,74869 menor que el SI (RLM) general de 4,04999, el SI (RLM) de Cine de 1,26753, el SI (RBN) de Cine de 1,620439415 y el SI (RBN) de Tráileres de 1,343009751.

4.5. Discusión de los resultados finales

La predicción se llevará a cabo con la función de predicción que más precisión ha demostrado en el experimento para cada caso concreto, tal y como se ha ilustrado en el apartado 4.4. Los valores devueltos por estas predicciones servirán como ayuda en la toma de decisiones.

Sin embargo, la precisión puede empeorar significativamente tanto por la presencia de valores anómalos en los datos de entrenamiento como en los datos de test, así como por factores que afecten a la variabilidad (R-cuadrado) no tenida en cuenta en las respectivas ecuaciones de predicción. Esto último significa que si una fórmula de predicción cuenta con un R-cuadrado de 18,3504%, como es el caso de la variable de páginas vistas por sesión (pageviewspersession_mean) de Tráileres (RLM), hay una variabilidad del 81,6496% no tenida en cuenta y que puede afectar de manera significativa a la precisión del modelo.

También se ha observado una sobredimensión de gran tamaño en algunos de los indicadores, especialmente los relacionados con el análisis de tendencias, ya que responden a datos anómalos cuyo origen no se ha podido estudiar en el presente estudio pero que, sin embargo, no se han podido eliminar del mismo debido a que forman parte de las características inherentes de la red social.

Para mejorar la precisión de las predicciones, se podría optar por analizar intervalos de tiempo mucho mayores, analizar los valores anómalos de manera separada al del resto de datos, tratar de encontrar nuevas variables de predicción que puedan aumentar la variabilidad (R-cuadrado) y seleccionar unidades temáticas mucho más específicas, siempre y cuando se cuente con una población suficiente de artículos para generar conclusiones estadísticas.

5. Conclusiones

Esta tesis se enmarca en el ámbito del marketing de contenidos y abre nuevas vías de investigación sobre el éxito de los contenidos online y su posible relación con el retorno de conversión en Internet. Se ha establecido como objetivo principal el diseño de una metodología cibermétrica de cálculo del éxito y su posible predicción de una publicación online. Todo ello integrado dentro de la estrategia de contenidos de un medio digital, que permite abordar el problema de investigación de optimizar la publicación de contenidos web en el ámbito de periodismo.

Esta metodología sirve como modelo de análisis de la comunicación en internet y la interrelación de las variables. Aporta un seguimiento y sirve como herramienta en la toma de decisiones de una estrategia de contenidos. Por lo tanto, esta metodología puede servir como base para nuevas ideas que den origen a otras investigaciones con hipótesis que varíen en su metodología o estudien diferentes recursos a los que aquí se han utilizado.

En los resultados expuestos en el capítulo anterior se puede observar el diseño y ejecución de dicha metodología, así como su validación mediante datos de test que han permitido evaluar la precisión de las predicciones. Esto permite, por tanto, mejorar la toma de decisiones del equipo editorial, ya que podrá optimizar tanto los recursos destinados a unos contenidos u otros, como el mismo contenido en sí.

Además, el estudio de tendencias y del interés de los usuarios en Twitter en el ámbito del entretenimiento, concretamente en las temáticas del sitio web que ha servido como caso de uso, ha posibilitado investigar cuáles son las nuevas tendencias de comunicación en dicho ámbito y la versatilidad de este, así como la necesidad de consumo de información de actualidad en los ámbitos estudiados.

A continuación, se mostrará una lista de objetivos específicos que han hecho posible el diseño de la metodología, así como una serie de conclusiones generales sobre la contribución de esta tesis doctoral tanto a la academia como a la industria del periodismo digital. Por último, se aportará una serie de posibles líneas de investigación que respondan a las limitaciones de este estudio o aborden las cuestiones teniendo en cuenta otros indicadores e incluso ámbitos de estudio.

5.1. Objetivos específicos

Para realizar el diseño de la metodología cibermétrica, ha sido necesario cumplir los siguientes objetivos específicos:

5.1.1. Investigar el concepto de éxito en periodismo digital, la red social Twitter, la analítica web y la publicidad en la web

Esta investigación se ha desarrollado en varios ámbitos multifactoriales, en los cuales el concepto de éxito depende de los objetivos y prioridades del medio en cuestión. Es por ello por lo que, para diseñar una metodología que sirva para casos de uso diferentes, ha sido necesario investigar las características de los ámbitos en lo que se ha trabajado.

Por tanto, se ha realizado una extensa revisión en forma de estado de la cuestión en el cual se ha hecho énfasis en el periodismo digital, Twitter, la difusión de las noticias en Twitter, la analítica web, la cibermetría, la analítica en Twitter, el análisis de tendencias en Twitter y la publicidad digital en la web. Todos estos ámbitos han aportado una visión global de qué ha abordado la academia y otras fuentes de autoridad hasta ahora en cuanto al éxito del contenido digital de una web. Éxito comprendido como la optimización y mejora del rendimiento de los indicadores aportados por los diferentes estudios que se han consultado.

Los indicadores cibermétricos aportados por las fuentes han permitido hacer una selección de éstos, incluida en el apartado 3.4. Debido a que el caso de uso a estudiar disponía de anuncios de Google AdSense, se ha hecho un especial énfasis en este sistema de publicidad web, siendo además el más utilizado actualmente.

Los indicadores seleccionados se han dividido en tres grupos: analítica del contenido en la web, analítica del contenido en la cuenta de Twitter del medio y análisis de tendencias en Twitter. Así, cada grupo ha respondido a una dimensión de éxito diferente.

5.1.2. Diseñar la metodología y determinar qué herramientas y reportes son necesarios

Una vez seleccionados los indicadores, se ha investigado qué herramientas podrían aportar su información. En el caso de uso en cuestión, se contaba con los datos de Google Analytics para los indicadores de la analítica del contenido en la web, por lo que se ha acudido a la API de informes de Google Analytics v4. Por otro lado, para obtener los

indicadores relacionados con Twitter, es decir, los datos de la cuenta de Twitter del medio y los del análisis de las tendencias que estuvieran relacionadas con los contenidos, se ha optado por Twitter Standard API.

Cada API cuenta con una serie de reportes que ha posibilitado la obtención de los datos necesarios para el estudio, aunque también ha impuesto limitaciones, sobre todo debido a las características de una cuenta gratuita, que han sido explicadas en el apartado 0.

Para el diseño de la metodología, se han evaluado diferentes métodos de regresión estadística que permiten predecir valores a partir de unos datos de entrenamiento. Estos métodos exigen que las variables cumplan una serie de condiciones, tal y como han indicado autores como Barón López y Téllez Montiel (s. f.), Bobbitt (2019) y Navaro et al. (2001).

5.1.3. Extraer los datos y procesarlos para obtener los indicadores

Para la extracción de los datos, ha sido necesario diseñar la base de datos y las tablas en las que se ha incluido la información obtenida de las APIs, tal y como se puede observar en el Apartado 3.6.1, así como los algoritmos de recogida de datos (Apartado 3.6.2), que se han debido ejecutar de manera que respetaran los límites de peticiones de cada API.

El uso de estas API ha posibilitado la obtención de 325.658 datos en total, cuya clasificación está en el Apartado 4.1.1. Su procesamiento ha permitido obtener los valores de los indicadores de las diferentes dimensiones de éxito.

5.1.4. Realizar regresiones que permitan obtener ecuaciones de predicción de las variables de éxito seleccionadas

La predicción se ha realizado mediante la Regresión Lineal Múltiple y la Regresión Binomial Negativa, incluyendo además la posibilidad de realizar en su lugar la de Poisson si no hubiera sobre dispersión en los datos de las variables de conteo.

Esto ha permitido no solo la obtención de las ecuaciones de predicción de las variables de éxito que se han seleccionado para el experimento, sino que además se ha podido observar qué variables explican cada variable de éxito para cada población, y cuáles no explican ninguna en absoluto y son por tanto redundantes para el caso de uso en cuestión.

Todo ello aporta información valiosa, ya no solo para tratar de predecir los indicadores de éxito, sino también para optimizar los contenidos de manera que éstos se vean reforzados.

Esto se puede hacer prestando atención a las variables que afectan a los indicadores de interés del medio, y tratando de potenciar estas mediante diferentes técnicas de elaboración de contenidos y marketing.

5.1.5. Validar las ecuaciones de predicción con datos de test y obtener su precisión

Las ecuaciones de predicción han sido validadas con los datos de test de otro periodo de tiempo, en el cual se ha analizado también los contenidos publicados por el medio. Así, se ha podido evaluar la precisión con la que las ecuaciones han predicho los indicadores con una población y por tanto unos datos diferentes.

Esta precisión sirve para establecer el grado de confianza que depositar en la fórmula de predicción, así como establecer una base para futuras investigaciones en las cuales tratar de mejorarla variando los múltiples factores que actúan en el experimento.

5.2. Conclusiones generales

Este estudio ha abordado el problema de investigación del éxito del contenido digital desde un prisma nuevo, ya que se ha optado por diseñar una metodología que se pueda integrar en la toma de decisiones del equipo editorial de un medio, y que además se pueda adaptar a otros indicadores y casos de uso. De esta manera, se responde a la necesidad de saber a priori el éxito de un contenido digital antes de invertir recursos en él.

Este ámbito de investigación resulta muy interesante, tanto a nivel académico como a nivel privado. La investigación de las redes sociales, el marketing digital y la analítica web están viéndose reforzadas en los últimos años, debido tanto a las continuas y frecuentes innovaciones tecnológicas como al predominio cada vez más fuerte del mundo digital en el día a día del ser humano. Todo ello también afecta a la industria, que se debe adaptar a las nuevas circunstancias y optimizar sus procesos ya no solo para aumentar o mantener los ingresos, sino incluso para no desaparecer en un entorno cada vez más globalizado.

Esta tesis prosigue el discurso científico sobre el rendimiento del contenido en internet, especializándose en el periodístico y de noticias de última hora, solo que lo inicia desde un punto de vista más general y teniendo en cuenta diferentes ámbitos. Como se puede observar en el estado de la cuestión, muchos investigadores han abordado la problemática en busca de los indicadores más adecuados para cada ámbito, pero siempre circunscribiéndose a un ámbito e incluso a unos indicadores en concreto, sin adaptarse a la visión global del contenido en varios ámbitos y a la estrategia de contenidos de un medio, con sus necesidades y peculiaridades.

A lo largo del estudio, se ha determinado que los indicadores propuestos por autores como Kaushik (2011), Gutiérrez Argüello (2013), Suh et al. (2010) y Thelwall y Cugelman (2017) sí que han aportado una parte de la variabilidad necesaria para explicar los indicadores de éxito. Sin embargo, en esta tesis ha sido destacable su número bajo en algunos casos, que ha provocado una disminución de la precisión, lo que lleva a proponer el uso de un número mayor de indicadores en una búsqueda de representación de una amplia variabilidad para diversas interpretaciones de éxito según el contexto. Esto responde al carácter fuertementemente multifactorial de internet, por lo que este experimento aporta un valor añadido con el descubrimiento de la presencia de relaciones entre las variables, ya no solo para su predicción, sino para su posible análisis y optimización.

Cabe señalar que las herramientas utilizadas para la extracción de datos, las API de Google Analytics (apartado 3.5.1) y Twitter (apartado 3.5.2), ofrecen limitaciones

principalmente técnicas que exigen replantear el marco temporal y los máximos de algunos valores.

Pese a ello, no solo se ha demostrado que se puede predecir parte de la información de los indicadores de éxito, sino que también invita a análisis más granulares y específicos, ya que se ha observado una mejora en la precisión de la predicción en poblaciones más específicas. Se trata, además, de una metodología cibermétrica que invita a la comunidad científica y tecnológica a aplicarla en casos de uso, muestras e indicadores diferentes, de manera que se pueda mejorar la calidad de las predicciones obtenidas.

El carácter fuertemente multifactorial del ámbito de Internet, por tanto, sugiere que hay mucho que investigar y que descubrir. Esta tesis sirve como punto de apoyo y fomenta un marco de investigación muy interesante y con mucho recorrido por delante: el de la predicción de éxito del contenido digital.

5.3. Investigaciones relacionadas

Una vez finalizado este estudio, se han realizado dos investigaciones propuestas como comunicaciones a congresos orientados a la comunicación, las redes sociales, las estrategias digitales y la sociedad de la información. Estas dos investigaciones han aplicado una pequeña parte de la metodología propuesta en esta tesis, pero aplicada a un conjunto de cuentas de Twitter de un sector determinado, con el objetivo de obtener la predicción de una variable estadística de mucho interés para cada sector, así como información descriptiva sobre el uso de los hashtags por parte de la población perteneciente al sector escogido en cada caso.

Se ha demostrado así que la metodología propuesta en esta tesis se puede adaptar en mayor o menor medida a poblaciones diferentes e, incluso, obtener información que permita generalizar conclusiones sobre un conjunto de medios de comunicación.

5.3.1. COMRED 2021, II Congreso Internacional Comunicación y Redes Sociales de la Sociedad de la Información

La primera de las comunicaciones presentadas se ha hecho en el marco de COMRED 2021, II Congreso Internacional Comunicación y Redes Sociales de la Sociedad de la Información, un congreso en el que se explora la influencia, alcance e impacto de las redes sociales en diferentes ámbitos de la sociedad de la información. Celebrado en la Universidad Autónoma de Lisboa el 31 de marzo y 1 de abril de 2021, pretende aunar diferentes niveles y disciplinas que ayuden a explicar las virtudes y los defectos de las redes sociales (EducationLab Consulting S.L., 2021).

Figura 196. Logo del congreso COMRED 2021

Al COMRED 2021 se envió un estudio en el cual se extrajo información estadística en modo de datos de conteo del uso de los hashtags en Twitter por parte de una muestra de 96 medios nativos digitales hispanos, de los cuales 48 son españoles y 48 son portugueses. Gracias a esos datos, se realizó un análisis estadístico con la hipótesis de que sería posible predecir el número de retuits 2 semanas después, desde el punto de vista de los medios nativos digitales españoles, los medios nativos digitales portugueses, los usuarios de Twitter españoles y los usuarios de Twitter portugueses.

Se realizaron regresiones lineales simples, debida a la presencia de correlaciones muy fuertes (cercana a uno) entre una variable inicial y la variable objetivo, y una regresión binomial negativa en el caso en que no ocurrió dicho fenómeno. Así, se pudo confirmar la hipótesis en tres de los cuatro conjuntos anteriormente mencionados, no siendo posible predecir el número de retuits en el caso de los usuarios de Twitter en español.

El estudio también aportó información sobre el uso de los hashtags en sí, cuáles han sido los más utilizados y los más exitosos en forma de retuits, así como la ratio de inclusión de una URL en los tuits por parte de los diferentes conjuntos.

5.3.2. CIMED 2021, I Congreso Internacional de Museos y Estrategias Digitales

La segunda de las comunicaciones se presentó en el marco de CIMED 2021, I Congreso Internacional de Museos y Estrategias Digitales, el cual nace de la necesidad del diseño y la integración de nuevas estrategias de comunicación en los museos e instituciones culturales. De esta manera, se abre un espacio para el intercambio de información y la discusión de las ventajas e inconvenientes de las tecnologías digitales cuando son aplicadas en los museos (REMED, 2021).

CIMED
I CONGRESO INTERNACIONAL DE
MUSEOS Y ESTRATEGIAS DIGITALES

Figura 197. Logo del congreso CIMED 2021

En este caso, se envió un estudio en el cual se obtuvo información en modo de datos de conteo sobre el uso de los hashtags en Twitter por parte de todos los museos con presencia en Twitter que pertenecen a la Red de Museos y Estrategias Digitales (REMED), entidad organizadora del congreso CIMED. La hipótesis propuesta fue la de obtener la predicción del número de favoritos de los hashtags cinco semanas después para el conjunto de los museos y el de los usuarios de Twitter en español.

Así pues, se detectó una correlación muy fuerte (cercana a uno) en el caso del conjunto de los museos, mientras que en el otro se usó la regresión binomial negativa. La hipótesis se confirmó en el caso de los favoritos de los hashtags desde el punto de vista de los museos, pero se rechazó en el conjunto de los usuarios de Twitter en español.

Asimismo, el estudio también ha incluido información descriptiva sobre el uso de los hashtags por parte de los museos pertenecientes a REMED, como es el caso de los hashtags más utilizados en cada conjunto, los hashtags que más favoritos obtuvieron y la detección de una diferencia significativa en la inclusión de una URL en los tuits procesados de ambos conjuntos.

5.4. Limitaciones y futuras líneas de investigación

A continuación, se van a describir posibles líneas de investigación que podrían enriquecer las conclusiones de la metodología propuesta en esta tesis, abordando cuestiones que responden a limitaciones metodológicas, nuevos ámbitos de estudio o indicadores que no se han tenido presentes en el estudio actual.

Se trata ciertamente de un tema que aborda cuestiones muy actuales que están en constante evolución. Esta evolución puede provocar cambios en el concepto de éxito de un contenido digital, los procesos mediante los cuales se comparte este y las características inherentes de la red social analizada. De manera transversal, también pueden cambiar las necesidades de los medios digitales en cuanto a su modelo de negocio, su modelo editorial, su estrategia de contenidos, la introducción de nuevas vías de competencia... Por otro lado, el análisis de tendencias en el consumo de contenidos abre las vías de investigación de otros campos que se han mencionado en alguna ocasión en el apartado 2 del estado de la cuestión, como la política, la bolsa, la psicología, la cultura, las necesidades sociológicas en el uso de las redes sociales... e incluso temáticas inherentes al periodismo digital que son muy cuestionadas en la actualidad, como la saturación de información, la desinformación, el *framing* o teoría del encuadre en la comunicación, etc. Todo lo anterior supone una muestra de que este es un terreno de investigación con múltiples facetas y muchísimo por explorar y descubrir.

En cuanto a la metodología propuesta, sería muy interesante abordar otros ámbitos de estudio con los que enriquecer la idea de éxito comprendido en los escenarios propuestos. Estos ámbitos de estudio ayudarían a segmentar el tráfico según los canales de procedencia: el SEO, el tráfico referido gracias a enlaces orgánicos en otros dominios e incluso el tráfico social en otras redes sociales además de Twitter.

En un escenario en el que se trate de obtener mayores ingresos en publicidad digital, se debe aproximar siempre a las fuentes de ingresos del caso de uso a analizar. En el presente estudio se han incorporado, por tanto, métricas relacionadas con anuncios web de Google Adsense, pero sería interesante incorporar anuncios de varios sistemas de anuncios y comparar cuáles producen más ingresos. Por otra parte, también se podrían realizar estudios futuros incluyendo otras fuentes digitales de ingresos, como anuncios o enlaces patrocinados en newsletters, enlaces patrocinados en artículos de la web, artículos con contenido patrocinado sobre una temática en concreto, vídeos cuyas visualizaciones generen ingresos, etc.

Las propias características de los anuncios se podrían investigar también si la modalidad del sistema de anuncios lo permite. En el caso de Google AdSense, se podrían utilizar bloques de anuncios personalizados e investigar con características como su localización y tamaño. Se han advertido valores anómalos en los ingresos estimados por cada mil impresiones (eCPM) en algunos artículos, por lo que sería conveniente estudiar qué contenidos se relacionan con anuncios cuyos clicks se pagan más.

Además, sería interesante incorporar al estudio la posibilidad de invertir dinero en promocionar contenidos, ya sea de forma online u offline. Por ejemplo, se podría hacer un experimento para comprobar qué contenidos tienen más éxito en este tipo de estrategias e incluso comprobar si la metodología para predecir el éxito orgánico obtiene las mismas conclusiones que el éxito de tráfico de pago. Esto es, que los contenidos con más posibilidades de tener éxito también tendrían más éxito que otros si se les ofreciera un presupuesto extra en forma de publicidad.

Otras variables interesantes para estudiar son las relacionadas con las características de la navegación del usuario, como la duración de la visita y las páginas vistas por sesión. Hay muchos elementos que podrían afectar y que podrían ayudar a mejorar la calidad de las sesiones: la presencia de elementos incrustados, las diferentes tipologías de esos contenidos incrustados (vídeos, publicaciones de redes sociales, reproductores de podcast, etc), el número de palabras, la presencia de enlaces internos en el contenido, etc.

La precisión obtenida en las regresiones que se han realizado en esta tesis ha sido, en algunos casos, pequeña, por lo que una línea de investigación interesante sería tratar de averiguar nuevas variables que aporten más información, lo cual haría más probable una mayor precisión en las predicciones.

En esta tesis se ha elegido Twitter como red social a estudiar, pero en la actualidad también hay presente un uso intensivo de otras redes sociales para el consumo de noticias de última hora u otros tipos de contenido digital. Es el caso de Facebook e Instagram, esta última especialmente interesante si se tiene en cuenta la capacidad de poner enlaces en las historias de 24h cuando se dispone de 10.000 seguidores o más. También es el caso de las redes sociales que estén directamente relacionadas con las temáticas del medio digital a analizar, así como las redes sociales que surjan en un futuro y que puedan aportar información valiosa de cara a predecir el éxito de las publicaciones.

El rango temporal en el que se efectúa el estudio cibermétrico también podría aportar conclusiones interesantes si este englobara un año completo de la redacción de un medio digital. De esta manera se podría tener en cuenta la repercusión que tiene el conjunto de

festividades del público objetivo en su consumo de contenidos digitales. Incluso se podría realizar este estudio durante varios años para comprobar que esas variaciones se repiten y no son provocadas por características inherentes al año en concreto, ya que como se ha comentado en el apartado 0 sobre limitaciones metodológicas, cada época puede presentar características que modifiquen el comportamiento de todo el público en general (una guerra, una pandemia...).

Otro estudio podría plantearse desde el punto de vista del instante de tiempo en el que se publica el artículo. Se podría investigar cuál es el mejor día y hora para publicar e incluso compartir un artículo, mediante el análisis del tráfico en el medio digital, el análisis de tendencias en las redes sociales y un proceso de experimentación. Esto ayudaría en la toma de decisiones al equipo editorial en una labor que probablemente ejecutan a diario. También se podría plantear ese estudio o incluso el de la presente tesis como una comparativa entre los artículos publicados entre semana y los que se publican en fin de semana, para averiguar si las conclusiones son diferentes debido a las diferencias inherentes de ambos periodos.

La metodología depende en todo momento de la detección de contenidos nuevos de manera manual, ya sea por fuentes fidedignas u originales. Se podría aumentar el alcance de este estudio añadiendo, como paso previo, la detección automática de tendencias. De esta manera se capturarían nuevos contenidos de manera automática para ser evaluados con la metodología actual , y asistir en la toma de decisiones del equipo editorial.

En esta ocasión se ha decidido estudiar la tipología de artículos de noticias de última hora. Sería muy interesante estudiar también otras tipologías de artículos, como las reseñas, los reportajes o las entrevistas. Si bien también pueden estar influenciadas por las tendencias de cada momento, su alcance suele ser mayor en el tiempo o incluso afectar a otro medio de información como puede ser el podcasting o radio online.

Otra línea de investigación interesante es el análisis del tráfico entre artículos. Las noticias de última hora suelen incluir enlaces a otras noticias relacionadas e incluso noticias que son de la misma sección, pero sin ninguna vinculación por términos concretos. El estudio del tráfico referido a nivel interno de estas noticias de última hora puede aportar información valiosa sobre las consecuencias indirectas de esta tipología de artículos. Incluso se podría comparar con otras tipologías para observar cuáles ayudan a impulsar este tipo de tráfico. Sería un estudio de interés, ya sea por su relación con la fidelización de la audiencia, al mantenerlos más tiempo en la web navegando entre páginas, como por sus beneficios positivos para el posicionamiento en buscadores (SEO).

Cada artículo, además, presenta una serie de términos relacionados cuya tendencia se analiza en la metodología. Esa serie de términos es seleccionada de manera manual por parte del redactor, por lo que se descartan así términos que en principio podrían ser considerados genéricos o no relacionados directamente con el contenido del artículo. El análisis de tendencias de estos términos, que podríamos catalogar de secundarios, también podría enriquecer el contexto del contenido. Se podría realizar, por tanto, una investigación que tenga en cuenta todos los términos vinculados a los contenidos de manera ponderada, para estudiar así su influencia contextual. También se podría ponderar la aparición de varios términos en un mismo tuit, dándole una mayor importancia a los datos de éstos con respecto a los que solo tengan un término en su contenido.

El análisis de tendencias se ha centrado en esta ocasión en dos momentos puntuales: el de la publicación del artículo y 14 días después. Sin embargo, otra línea de investigación podría dedicarse al análisis de las tendencias con el paso del tiempo y averiguar qué parámetros pueden servir para detectar aquellas que se mantengan un mayor periodo de tiempo en activo, incrementando así las posibilidades de elegir el contenido con mayores beneficios a largo plazo. Por otro lado, también se podría incorporar a la metodología actual el seguimiento de tendencias sobre las cuales se va a publicar contenido, de manera que se puedan tener también datos de 14 días antes de la publicación de un artículo, y comprobar si su fuerza aumenta con el surgimiento de novedades como noticias de última hora.

Se ha advertido además que las tendencias con un mayor número total de retuits suelen provenir de tendencias con un número inicialmente muy alto, aunque en ocasiones son tendencias con un gran crecimiento en los 14 días estudiados en esta tesis. Sería interesante estudiar qué características tienen estas tendencias y a qué se debe ese crecimiento.

El análisis de tendencias también podría dedicar su atención a detectar y tratar de predecir el momento de pico máximo de una tendencia, para así realizar la publicación del artículo en el medio digital o su compartición en la red social analizada, aprovechando el instante álgido de atención al respecto por parte de la audiencia. Vinculado a esto último, también se podría estudiar la influencia de la hora de publicación y la hora de compartición en redes sociales en el éxito de un contenido, y experimentar con ello para averiguar el instante de tiempo óptimo en el que realizar ambas acciones, ya sea por el histórico general del medio, el porcentaje de audiencia conectado en cada momento o el análisis de tendencias previamente mencionado.

Por otro lado, la utilización de API externas ha provocado una serie de limitaciones en el estudio, ya que las funcionalidades están sujetas a las cuotas y los límites impuestos por los desarrolladores de las API. Lo que en un principio es un inconveniente, también supone la ventaja de que las investigaciones se tendrán que adecuar a la evolución de las API y podrán aprovecharlas para adquirir datos más avanzados o completos. Por tanto, se podrían realizar investigaciones futuras con versiones más avanzadas o incluso de pago de las API de Google Analytics y Twitter, con características y cuotas diferentes. En el caso concreto de Twitter, está actualmente preparando la versión 2 de su API que constará de muchos cambios, entre los cuales se encuentra el acceso a indicadores como el número de respuestas (reply_count) y el número de citas (quote_count), ahora mismo no disponibles en la versión estándar.

Por último, en esta tesis se aporta una metodología cibermétrica para obtener ecuaciones de predicción de variables que se consideren éxito para un medio de comunicación digital en concreto. Por tanto, las conclusiones que se obtienen tienen un componente muy práctico y dirigido al caso de uso en el que se aplican. Si se deseara generalizar alguna de las conclusiones, ya sea sobre las características de algunas variables, la relación entre unas y otras, la predicción del éxito, el estudio de los valores anómalos… sería necesario aplicar la metodología a una muestra suficiente de diferentes medios de comunicación, incluso pudiéndose abordar esa muestra según tipologías (periódicos, blogs, blogs de tiendas online…) o sectores industriales concretos. El propósito sería encontrar similitudes que den lugar a conclusiones generales que puedan servir para enmarcar y explicar con más detalle el consumo de información en internet.

Con todo ello, el modelo que se propone en esta tesis presenta múltiples facetas que pueden ser extendidas y exploradas, ya que se trata de un campo de investigación con muchas posibilidades debido tanto a su actualidad como a su diversidad. Presenta una gran amplitud de miras e intereses, ya sea en el ámbito del periodismo digital, la comunidad científica y el sector privado, por lo que se pueden integrar otras disciplinas académicas y realizar experimentos realmente interesantes.

Si Bill Gates aseguró en 1996 que el contenido era el rey (Bailey, 2010), en una actualidad cada vez más digital y automatizada, el contenido se ha convertido en un pilar fundamental de la realidad diaria de la sociedad, con muchísimo potencial por descubrir.

Bibliografía

Álvarez Intriago, V., Agreda Fernández, L. & Cevallos Gamboa, A., 2016. Análisis de la estrategia de marketing digital mediante herramientas de analítica web. *INVESTIGATIO,* Issue 7, pp. 81-97.

Abel, F., Gao, Q., Houben, G.-J. & Tao, K., 2011. *Analyzing User Modeling on Twitter for Personalized News Recommendations.* Berlin, Heidelberg, Springer, pp. 1-12.

Acevedo, J., 2019. *¿Cómo se construyen las noticias? La teoría del 'framing' tiene la respuesta.* [En línea]
Available at: https://www.caninomag.es/se-construyen-las-noticias-la-teoria-del-framing-la-respuesta/
[Último acceso: 17 junio 2020].

Aggarwal, C. C., Wolf, J. L. & Yu, P. S., 1998. *A framework for the optimizing of WWW advertising.* New York, Springer, pp. 1-10.

Aguillo, I. F., 2000. A new generation of tools for search, recovery and quality evaluation of World Wide Web medical resources. *Online Information Review,* 24(2), pp. 138-143.

Aguillo, I. F., 2012. La necesaria evolución de la cibermetría. *Anuario ThinkEPI,* Volumen 6, pp. 119-122.

Ahmad, A. N., 2010. Is Twitter a useful tool for journalists?. *Journal of Media Practice,* 11(2), pp. 145-155.

Aiello, L. M. y otros, 2013. Sensing trending topics in twitter. *IEEE Transactions on Multimedia,* 15(6), pp. 1268-1282.

Aimiuwu, E., Bapna, S., Ahmed, A. & Aimiuwu, E. E., 2012. *How Web Search and Social Media Affect Google AdSense Performance.* s.l., s.n.

Albero-Gabriel, G., 2014. Twitter, #primaveravalenciana y generación de noticias. *Cuadernos de Información y Comunicación,* Volumen 19, pp. 253-269.

Albertos, J. L. M., 2001. El mensaje periodístico en la prensa digital. *Estudios sobre el Mensaje Periodístico,* Volumen 7, pp. 19-32.

Albishry, N., Crick, T., Tryfonas, T. & Fagade, T., 2018. *An Evaluation of Performance and Competition in Customer Services on Twitter: A UK Telecoms Case Study.* Lyon, ACM, pp. 1713-1720.

Albornoz, L. A., 2006. Características de los diarios online. En: *Periodismo Digital. Los Grandes Diarios En La Red.* Buenos Aires: Ediciones La Cujía, pp. 201-256.

Almind, T. C. & Ingwersen, P., 1997. Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation,* 53(4), pp. 404-426.

Alonso Berrocal, J. L., Figuerola, C. G. & Zazo Rodríguez, Á. F., 2001. Cibermetría del Web: Las leyes de exponenciación. *Revista general de información y documentación,* 11(1), pp. 201-209.

Alonso Berrocal, J. L., García Figuerola Paniagua, C. & Zazo Rodríguez, Á. F., 2008. Recuperación de información Web: 10 años de cibermetría. *Ibersid: revista de sistemas de información y documentación,* Volumen 2, pp. 69-78.

Al-Rawi, A., 2019. Viral News on Social Media. *Digital Journalism,* 7(1), pp. 63-79.

Alvanaki, F., Sebastian, M., Ramamritham, K. & Weikum, G., 2011. *EnBlogue: emergent topic detection in web 2.0 streams.* New York, ACM Press, pp. 1271-1274.

Alvarez Intriago, V., Agreda Fernández, L. & Cevallos Gamboa, A., 2016. Análisis de la estrategia de marketing digital mediante herramientas de analítica web. *Investigatio,* 7(7), pp. 81-97.

Amat Rodrigo, J., 2017. *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE.* [En línea] Available at: https://www.cienciadedatos.net/documentos/35_principal_component_analysis [Último acceso: 20 diciembre 2020].

Amiri, A. & Menon, S., 2006. Scheduling web banner advertisements with multiple display frequencies. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans,* 36(2), pp. 245-251.

An, D., 2006. A content analysis of multinational advertisers' localisation strategy in web advertising. *International Journal of Internet Marketing and Advertising,* 3(2), pp. 120-141.

Anderson, C. W., 2011. Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism: Theory, Practice & Criticism,* 12(5), pp. 550-566.

Anderson, D., 2005. Pop-up Ads are No Longer as Popular with Marketers. *Brandweek,* 46(5), p. 13.

Anderson, P., 2007. What is Web 2.0? Ideas, technologies and implications for education. *JISC. Technology and Standards Watch.*

Anger, I. & Kittl, C., 2011. *Measuring influence on Twitter.* New York, ACM Press, pp. 1-4.

Anon., 2005. *Gráfico de Caja y Bigotes.* [En línea]
Available at: http://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Grafico%20de%20Caja%20y%20Bigotes.pdf
[Último acceso: 7 enero 2021].

Anon., 2006. *Regresión Múltiple.* [En línea]
Available at: https://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Regresion%20Multiple.pdf
[Último acceso: 7 enero 2021].

Anon., 2007. *Componentes Principales.* [En línea]
Available at: http://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Componentes%20Principales.pdf
[Último acceso: 7 enero 2021].

Anon., 2007. *Gráfico Normal de Probabilidad.* [En línea]
Available at: http://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Grafico%20de%20Probabilidad%20Normal.pdf
[Último acceso: 7 enero 2021].

Anon., 2011. *Características.* [En línea]
Available at: https://statgraphics.net/caracteristicas/
[Último acceso: 7 enero 2021].

Anon., 2013. *What Are the Effects of Multicollinearity and When Can I Ignore Them?.* [En línea]
Available at: https://blog.minitab.com/blog/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them
[Último acceso: 20 diciembre 2020].

Anon., 2015. *No todo es normal. Manejo de datos no normales.* [En línea] Available at: https://anestesiar.org/2015/no-todo-es-normal-manejo-de-datos-no-normales/
[Último acceso: 20 diciembre 2020].

Anon., 2017. En: *Multiple Regression Analysis.* s.l.:SAGE Publications, Inc., pp. 157-205.

Anon., 2020. *Qué es Google Trends - Definición, significado y ejemplos.* [En línea] Available at: https://www.arimetrics.com/glosario-digital/google-trends
[Último acceso: 8 enero 2021].

Anupam, V. y otros, 1999. On the security of pay-per-click and other Web advertising schemes. *Computer Networks,* 31(11), pp. 1091-1100.

Ardèvol-Abreu, A., 2015. Framing o teoría del encuadre en comunicación. Orígenes, desarrollo y panorama actual en España. *Revista Latina de Comunicación Social,* Volumen 70, pp. 423-450.

Ardon, S. y otros, 2013. *Spatio-temporal and events based analysis of topic popularity in twitter.* New York, ACM Press, pp. 219-228.

Armañanzas, E., 1996. La cultura, una parcela para periodistas especializados. *ZER: Revista de Estudios de Comunicación.*

Armentia Caminos, o. I., Caminos, J. M., Elexgaray, J. & Merchán, I., 2000. La información en la prensa digital: redacción, diseño y hábitos de lectura. *ZER,* 5(8).

Arrabal-Sánchez, G. & De-Aguilera-Moyano, M., 2016. Communicating in 140 Characters. How Journalists in Spain use Twitter. *Comunicar,* 24(46), pp. 9-17.

Arrese, Á., 2013. Algunas reconquistas pendientes del periodismo. *Revista de Comunicación,* 12(2010), pp. 197-219.

Arroyo Vázquez, N. & Pareja Pérez, V. M., 2003. Metodología para la obtención de datos con fines cibermétricos.

Arroyo Valencia, M. E. & Ceron Mayor, J. C., 2009. Estudio de los conceptos de la publicidad online aplicados al sitio web entretenete. com en cali en el año 2008.

Asur, S., Huberman, B. A., Szabo, G. & Wang, C., 2011. Trends in Social Media: Persistence and Decay. *SSRN Electronic Journal.*

Atefeh, F. & Khreich, W., 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence,* 31(1), pp. 132-164.

Awichanirost, J. & Phumchusri, N., 2020. *Analyzing The Effects of Sessions on Unique Visitors and Unique Page Views with Google Analytics: A case study of a Tourism Website in Thailand.* s.l., IEEE.

Azam, N., Jahiruddin, Abulaish, M. & Haldar, N. A.-H., 2015. *Twitter Data Mining for Events Classification and Analysis.* s.l., IEEE, pp. 79-83.

Bailey, C., 2010. *Content Is King by Bill Gates.* [En línea] Available at: https://www.craigbailey.net/content-is-king-by-bill-gates/ [Último acceso: 10 junio 2019].

Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J., 2011. *Everyone's an Influencer: Quantifying Influence on Twitter.* Hong Kong, ACM, pp. 65-74.

Balusamy, B., Krishna P, V. & Sridhar, J., 2016. Web Analytics: Assessing the Quality of Websites Using Web Analytics Metrics. En: *Design Solutions for Improving Website Quality and Effectiveness.* s.l.:s.n., pp. 253-275.

Barbieri, F. & Saggion, H., 2015. *Modelling Irony in Twitter.* Gothenburg, Association for Computational Linguistics (ACL), pp. 56-64.

Bardoel, J., 1996. Beyond Journalism: A Profession between Information Society and Civil Society. En: *Communication Theory and Research.* London, Thousand Oaks, New Delhi: SAGE Publications Ltd, pp. 178-190.

Bardoel, J. & Deuze, M., 2001. "Network journalism": Converging competencies of old and new media. *Australian Journalism Review,* 23(2), p. 91.

Barnard, S. R., 2012. *Twitter and the journalistic field: how the growth of a new (s) medium is transforming journalism,* s.l.: s.n.

Barnston, A. G., 1992. *Correspondence among the Correlation [root mean square error] and Heidke Verification Measures; Refinement of the Heidke Score,* Washington, D. C.: Notes and Correspondance.

Barón López, F. J. & Téllez Montiel, F., s.f. *Regresión múltiple.* [En línea] Available at: https://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf [Último acceso: 18 diciembre 2020].

Bastos, M. T., 2015. Shares, Pins, and Tweets: News readership from daily papers to social media. *Journalism Studies,* 16(3), pp. 305-325.

Belair-Gagnon, V. & Holton, A. E., 2018. Boundary Work, Interloper Media, And Analytics In Newsrooms: An analysis of the roles of web analytics companies in news production. *Digital Journalism,* 6(4), pp. 492-508.

Benevenuto, F., Magno, G., Rodrigues, T. & Almeida, V., 2010. *Detecting Spammers on Twitter.* Redmond, s.n.

Benway, J. P., 1998. *Banner Blindness: The Irony of Attention Grabbing on the World Wide Web.* s.l., Human Factors and Ergonomics Society, Inc., pp. 463-467.

Berger, J. & Milkman, K. L., 2012. What Makes Online Content Viral?. *Journal of Marketing Research,* 49(2), pp. 192-205.

Berrocal, J. L., Figuerola, C. G., Zazo, A. F. & Rodríguez, E., 2002. La cibermetría en la recuperación de información en el Web.

Berthon, P., Pitt, L. F. & Watson, R. T., 1996. The World Wide Web as an advertising medium: Toward an understanding of conversion efficiency. *Journal of Advertising Research,* 36(1), pp. 43-54.

Bierhoff, J., Deuze, M. & de Vreese, C., 2000. *Media innovation, professional debate and media training. A European analysis,* Maastricht: European Journalism Centre.

Bifet, A. & Frank, E., 2010. *Sentiment Knowledge Discovery in Twitter Streaming Data.* Springer, Berlin, Heidelberg, s.n., pp. 1-15.

Björneborn, L. & Ingwersen, P., 2001. Perspectives of webometrics. *Scientometrics,* 50(1), pp. 65-82.

Björneborn, L. & Ingwersen, P., 2004. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology,* 55(14), pp. 1216-1227.

Blanchett Neheli, N., 2018. News by Numbers: The evolution of analytics in journalism. *Digital Journalism,* 6(8), pp. 1041-1051.

Blumler, J. G., 2010. FOREWORD, the two-legged crisis of journalism. *Journalism Studies,* 11(4), pp. 439-441.

Bobbitt, Z., 2019. *A Gentle Introduction to Poisson Regression for Count Data.* [En línea] Available at: https://www.statology.org/poisson-regression/ [Último acceso: 20 diciembre 2020].

Bollen, J., Pepe, A. & Mao, H., 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.

Boone, G., Secci, J. & Gallant, L., 2010. Emerging Trends in Online Advertising. *Doxa Comunicación. Revista interdisciplinar de estudios de comunicación y ciencias sociales,* Volumen 5, p. 241.

Bornmann, L., 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics,* 8(4), pp. 895-903.

Boyd, D., Golder, S. & Lotan, G., 2010. *Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.* Hawaii, IEEE, pp. 1-10.

Brain, M., 2002. How Web Advertising Works. *Retrieved December,* Volumen 1.

Briggs, R. & Hollis, N., 1997. Advertising on the Web: is there response before click-through?. *Journal of Advertising Research,* 37(2), pp. 33-46.

Bright, J., 2016. The Social News Gap: How News Reading and News Sharing Diverge. *Journal of Communication,* 66(3), pp. 343-365.

Broersma, M. & Graham, T., 2013. Twitter as a news source. *Journalism Practice,* 7(4), pp. 446-464.

Brolin, P., Svedström, A. & Monstad, T., 2017. *Web Analytics and Online Journalism.* Tampere, s.n., pp. 17-19.

Bruner, R. E. & Gluck, M., 2006. Best Practices for Optimizing Web Advertising Effectiveness. *DoubleClick Inc. White Paper.*

Bruns, A., 2012. How long is a tweet? Mapping Dynamic Conversation Networks on Twitter using Gawk and Gephi. *Information, Communication & Society,* 15(9), pp. 1323-1351.

Bruns, A., 2012. Journalists and Twitter: How Australian News Organisations Adapt to a New Medium. *Media International Australia,* 144(1), pp. 97-107.

Bruns, A. & Stieglitz, S., 2012. Quantitative Approaches to Comparing Communication Patterns on Twitter. *Journal of technology in human services,* 30(3-4), pp. 160-185.

Bygrave, L. A., 2001. Automated profiling - Minding the machine: Article 15 of the EC data protection directive and automated profiling. *Computer Law and Security Report,* 17(1), pp. 17-24.

Cai, H., Yang, Y., Li, X. & Huang, Z., 2015. *What are Popular: Exploring Twitter Features for Event Detection, Tracking and Visualization.* New York, ACM Press, pp. 89-98.

Caminos Marcel, J. M., Marín Murillo, F. & Armentia Vizuete, J. I., 2006. Las audiencias ante los cambios en el ciberperiodismo. *Revista Latina de Comunicación Social,* 9(61).

Campos, R. M. & Soares, C. G., 2016. Comparison and assessment of three wave hindcasts in the North Atlantic Ocean. *Journal of Operational Oceanography,* 9(1), pp. 26-44.

Canavilhas, J., 2007. *Webnoticia propuesta de modelo periodístico para la WWW.* Covilhã: Livros LabCom.

Cao, J., 1999. *Evaluation of advertising effectiveness using agent-based modeling and simulation.* Bristol, s.n.

Carlson, M., 2018. Confronting Measurable Journalism. *Digital Journalism,* 6(4), pp. 406-417.

Carrillo, M. V. & Castillo, A., 2005. La Nueva Publicidad Digital (NPD): Servicios Digitales y Contenidos Interactivos que Generen 'Experiencias' en los Consumidores. *Razón y Palabra,* 10(45).

Castelló Martínez, A., Del Pino Romero, C. & Ramos Soler, I., 2014. Twitter como canal de comunicación corporativa y publicitaria. *Comunicación y Sociedad,* 27(2), pp. 21-54.

Castillo, C., Mendoza, M. & Poblete, B., 2011. *Information Credibility on Twitter.* Hyderabad, ACM, pp. 675-684.

Cataldi, M., Caro, L. D. & Schifanella, C., 2010. *Emerging topic detection on twitter based on temporal and social terms evaluation.* Washington, D.C., ACM, pp. 1-10.

Cebrián Herreros, M., 2009. Comunicación interactiva en los cibermedios. *Revista Comunicar,* 17(33).

Chaffey, D. & Patron, M., 2012. From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics. *Journal of Direct, Data and Digital Marketing Practice,* 14(1), pp. 30-45.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K. P., 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *Icwsm,* 10(10-17), p. 30.

Chandramouli, B., Goldstein, J. & Duan, S., 2012. *Temporal Analytics on Big Data for Web Advertising.* s.l., IEEE, pp. 90-101.

Chatterjee, P., 2001. Beyond CPMs and Clickthroughs: Understanding Consumer Interaction with Web Advertising. En: *Internet Marketing Research: Theory and Practice.* Hershey: Ook Lee, pp. 209-216.

Chatterjee, P., Hoffman, D. L., Novak, T. P. & Graduate, O., 2003. Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science,* 22(5), pp. 520-541.

Chen, J. & Stallaert, J., 2014. An economic analysis of online advertising using behavioral targeting. *Mis Quarterly,* 38(2), pp. 429-450.

Chen, K.-S. & Chen, M.-H., 2010. *EC 2.0: Can You Get Profit by Writing Blog? An Empirical Study in Google Adsense.* s.l., IEEE, pp. 1-7.

Chinea, J. D., 2008. *Ajustes de datos: transformación de datos,* s.l.: s.n.

Clerwall, C., 2014. Enter the Robot Journalist. *Journalism Practice,* 8(5), pp. 519-531.

Coleman, R., 2016. Web analytics 101: How to use statistics to drive online engagement to your institutional page or research project. *Impact of Social Sciences Blog.*

Comm, J., 2006. *The AdSense Code: What Google Never Told You about Making Money with Adsense.* Garden City, New York: Morgan James Publishing.

Congosto, M. L., Deltell, L., Claes, F. & Osteso, J. M., 2013. Análisis de la audiencia social por medio de Twitter Caso de estudio: los premios Goya 2013. *Revista de comunicación y tecnologías emergentes,* 11(2), pp. 53-82.

Costas, R., 2017. Hacia los estudios de medios sociales de la ciencia: las métricas de los medios sociales, presente y futuro. *Bibliotecas. Anales de Investigación,* 13(1), pp. 1-5.

Costera Meijer, I. & Groot Kormelink, T., 2015. Checking, Sharing, Clicking and Linking. *Digital Journalism,* 3(5), pp. 664-679.

Coviello, L. y otros, 2014. Detecting Emotional Contagion in Massive Social Networks. *PLoS ONE,* 9(3).

Crawford, K., 2009. Following you: Disciplines of listening in social media. *Continuum,* 23(4), pp. 525-535.

Cristian, M., Morphl, P. & Morphl, A. A., 2019. Predicting next shopping stage using Google Analytics data for E-commerce applications.

Crossfield, J., 2016. *Is Your Social Media Content as Popular as You Think?.* [En línea] Available at: https://contentmarketinginstitute.com/2016/03/social-media-content-popular/ [Último acceso: 21 mayo 2019].

Cui, A. y otros, 2012. *Discover breaking events with popular hashtags in twitter.* New York, ACM Press, pp. 1794-1798.

Dörr, K. N., 2016. Mapping the field of Algorithmic Journalism. *Digital Journalism,* 4(6), pp. 700-722.

Dahlgren, P., 1996. Media Logic in Cyberspace: Repositioning Journalism and its Publics. *Javnost - The Public,* 3(3), pp. 59-72.

D'Andrea, E., Ducange, P., Lazzerini, B. & Marcelloni, F., 2015. Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems,* 16(4), pp. 2269-2283.

Danman Fugl, L., 2001. Fundamental methodologies and tools for the employment of webometric analyses: a discussion and proposal for improving the foundation of webometrics. *Royal School of Library and Information Science.*

Darn, S., 2010. Twitter content classification. *First Monday,* 15(12).

Davidov, D., Tsur, O. & Rappoport, A., 2010. *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys.* s.l., s.n., pp. 241-249.

Davis, D., 2020. *Twitter Hashtags - Here's Everything You Need To Know.* [En línea] Available at: https://socialplanner.io/blog/twitter-hashtags/ [Último acceso: 11 mayo 2021].

Davis, H., 2006. *Google Advertising Tools.* Sebastopol: O'Reilly Media, Inc..

Dean, B., 2021. *How Many People Use Twitter in 2021? [New Twitter Stats].* [En línea] Available at: https://backlinko.com/twitter-users [Último acceso: 7 mayo 2021].

Deller, R., 2011. Twittering on: Audience research and participation using Twitter. *Participations,* 8(1).

Desnica, M., Sayadchi, M., Umair, K. & Szabó, P., 2014. *Google AdSense. User Modeling and Recommender Systems − Case exercise,* s.l.: s.n.

Deuze, M., 1998. The WebCommunicators: Issues in research into online journalism and journalists. *First Monday,* 3(12).

Deuze, M., 1999. Journalism and the web: An Analysis of Skills and Standards in an Online Environment. *Gazette,* 61(5), pp. 373-390.

Deuze, M., 2001. Online journalism: Modelling the first generation of news media on the World Wide Web. *First Monday,* 6(10).

Deuze, M., 2017. Considering a possible future for Digital Journalism. *Revista Mediterránea de Comunicación,* 8(1).

Diaz Noci, J., 2008. Definición teórica de las características del ciberperiodismo: elementos de la comunicación digital. *Doxa Comunicación. Revista interdisciplinar de estudios de comunicación y ciencias sociales,* Issue 6, pp. 53-91.

Díaz-Campo, J. & Segado-Boj, F., 2015. Journalism ethics in a digital environment: How journalistic codes of ethics have been adapted to the Internet and ICTs in countries around the world. *Telematics and Informatics,* Volumen 32, pp. 735-744.

Dickerson, J., 2008. Don't Fear Twitter. *Nieman Reports,* 63(2), pp. 5-6.

Dimitrova, D. V. & Neznanski, M., 2006. Online Journalism and the War in Cyberspace: A Comparison Between U.S. and International Newspapers. *Journal of Computer-Mediated Communication,* 12(1), pp. 248-263.

Dinh Le, T. & Ho Nguyen, B.-T., 2014. Attitudes toward mobile advertising: A study of mobile web display and mobile app display advertising. *Asian Academy of Management Journal,* 19(2), pp. 87-103.

Drivas, I. C., Sakas, D. P., Giannakopoulos, G. A. & Kyriaki-Manessi, D., 2020. Big data analytics for search engine optimization. *Big Data and Cognitive Computing,* 4(5), pp. 1-22.

Dunteman, G. H., 1989. *Principal Components Analysis.* Newbury Park(California): SAGE Publications, Inc..

Eddy, T., Cork, B. C., Lebel, K. & Hickey, E. H., 2021. Examining Engagement With Sport Sponsor Activations on Twitter. 14(1), pp. 79-108.

Edmonds, A., White, R. W., Morris, D. & Drucker, S. M., 2007. Instrumenting the dynamic web. *Journal of Web Engineering,* 6(3), pp. 244-260.

Edo, C., 2006. Cyberjournalism and The Scientific Bases of Newswriting. *Brazilian Journalism Research,* 2(2), pp. 115-132.

EducationLab Consulting S.L., 2021. *COMRED, II Congreso Internacional Comunicación y Redes Sociales de la Sociedad de la Información.* [En línea] Available at: https://comred.autonoma.pt/es/
[Último acceso: 17 febrero 2021].

e-intelligent.es, 2014. *Medir el éxito de nuestro Marketing de Contenidos.* [En línea] Available at: https://www.e-intelligent.es/es/blog/medir-el-exito-de-nuestro-marketing-de-contenidos
[Último acceso: 21 mayo 2019].

Ekanem, A., 2016. *How to Profit from Google AdSense.* s.l.:s.n.

Elías Pérez, C., 2009. La "cultura convergente" y la filosofía Web 2.0 en la reformulación de la comunicación científica en la era del ciberperiodismo. *Arbor : Ciencia, Pensamiento y Cultura,* Volumen 737, pp. 623-634.

Eren Erdogmus, I. & Cicek, M., 2012. The impact of social media marketing on brand loyalty. *Procedia-Social and Behavioral Sciences,* Volumen 58, pp. 1353-1360.

Fainholc, B., 2011. Un análisis contemporáneo del Twitter. *RED,* Volumen 26.

Fan, W. & Gordon, M. D., 2014. The power of social media analytics. *Communications of the ACM,* 57(6), pp. 74-81.

Farhan, M. & Yousaf, A., 2016. Exploring Factors Affecting the Effectiveness of Web-Advertising. *Indian Journal of Marketing,* 46(8), pp. 51-57.

Farhi, P., 2007. Salvation? The embattled newspaper business is betting heavily on Web advertising revenue to secure its survival. But that wager is hardly a sure thing. *American Journalism Review,* 29(6), pp. 18-24.

Farías de Estany, J. & María Prieto, C., 2009. Ciberperiodismo: hacia un modelo de producción de contenidos en el ciberespacio. *Quórum Académico,* 6(1), pp. 11-37.

Feixa, C. & Fernández-Planells, A., 2014. Generación @ versus Generación #. La juventud en la era hiperdigital. En: *Audiencias juveniles y cultura digital.* Barcelona: Institut de la Comunicació (InCom-UAB), pp. 35-54.

Fenton, N., 2009. News in the digital area. En: S. Allan, ed. *The Routledge Companion to News and Journalism.* Londres: Routledge, pp. 601-611.

Ferrara, E. & Yang, Z., 2015. Measuring Emotional Contagion in Social Media. *PLoS One,* 10(11).

Ferrara, E. & Yang, Z., 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science,* 1(9).

Ferrer-Conill, R. & Tandoc, E. C., 2018. The Audience-Oriented Editor: Making sense of the audience in the newsroom. *Digital Journalism,* 6(4), pp. 436-453.

Flew, T., Spurgeon, C., Daniel, A. & Swift, A., 2012. The Promise of Computational Journalism. *Journalism Practice,* 6(2), pp. 157-171.

Fox, D., Smith, A., Chaparro, B. S. & Shaikh, A. D., 2009. *Optimizing Presentation of AdSense Ads within Blogs.* Los Angeles, SAGE Publications Ltd, pp. 1267-1271.

Fox, S. & Lenhart, A., 2009. *Twitter and status updating,* Washington, D.C.: Pew Internet & American Life Project.

Fredin, E. S. & David, P., 1998. Browsing and the hypermedia interaction cycle: A model of self-efficacy and goal dynamics. *Journalism and Mass Communication Quaterly,* 75(1), pp. 35-54.

Fu, C. S. & Wu, W. Y., 2010. The means-end cognitions of web advertising: A cross-cultural comparison. *Online Information Review,* 34(5), pp. 686-703.

Gabarron, E. y otros, 2020. Factors Engaging Users of Diabetes Social Media Channels on Facebook, Twitter, and Instagram: Observational Study. *Journal of Medical Internet Research,* 22(9).

Gaglio, S., Lo Re, G. & Morana, M., 2016. A framework for real-time Twitter data analysis. *Computer Communications,* Volumen 73, pp. 236-242.

Gallagher, K., Foster, K. D. & Parsons, J., 2001. The medium is not the message: Advertising effectiveness and content evaluation in print and on the web. *Journal of Advertising Research,* 41(4), pp. 57-70.

Ghose, A., Han, S. P. & Park, S., 2013. An Empirical Analysis of Digital Advertising Analyzing the Interdependence between Web and Mobile Advertising.

Ghosh, S. y otros, 2012. *Understanding and Combating Link Farming in the Twitter Social Network.* Lyon, ACM, pp. 61-70.

Gilani, Z. y otros, 2017. *An in-depth characterisation of Bots and Humans on Twitter,* s.l.: s.n.

Glen, S., 2020. *RMSE: Root Mean Square Error.* [En línea] Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
[Último acceso: 24 diciembre 2020].

Go, A., Bhayani, R. & Huang, L., 2009. Twitter Sentiment Classification using Distant Supervision. *CS224N project report,* 1(12).

Goldsmith, R. E. & Lafferty, B. A., 2002. Consumer response to Web sites and their influence on advertising effectiveness. *Internet Research,* 12(4), pp. 318-328.

González Mina, J., 2004. *Repensar el periodismo. Transformaciones y emergencias del periodismo actual,* Cali: Programa Editorial, Universidad del Valle.

González Rodríguez, B. J. y otros, 2013. *Muestreo y estimación,* La Laguna: Universidad de La Laguna.

González-Ibáñez, R., Muresan, S. & Wacholder, N., 2011. *Identifying Sarcasm in Twitter: A Closer Look.* Portland, Omnipress, Inc., pp. 581-586.

Google AdSense, s.f. *Acerca de los anuncios automáticos.* [En línea] Available at: https://support.google.com/adsense/answer/9261805?hl=es#auto_ads
[Último acceso: 27 octubre 2020].

Google APIs, s.f. *googleapis/google-api-php-client: A PHP client library for accessing Google APIs.* [En línea] Available at: https://github.com/googleapis/google-api-php-client
[Último acceso: 30 junio 2020].

Google Cloud, 2021. *Bibliotecas cliente de Cloud.* [En línea] Available at: https://cloud.google.com/apis/docs/cloud-client-libraries
[Último acceso: 03 septiembre 2021].

Google Developers, s.f. *API de informes de Google Analytics.* [En línea] Available at: https://developers.google.com/analytics/devguides/reporting/core/v4/rest [Último acceso: 6 agosto 2020].

Google Developers, s.f. *Dimensions & Metrics Explorer — Google Analytics Demos & Tools.* [En línea] Available at: https://ga-dev-tools.appspot.com/dimensions-metrics-explorer/ [Último acceso: 29 junio 2020].

Google Developers, s.f. *Google Ad Manager reporting.* [En línea] Available at: https://support.google.com/analytics/answer/6183313?hl=en [Último acceso: 22 junio 2020].

Google Developers, s.f. *Límites y cuotas en las solicitudes a API.* [En línea] Available at: https://developers.google.com/analytics/devguides/reporting/core/v3/limits-quotas [Último acceso: 29 junio 2020].

Google Developers, s.f. *Más información sobre Google Analytics.* [En línea] Available at: https://developers.google.com/analytics/devguides/platform [Último acceso: 29 junio 2020].

Google Developers, s.f. *Método: reports.batchGet.* [En línea] Available at: https://developers.google.com/analytics/devguides/reporting/core/v4/rest/v4/reports/batchGet [Último acceso: 6 agosto 2020].

Google Developers, s.f. *Parámetros de consulta estándar.* [En línea] Available at: https://developers.google.com/analytics/devguides/reporting/core/v4/parameters [Último acceso: 6 agosto 2020].

Google Developers, s.f. *Respuestas de error.* [En línea] Available at: https://developers.google.com/analytics/devguides/config/mgmt/v3/errors [Último acceso: 15 julio 2020].

Grace-Martin, K., 2008. *Assessing the Fit of Regression Models.* [En línea] Available at: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/ [Último acceso: 6 enero 2021].

Graefe, A., 2016. Guide to Automated Journalism. *Tow Center for Digital Journalism.*

Graham, S., 2014. *Indicadores Clave de Éxito para medir canales digitales KPI.* [En línea] Available at: https://es.slideshare.net/sallygrahams/kp-is-como-optimizar-la-efectividad-de-los-canales-digitales
[Último acceso: 4 febrero 2020].

Gratt, M., 2012. *Competitive Content Intelligence: Learning from the Competition.* [En línea] Available at: https://www.buzzstream.com/blog/competitive-content-intelligence-learning-from-the-competition.html
[Último acceso: 21 mayo 2019].

Groselj, D., 2014. A webometric analysis of online health information: Sponsorship, platform type and link structures. *Online Information Review,* 38(2), pp. 209-231.

Grueskin, B., Seave, A. & Graves, L., 2011. *The Story So Far: What We Know About the Business of Digital Journalism.* New York: Columbia University Press.

Guallar, J., 2015. Prensa digital en 2013-2014. *El profesional de la información,* Volumen 9, pp. 153-160.

Guille, A. & Favre, C., 2014. *Mention-anomaly-based Event Detection and Tracking in Twitter.* s.l., IEEE, pp. 375-382.

Guille, A. & Hacid, H., 2012. *A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks.* Lyon, ACM, pp. 1145-1152.

Guillory, J. y otros, 2020. Using Social Media to Conduct Outreach and Recruitment for Expanded Newborn Screening. *Frontiers in Communication,* Volumen 5, p. 21.

Gómez Nieto, B., 2016. Análisis de la Publicidad digital en los sitios web españoles de mayor audiencia. *Razón y Palabra,* Volumen 93, pp. 374-396.

Gómez, L. & Martínez, S., 2011. 'Twitter no es una red social, es una red de información abierta'. *El Mundo*, 26 febrero.

Gupta, A., Kumaraguru, P., Castillo, C. & Meier, P., 2014. *TweetCred: Real-Time Credibility Assessment of Content on Twitter.* Cham, Springer, pp. 228-243.

Gupta, P. y otros, 2013. *WTF: The Who to Follow Service at Twitter.* Rio de Janeiro, ACM, pp. 505-514.

Gutiérrez Argüello, C., 2013. *Modelos de Comunicación y Negocios en Periódicos Digitales.* [En línea]
Available at: https://es.slideshare.net/CarlosGutArg/modelos-de-comunicacin-y-negocios-en-peridicos-digitales
[Último acceso: 6 mayo 2019].

Haak, B. V. D., Parks, M. & Castells, M., 2012. The future of journalism: Networked journalism. *International Journal of Communication,* Volumen 6.

Haile, T., 2014. *Chartbeat CEO Tony Haile: What You Get Wrong about the Internet.* [En línea]
Available at: https://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/
[Último acceso: 21 mayo 2019].

Hall, S., Hobson, D., Lowe, A. & Willis, P., 1980. Encoding/Decoding. *Culture, Media, Language,* pp. 128-138.

Hamidian, S. & Diab, M., 2019. Rumor Detection and Classification for Twitter Data.

Hamilton, A., 2007. Why Everyone's Talking about Twitter. *Time*, 27 marzo.

Hannon, J., Bennett, M. & Smyth, B., 2010. *Recommending twitter users to follow using content and collaborative filtering approaches.* New York, ACM Press, pp. 199-206.

Hansen, L. K. y otros, 2011. *Good Friends, Bad News Affect and Virality in Twitter.* s.l., s.n., pp. 34-43.

Hanusch, F., 2017. Web analytics and the functional differentiation of journalism cultures: Individual, organizational and platform-specific influences on newswork. *Information Communication and Society,* 20(10), pp. 1571-1586.

Hanusch, F. & Tandoc, E. C., 2019. Comments, analytics, and social media: The impact of audience feedback on journalists' market orientation. *Journalism,* 20(6), pp. 695-713.

Harinarayana, N. S., 2015. Webometrics, Cybermetrics and Nettometrics.

Harper, C., 1998. *And That's the Way It Will Be: News and Information in a Digital World.* Nueva York y Londres: New York University Press.

Hasan, M., Orgun, M. A. & Schwitter, R., 2016. *TwitterNews+: A Framework for Real Time Event Detection from the Twitter Data Stream.* Cham, Springer, pp. 224-239.

Haustein, S., 2019. *Scholarly Twitter metrics.* s.l., s.n., pp. 729-760.

Hedman, U. & Djerf-Pierre, M., 2013. THE SOCIAL JOURNALIST: Embracing the social media life or creating a new digital divide?. *Digital Journalism,* 1(3), pp. 368-385.

Heikkilä, H. & Ahva, L., 2015. The relevance of journalism: Studying news audiences in a digital era. *Journalism Practice,* 9(1), pp. 50-64.

Herbert, J., 1999. *Journalism in the Digital Age: Theory and practice for broadcast, print and online media.* Nueva York: Routledge.

Hermida, A., 2010. From TV to Twitter: How Ambient News Became Ambient Journalism. *Media/Culture Journal,* 13(2).

Hermida, A., 2010. Twittering the news. *Journalism Practice,* 4(3), pp. 297-308.

Hermida, A., 2013. #Journalism: Reconfiguring journalism research about Twitter, one tweet at a time. *Digital Journalism,* 1(3), pp. 295-313.

Hermida, A., 2016. Social Media and the News. En: *The SAGE Handbook of Digital Journalism.* London: SAGE Publications Ltd, pp. 81-94.

Hernandez, R. K. & Rue, J., 2016. *The principles of multimedia journalism: Packaging digital news.* New York: Routledge.

Herrera, S. & Requejo, J. L., 2012. 10 Good Practices for News Organizations Using Twitter. *Journal of Applied Journalism & Media Studies,* 1(1), pp. 79-95.

Herrero Gutiérrez, F. J., López Ornelas, M. & Álvarez Nobell, A., 2012. Análisis cibermétrico de cinco revistas emergentes de Comunicación en sus dos primeros años en línea. *Index. comunicación: Revista científica en el ámbito de la Comunicación Aplicada,* 2(1), pp. 69-90.

Hicks Maynard, N., 2000. *Mega Media: How Market Forces are Transforming News.* Nueva York: Maynard Partners Incorporated.

Hochuli, D., 2015. *Search vs. Social Media: How Audience 'Intent' Can Affect Content Marketing Performance.* [En línea]
Available at: https://contentmarketinginstitute.com/2015/11/search-social-content-performance/
[Último acceso: 21 mayo 2019].

Hofacker, C. F. & Murphy, J., 2000. Clickable world wide web banner ads and content sites. *Journal of Interactive Marketing,* 14(1), pp. 49-59.

Hoffman, D. L. & Novak, T. P., 2000. Advertising Pricing Models for the World Wide Web. *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property,* Volumen 5, pp. 1-22.

Holmström, J. y otros, 2019. Do we Read what we Share? Analyzing the Click Dynamic of News Articles Shared on Twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* pp. 420-425.

Holton, A. & Lewis, S. C., 2011. Journalists, social media, and the use of humor on Twitter. *Electronic Journal of Communication,* Volumen 21, pp. 1-22.

Honeycutt, C. & Herring, S. C., 2009. *Beyond Microblogging: Conversation and Collaboration via Twitter.* Hawaii, IEEE.

Hong, L., Dan, O. & Davison, B. D., 2011. *Predicting Popular Messages in Twitter.* Hyderabad, ACM, pp. 57-58.

Huang, J. y otros, 2020. Predicting the active period of popularity evolution: A case study on Twitter hashtags. *Information Sciences,* Volumen 512, pp. 315-326.

Huang, J., Thornton, K. M. & Efthimiadis, E. N., 2010. *Conversational Tagging in Twitter.* Toronto, HT'10, pp. 173-178.

Huberman, B. A., Romero, D. M. & Wu, F., 2009. Social networks that matter: Twitter under the microscope. *First Monday,* 14(1).

Hussain, R., Ferdous, A. S. & Mort, G. S., 2018. Impact of web banner advertising frequency on attitude. *Asia Pacific Journal of Marketing and Logistics,* 30(2), pp. 380-399.

Hu, Y. y otros, 2017. Predicting Key Events in the Popularity Evolution of Online Information. *PLoS One,* 12(1).

IAB, 2019. *IAB internet advertising revenue report. 2018 full year results.* [En línea] Available at: https://www.iab.com/wp-content/uploads/2019/05/Full-Year-2018-IAB-Internet-Advertising-Revenue-Report.pdf
[Último acceso: 20 octubre 2020].

Ibrahim, F., 2011. Cyber Journalism: Bridging the Gap between Professionalism and Epistemology. *Journal of Media and Information Warfare,* 4(4).

Ingwersen, P., 1998. The Calculation of Web Impact Factors. *Journal of Documentation,* 54(2), pp. 236-243.

Ingwersen, P., 2009. Scientometric and Webometric Methods. *Document, Information and Knowledge,* pp. 1-11.

Ingwersen, P. & Björneborn, L., 2004. Methodological issues of webometric studies. En: *Handbook of Quantitative Science and Technology Research.* s.l.:Kluwer Academic Publishers, pp. 339-369.

Internet World Stats, 2021. *World Internet Users Statistics and 2021 World Population Stats.* [En línea]
Available at: https://internetworldstats.com/stats.htm
[Último acceso: 10 mayo 2021].

Jackson, A. M. y otros, 2018. #CDCGrandRounds and #VitalSigns: A Twitter Analysis. *Annals of Global Health,* 84(4), p. 710.

Jana, S. & Chatterjee, S., 2004. Quantifying Web-site visits using Web statistics: An extended cybermetrics study. *Online Information Review,* 28(3), pp. 191-199.

Java, A., Song, X., Finin, T. & Tseng, B., 2009. *Why We twitter: An analysis of a microblogging community.* Springer, Berlin, Heidelberg, s.n., pp. 118-138.

Johnson, E., 2016. Most students get breaking news first from Twitter. *Article in Newspaper Research Journal,* 37(2), pp. 153-166.

Ju, A., Jeong, S. H. & Chyi, H. I., 2014. Will Social Media Save Newspapers?. *Journalism Practice,* 8(1), pp. 1-17.

Ka Po, N., 2006. Factors Affecting Attitude toward Web Advertising.

Kalsnes, B. & Larsson, A. O., 2018. Understanding News Sharing Across Social Media. *Journalism Studies,* 19(11), pp. 1669-1688.

Kaushik, A., 2006. *Excellent Analytics Tip #2: Segment Absolutely Everything.* [En línea]
Available at: https://www.kaushik.net/avinash/web-analytics-segments-three-category-recommendations/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2006. *Experimentation and Testing: A Primer.* [En línea]
Available at: https://www.kaushik.net/avinash/experimentation-and-testing-a-primer/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2007. *Emetrics SFO Reflections: Deliberate, Dig, Understand, Throw A Feast!.* [En línea]
Available at: https://www.kaushik.net/avinash/emetrics-sfo-reflections-deliberate-dig-understand-throw-a-feast/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2007. *Excellent Analytics Tip #11: Measure Effectiveness Of Your Web Pages.* [En línea]
Available at: https://www.kaushik.net/avinash/excellent-analytics-tip-11-measure-effectiveness-of-your-web-pages/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2007. *I Got No Ecommerce. How Do I Measure Success?.* [En línea]
Available at: https://www.kaushik.net/avinash/i-got-no-ecommerce-how-do-i-measure-success/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2007. *Kick Butt With Internal Site Search Analytics.* [En línea]
Available at: https://www.kaushik.net/avinash/kick-butt-with-internal-site-search-analytics/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2009. *Social Media Analytics: Twitter: Quantitative & Qualitative Metrics.* [En línea]
Available at: https://www.kaushik.net/avinash/social-media-analytics-twitter-quantitative-qualitative-analysis/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2010. *3 Awesome, Downloadable, Custom Web Analytics Reports.* [En línea]
Available at: https://www.kaushik.net/avinash/best-downloadable-custom-web-analytics-reports/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2010. *Web Analytics Segments: 3 Key Category Recommendations.* [En línea]
Available at: https://www.kaushik.net/avinash/web-analytics-segments-three-category-recommendations/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2011. *Best Social Media Metrics: Conversation, Amplification, Applause, Economic Value.* [En línea]
Available at: https://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/
[Último acceso: 21 mayo 2019].

Kaushik, A., 2011. *Best Web Metrics / KPIs for a Small, Medium or Large Sized Business.* [En línea]
Available at: https://www.kaushik.net/avinash/best-web-metrics-kpis-small-medium-large-business/
[Último acceso: 5 febrero 2020].

Kaushik, A., 2011. *Digital Marketing and Measurement Model: Web Analytics.* [En línea]
Available at: https://www.kaushik.net/avinash/digital-marketing-and-measurement-model/
[Último acceso: 4 febrero 2020].

Kaushik, A., 2011. *Identify Website Goal Values & Win: Excellent Analytics Tips # 19.* [En línea]
Available at: https://www.kaushik.net/avinash/web-analytics-tips-identify-website-goal-values/
[Último acceso: 5 febrero 2020].

Kaushik, A., 2011. *Your Web Metrics: Super Lame or Super Awesome?.* [En línea]
Available at: https://www.kaushik.net/avinash/web-metrics-super-lame-super-awesome/
[Último acceso: 5 febrero 2020].

Kaushik, A., 2014. *Benchmarking Analytics Performance: The Options, Dos, Don'ts.* [En línea]
Available at: https://www.kaushik.net/avinash/benchmarking-digital-analytics-performance-metrics/
[Último acceso: 21 mayo 2019].

Kawamoto, K., 2003. *Digital Journalism: Emerging Media and the Changing Horizons of Journalism.* Lanham: Rowman & Littlefield Publishers, Inc..

Kaye, J. & Quinn, S., 2010. *Funding journalism in the digital age : business models, strategies, issues and trends.* New York: Peter Lang Publishing, Inc..

Kazienko, P. & Adamski, M., 2004. *Personalized Web Advertising Method.* Berlin, Heidelberg, Springer, pp. 146-155.

Kazienko, P. & Adamski, M., 2007. AdROSA - Adaptive personalization of web advertising. *Information Sciences,* 177(11), pp. 2269-2295.

Kempe, D., Kleinberg, J. & Tardos, E., 2015. Maximizing the Spread of Influence through a Social Network. *Theory of Computing,* 11(1), pp. 105-147.

Kiani, G. R., 1998. Marketing opportunities in the digital world. *Internet Research,* 8(2), pp. 185-194.

Kierzkowski, A., McQuade, S., Waitman, R. & Zeisser, M., 1996. Marketing to the digital consumer. *The McKinsey Quarterly,* Volumen 3.

King, A. B., 2008. *Website Optimization.* Sebastopol: O'Reilly Media, Inc..

King, A. B., 2008. *Website Optimization Editor.* Sebastopol: O'Reilly Media, Inc..

Kirchhoff, S. M., 2009. Advertising Industry in the Digital Age Congressional Research Service. *Congressional Research Service.*

Kong, S. y otros, 2019. Web advertisement effectiveness evaluation: Attention and memory. *Journal of Vacation Marketing,* 25(1), pp. 130-146.

Kontopoulos, E., Berberidis, C., Dergiades, T. & Bassiliades, N., 2013. Ontology-based sentiment analysis of twitter posts. *Expert Systems With Applications,* Volumen 40, pp. 4065-4074.

Korgaonkar, P., Silverblatt, R. & O'Leary, B., 2001. Web advertising and Hispanics. *Journal of Consumer Marketing,* 18(2), pp. 134-150.

Kossinets, G., Kleinberg, J. & Watts, D., 2008. *The Structure of Information Pathways in a Social Communication Network.* s.l., s.n., pp. 435-443.

Kouloumpis, E., Wilson, T. & Moore, J., 2011. *Twitter Sentiment Analysis: The Good the Bad and the OMG!.* Edinburgh, s.n.

Krishnamurthy, B., Gill, P. & Arlitt, M., 2008. *A Few Chirps About Twitter.* Seattle, WOSN'08, pp. 19-24.

Kumar, S., 2016. *Evolution of Web Advertising.* Cham, Springer, pp. 1-7.

Kumar, S., Liu, H., Mehta, S. & Subramaniam, L. V., 2014. From Tweets to Events: Exploring a Scalable Solution for Twitter Streams.

Kumar, S., Morstatter, F. & Liu, H., 2014. Analyzing Twitter Data. En: *Twitter Data Analytics.* New York: Springer, pp. 35-48.

Kumar, S. & Sethi, S. P., 2009. Dynamic pricing and advertising for web content providers. *European Journal of Operational Research,* 197(3), pp. 924-944.

Kurniawan, D. A., Wibirama, S. & Setiawan, N. A., 2016. *Real-time traffic classification with Twitter data mining.* Yogyakarta, IEEE, pp. 1-5.

Kursuncu, U. y otros, 2019. *Predictive Analysis on Twitter: Techniques and Applications.* Cham, Springer, pp. 67-104.

Kwak, H., Lee, C., Park, H. & Moon, S., 2010. *What is Twitter, a Social Network or a News Media?.* Raleigh, International World Wide Web Conference Committee (IW3C2), pp. 591-600.

Labrecque, L. I., Swani, K. & Stephen, A. T., 2020. The impact of pronoun choices on consumer engagement actions: Exploring top global brands' social media communications. *Psychology & Marketing,* 37(6), pp. 796-814.

Lamberti, F., Paravati, G., Gatteschi, V. & Cannavo, A., 2017. Supporting Web Analytics by Aggregating User Interaction Data From Heterogeneous Devices Using Viewport-DOM-Based Heat Maps. *IEEE Transactions on Industrial Informatics,* 13(4), pp. 1989-1999.

Langheinrich, M. y otros, 1999. Unintrusive customization techniques for Web advertising. *Computer Networks,* 31(11), pp. 1259-1272.

Larson, R. R., 1996. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *Proceedings of the Annual Meeting-American Society for Information Science,* Volumen 33, pp. 71-78.

Lasorsa, D. L., Lewis, S. C. & Holton, A. E., 2010. Normalizing Twitter Journalism Practice in an Emerging Communication Space. *Journalism Studies,* 13(1), pp. 19-36.

Lavinsky, D., 2013. *Stop SWOT'ing The Small Stuff.* [En línea] Available at: https://www.forbes.com/sites/davelavinsky/2013/03/20/stop-swoting-the-small-stuff/
[Último acceso: 6 febrero 2020].

Lee, A. M., Lewis, S. C. & Powers, M., 2014. Audience Clicks and News Placement. *Communication Research,* 41(4), pp. 505-530.

Lee, K. y otros, 2011. *Twitter Trending Topic Classification.* Vancouver, IEEE, pp. 251-258.

Lee, S. Y., 2010. Do Web Users Care About Banner Ads Anymore? The Effects of Frequency and Clutter in Web Advertising. *Journal of Promotion Management,* 16(3), pp. 288-302.

Lehmann, J., Castillo, C., Lalmas, M. & Zuckerman, E., 2013. *Finding News Curators in Twitter.* Rio de Janeiro, ACM, pp. 863-870.

Lei, R. M., 2000. An assessment of the World Wide Web as an advertising medium. *The Social Science Journal,* 37(3), pp. 465-471.

Lewandowska (Tomaszewska), A. & Jankowski, J., 2017. The negative impact of visual web advertising content on cognitive process: towards quantitative evaluation. *International Journal of Human Computer Studies,* Volumen 108, pp. 41-49.

Li, J. & Cardie, C., 2014. *Timeline Generation: Tracking individuals on Twitter.* Seoul, International World Wide Web Conference Committee (IW3C2), pp. 643-652.

Li, L., Mei, T., Niu, X. & Ngo, C.-W., 2010. *PageSense: Style-wise Web Page Advertising.* Raleigh, ACM, pp. 1273-1276.

Linden, C. G., 2017. Algorithms for journalism: The future of news work. *The Journal of Media Innovations,* Volumen 4, pp. 60-76.

Li, R., Lei, K. H., Khadiwala, R. & Chang, K. C.-C., 2012. *TEDAS: A Twitter-based Event Detection and Analysis System.* Washington, D.C., IEEE, pp. 1273-1276.

Li, W. y otros, 2010. *Exploitation and Exploration in a Performance based Contextual Advertising System.* Washington, D.C., AMC, pp. 27-36.

Li, X., 1998. Web page design and graphic use of three U.S. Newspapers. *Journalism and Mass Communication Quaterly,* 75(2), pp. 353-365.

Londoño Russi, K., 2013. *Periodismo de entretenimiento en el canal RCN y su discusión con el periodismo rosa difundido por los demás noticieros nacionales.* Pereira: s.n.

Luo, Z., Osborne, M., Tang, J. & Wang, T., 2013. *Who Will Retweet Me? Finding Retweeters in Twitter.* s.l., s.n., pp. 869-872.

López Jiménez, D. & Martínez López, J. F., 2010. Nuevas coordenadas en el ámbito de la web 2.0: el caso de la publicidad comportamental. *Revista de Estudios Económicos y Empresariales,* Volumen 22, pp. 101-134.

López, J. F., 2017. *Economipedia.* [En línea] Available at: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html
[Último acceso: 22 11 2020].

López, J. F., 2017. *Homecedasticidad.* [En línea] Available at: https://economipedia.com/definiciones/homocedasticidad.html
[Último acceso: 18 diciembre 2020].

López, M. J., Continente, X., Sánchez, E. & Bartroli, M., 2017. Intervenciones que incluyen webs y redes sociales: herramientas e indicadores para su evaluación. *Gaceta Sanitaria,* 31(4), pp. 346-348.

Lu, R. & Yang, Q., 2012. Trend Analysis of News Topics on Twitter. *International Journal of Machine Learning and Computing,* 2(3), p. 327.

Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S., 2017. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS one,* 12(9).

Macias, W., 2003. A Preliminary Structural Equation Model of Comprehension and Persuasion of Interactive Advertising Brand Web Sites. *Journal of interactive advertising,* 3(2), pp. 36-48.

Mafra, É., 2020. *Engagement: ¿qué es y para qué sirve en el Marketing?.* [En línea] Available at: https://rockcontent.com/es/blog/que-es-engagement/
[Último acceso: 2 junio 2021].

Mahmoud, A. B., 2014. Linking information motivation to attitudes towards Web advertising. *Journal of Islamic Marketing,* 5(3), pp. 396-413.

Maier, T., 2013. *How much does AdSense pay?.* [En línea] Available at: https://webgilde.com/en/how-much-does-adsense-pay/
[Último acceso: 22 octubre 2020].

Maina, S., 2012. Webometrics and journal websites. *European Science Editing,* 38(3), pp. 65-66.

Malik, M. M. & Pfeffer, J., 2016. A macroscopic analysis of news content in Twitter. *Digital Journalism,* 4(8), pp. 955-979.

Malinsky, R. & Jelínek, I., 2010. *Improvements of Webometrics by Using Sentiment Analysis for Better Accessibility of the Web.* Springer, Berlin, Heidelberg, s.n., pp. 581-586.

Mantilla Díez, K., 2018. *Utilización de técnicas de analítica web e inbound marketing: Caso práctico aplicado a una empresa de ilustraciones,* s.l.: Universidad de Cantabria.

Manzano Zambruno, L. y otros, 2019. *Investigar las redes sociales. Un acercamiento interdisciplinar.* s.l.:Ediciones Egregius.

Marcus, A. y otros, 2011. *Twitinfo: aggregating and visualizing microblogs for event exploration.* New York, ACM Press, pp. 227-236.

Marín Diazaraque, J. M., 2004. *Transformaciones de variables,* s.l.: s.n.

Martín de Antonio, R., 2000. Internet como medio publicitario. *Artes Liberales Serie Quadrivium,* Volumen 18.

Martínez Rodríguez, A., 2006. Indicadores cibermétricos: Nuevas propuestas para medir la información en el entorno digital. *Revista Cubana de Información en Ciencias de la Salud,* 14(4).

Marwick, A. E., 2010. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media Society,* 20(10), pp. 1-20.

Mathews, P., Mitchell, L., Nguyen, G. & Bean, N., 2017. *The Nature and Origin of Heavy Tails in Retweet Activity.* Perth, Australia, International World Wide Web Conference Committee (IW3C2), pp. 1493-1498.

Mathioudakis, M. & Koudas, N., 2010. *Twittermonitor: trend detection over the twitter stream.* New York, ACM Press, pp. 1155-1158.

Ma, Z., Sun, A., Yuan, Q. & Cong, G., 2014. *Tagging Your Tweets: A Probabilistic Modeling of Hashtag Annotation in Twitter.* Shanghai, ACM, pp. 999-1008.

McCandless, M., 1998. Web advertising. *IEEE Intelligent Systems and their Applications,* 13(3), pp. 8-9.

McKiernan, G., 2005. Bibliometrics, Cybermetrics, Informetrics, and Scientometrics Sites and Sources. *Science & Technology Libraries,* 26(2), pp. 107-115.

McMinn, A. J. & Jose, J. M., 2015. *Real-time entity-based event detection for twitter.* Cham, Springer, pp. 65-77.

Mediavilla, J. C. & Abuín Vences, N., 2007. GT-3: Diversificación y modelos de negocio de los medios en la red.

Melnikov, N. & Schönwälder, J., 2010. *Cybermetrics: User Identification through Network Flow Analysis.* Springer, Berlin, Heidelberg, s.n., pp. 167-170.

Mendoza, M., Poblete, B. & Castillo, C., 2010. *Twitter Under Crisis: Can we trust what we RT?.* Washington, D.C., ACM, pp. 71-79.

Menon, S. & Soman, D., 2002. Managing the power of curiosity for effective web advertising strategies. *Journal of Advertising,* 31(3), pp. 1-14.

Messias, J., Schmidt, L., Oliveira, R. & Benevenuto, F., 2013. You followed my bot! Transforming robots into influential users in Twitter. *First Monday,* 18(7).

Meyer, J., 2018. *Poisson or Negative Binomial? Using Count Model Diagnostics to Select a Model.* [En línea]
Available at: https://www.theanalysisfactor.com/poisson-or-negative-binomial-using-count-model-diagnostics-to-select-a-model/
[Último acceso: 21 enero 2021].

Milenta, T. & Pestano, J., 2009. *El tratamiento de las industrias culturales emergentes en la prensa española: el caso de los videojuegos.* La Laguna, s.n.

Millán, V., 2021. *Google y el fin de las cookies: ¿más privacidad o más monopolio?.* [En línea]
Available at: https://hipertextual.com/2021/03/google-fin-cookies
[Último acceso: 27 mayo 2021].

Miroglio, B., Zeber, D., Kaye, J. & Weiss, R., 2018. *The Effect of Ad Blocking on User Engagement with the Web.* Lyon, ACM, pp. 813-821.

Mischaud, E., 2007. *Twitter: Expressions of the Whole Self. An investigation into user appropriation of a web-based communications platform,* London: Media@lse, London School of Economics and Political Science ("LSE").

Mitchelstein, E. & Boczkowski, P. J., 2009. Between tradition and change: A review of recent research on online news production. *Journalism: Theory, Practice & Criticism,* 10(5), pp. 562-586.

Mobasher, B., 2005. Web Usage Mining. En: *Encyclopedia of Data Warehousing and Mining.* s.l.:IGI Global, pp. 1216-1220.

Molyneux, L., 2014. What journalists retweet: Opinion, humor, and brand development on Twitter. *Journalism,* 16(7), pp. 920-935.

Montero, F., Lorenzo Romero, C. & Alarcón del Amo, M. d. C., 2010. *Analítica web: pasado, presente y futuro,* Albacete: Universidad de Castilla-La Mancha.

Montiel, M. & Villalobos, F., 2005. La enseñanza del periodismo en el siglo XXI: un desafío entre lo impreso y lo digital. *Telos,* 7(3).

Moore, B., 2021. *Effect of Digital Advertising on Website Traffic at a Kentucky Comprehensive Regional University.* s.l.:s.n.

Morgan, J. S., Shafiq, M. Z. & Lampe, C., 2013. *Is News Sharing on Twitter Ideologically Biased?.* San Antonio, Texas, USA, ACM.

Mozafari, N. & Hamzeh, A., 2015. An enriched social behavioural information diffusion model in social networks. *Journal of Information Science,* 41(3), pp. 273-283.

Mulvenna, M. D., Anand, S. S. & Büchner, A. G., 2000. Personalization on the Net using Web mining: Introduction. *Communications of the ACM,* 43(8), pp. 122-125.

Murthy, D., 2011. Twitter: Microphone for the masses?. *Media, Culture & Society,* 33(5), pp. 779-789.

Myers, S., Zhu, C. & Leskovec, J., 2012. *Information Diffusion and External Influence in Networks.* Beijing, ACM, pp. 33-41.

Navarro Zamora, L., 2000. El periódico on line. *Estudios sobre el Mensaje Periodístico,* Issue 6, pp. 273-287.

Navarro, A. y otros, 2001. Negative binomial distribution versus Poisson in the analysis of recurrent phenomena. *Gaceta sanitaria / S.E.S.P.A.S,* 15(5), pp. 447-452.

Navas, A., 2018. *Modelo de variables de desempeño e impacto en Twitter. Un análisis comunicacional.* Pamplona: s.n.

Nelson, J. L., 2020. The enduring popularity of legacy journalism: An analysis of online audience data. *Media and Communication,* 8(2), pp. 40-50.

Nesterenko, H., 2017. *14 Ways to Do Competitive Marketing Analysis.* [En línea] Available at: https://writtent.com/blog/competitive-marketing-analysis-14-ways-to-monitor-and-beat-your-competitors/
[Último acceso: 21 mayo 2019].

Neuberger, C., Tonnemacher, J., Biebl, M. & Duck, A., 1998. Online-The Future of Newspapers? Germany's Dailies on the World Wide Web. *Journal of Computer-Mediated Communication,* 4(1).

Newman, E. J., Stem, D. E. & Sprott, D. E., 2004. Banner advertisement and Web site congruity effects on consumer Web site perceptions. *Industrial Management and Data Systems,* 104(3), pp. 273-281.

Newman, M. E. J., 2004. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics,* 46(5), pp. 323-351.

Ngai, E. W., 2003. Selection of web sites for online advertising using the AHP. *Information and Management,* 40(4), pp. 233-242.

Noguera, J. M., 2009. La convergencia entre cibermedios y blogosfera: apuntes para entrar de forma natural en la conversación. *Temas de Comunicación,* Volumen 18, pp. 35-151.

Ojeda-Zapata, J., 2008. *Twitter Means Business: How Microblogging Can Help Or Hurt Your Company.* Silicon Valley: s.n.

Okazaki, M. & Matsuo, Y., 2008. *Semantic Twitter: Analyzing tweets for real-time event notification.* Berlin, Heidelberg, Springer, pp. 63-74.

Oliva Rodríguez, J., 2014. Motor de Recomendación e Integración con Ad server.

Orduña-Malea, E., 2013. The web trail – Using cybermetrics to build reputation. *University World News,* Volumen 267.

Orduña-Malea, E., 2019. Google Trends: analítica de búsquedas al servicio del investigador, del profesional y del curioso. *Anuario ThinkEPI,* Volumen 13, pp. 1-14.

Orduña-Malea, E. & Aguillo, I. F., 2015. *Cibermetría: midiendo el espacio red.* Barcelona: Editorial UOC - El Profesional de la Información.

Orduña-Malea, E. & Alonso-Arroyo, A., 2018. *Cybermetric Techniques to Evaluate Organizations Using Web-Based Data.* Cambridge, Kidlington: Chandos Publishing.

Organista Sandoval, J. & Cordero Arroyo, G., 2001. Indicadores cibermétricos para el caso de una revista electrónica de investigación educativa. *Biblioteca Universitaria,* 4(2), pp. 67-76.

Ortega Priego, J. L., 2004. Análisis del consumo de información de una revista electrónica: Análisis de ficheros log de cybermetrics. *Revista española de Documentación Científica,* 27(4), pp. 455-468.

Ortega Priego, J. L. & Aguillo, I. F., 2009. Minería del uso de webs. *El profesional de la información,* 18(1), pp. 20-26.

Ortega, C., 2019. *¿Qué es el coeficiente de correlación de Pearson?.* [En línea] Available at: https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/ [Último acceso: 7 enero 2021].

Osborne, M. & Dredze, M., 2014. *Facebook, Twitter and Google Plus for Breaking News: Is There a Winner?.* s.l., Association for the Advancement of Artificial Intelligence.

Pak, A. & Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc,* 10(2010), pp. 1320-1326.

Pallares, A., 2014. *10 Indicadores de éxito en tus Redes Sociales.* [En línea] Available at: https://smartupmarketing.com/10-indicadores-de-exito-en-tus-redes-sociales/ [Último acceso: 21 mayo 2019].

Parsons, J., Gallagher, K. & Foster, K. D., 2000. *Messages in the Medium: An Experimental Investigation of Web Advertising Effectiveness and Attitudes toward Web Content.* Hawaii, IEEE.

Pavlik, J. V., 2000. The Impact of Technology on Journalism. *Journalism Studies*, 12 diciembre, 1(2), pp. 229-237.

Pavlik, J. V., 2013. Innovation and the Future of Journalism. *Digital Journalism*, 15 enero, pp. 181-193.

Pennacchiotti, M. & Popescu, A.-M., 2011. *A Machine Learning Approach to Twitter User Classification.* s.l., s.n.

Peñafiel Sáiz, C., 2016. Reinvención del periodismo en el ecosistema digital y narrativas transmedia. *AdComunica,* Volumen 12, pp. 163-182.

Peterson, E. T., 2004. *Web Analytics Demystified: A Marketer's Guide to Understanding how Your Web Site Affects Your Business.* s.l.:Celilo Group Media y CafePress.

Phelan, O., McCarthy, K. & Smyth, B., 2009. *Using twitter to recommend real-time topical news.* New York, ACM Press, pp. 385-388.

Phippen, A., Sheppard, L. & Furnell, S., 2004. A practical evaluation of Web analytics. *Internet Research,* 14(4), pp. 284-293.

Picasso, N., 2017. *Regresión Múltiple.* [En línea] Available at: https://blablanegocios.com/regresion-multiple/ [Último acceso: 20 diciembre 2020].

Pradhan, S. S., 2019. Webometric Analysis of Websites of Hindi Newspapers in India. *SRELS Journal of Information Management,* 56(6), p. 306–310.

Priem, J. & Costello, K. L., 2010. How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology,* 47(1), pp. 1-4.

Priem, J., Taraborelli, D., Groth, P. & Neylon, C., 2010. *Altmetrics: A manifesto.* [En línea] Available at: http://altmetrics.org/manifesto/ [Último acceso: 12 abril 2020].

Qi, J., Liang, X., Wang, Y. & Cheng, H., 2018. Discrete time information diffusion in online social networks: micro and macro perspectives. *Scientific Reports,* 8(1), pp. 1-15.

Quandt, T., 2008. News on the world wide web?: A comparative content analysis of online news in europe and the United States. *Journalism Studies,* 9(5), pp. 717-738.

Quinn, S., 2012. *Online Newsgathering: Research and Reporting for Journalism.* New York: Routledge.

Quintero, E., 2008. *Test A/B o Multivariante, qué son y cuál aplicar.* [En línea] Available at: https://www.doctormetrics.com/test-ab-o-multivariante-que-son-y-cual-aplicar/ [Último acceso: 28 mayo 2021].

Rödlund, A., 2020. *Social media and stress: A quantitative study of social media habits and stress in an adult population,* s.l.: s.n.

Rasmussen, T., 1997. Social interaction and the new media: The construction of communicative contexts. *NordicomReview,* 18(2), pp. 63-76.

Raso, A., 2016. *How to Measure Engagement the Right Way.* [En línea] Available at: https://contentmarketinginstitute.com/2016/03/measure-engagement-right/ [Último acceso: 21 mayo 2019].

Rathimala, K. & Marthandan, G., 2010. Exploring hyperlink structure of electronic commerce websites: a Webometric study. *International Journal of Electronic Business,* 8(4-5), pp. 391-404.

Rausell Köster, C., 2004. La publicidad en la web: Hacia el patrocinio y el banner narrativo audiovisual. *Questiones Publicitarias,* 1(9), pp. 103-127.

Rayson, S., 2014. *How To Track Competitor Content Marketing In 5 Steps: A Case Study.* [En línea]
Available at: https://buzzsumo.com/blog/5-steps-track-competitor-content-buzzsumo-case-study/
[Último acceso: 21 mayo 2019].

Recuero de los Santos, P., 2020. *Datos de entrenamiento vs datos de test.* [En línea] Available at: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/ [Último acceso: 22 11 2020].

Recuero de los Santos, P., 2020. *Datos de entrenamiento vs datos de test.* [En línea] Available at: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/ [Último acceso: 3 diciembre 2020].

Regan, T., 1997. The digital journalist. *Nieman Reports,* p. 80.

REMED, 2021. *CIMED 2021, I Congreso Internacional de Museos y Estrategias Digitales.* [En línea]
Available at: https://remed.webs.upv.es/congreso/
[Último acceso: 17 febrero 2021].

Ribeiro-Neto, B., Cristo, M., Golgher, P. B. & Silva De Moura, E., 2005. *Impedance Coupling in Content-targeted Advertising.* Salvador, Brazil, AMC, pp. 496-503.

Rifon, N. J., 2002. Antecedents and Consequences of Web Advertising Credibility: A Study of Consumer Response to Banner Ads. *Journal of interactive Advertising,* 3(1), pp. 12-24.

Riley, P. y otros, 1998. Community or colony: The case of online newspapers and the Web. *Journal of Computer Mediated Communication,* 4(1).

Robinson, H., Wysocka, A. & Hand, C., 2007. Internet advertising effectiveness. *International Journal of Advertising,* 26(4), pp. 527-541.

Rodríguez Martínez, L., 2013. *5 indicadores clave para medir la interacción en las redes sociales.* [En línea]
Available at: https://www.puromarketing.com/42/16086/indicadores-clave-para-medir-interaccion-redes-sociales.html
[Último acceso: 21 mayo 2019].

Rodríguez-Martínez, R., Codina, L. & Pedraza-Jiménez, R., 2012. Indicadores para la evaluación de la calidad en cibermedios: análisis de la interacción y de la adopción de la Web 2.0. *Revista española de Documentación Científica,* 35(1), pp. 61-93.

Rodriguez-Martinez, R. & Pedraza-Jimenez, R., 2009. Prensa digital y Web 2.0. *Hipertext.net,* Volumen 7.

Rogstad, I., 2016. Is Twitter just rehashing? Intermedia agenda setting between Twitter and mainstream media. *Journal of Information Technology & Politics,* 13(2), pp. 142-158.

Rohloff, T., Oldag, S., Renz, J. & Meinel, C., 2019. *Utilizing web analytics in the context of learning analytics for large-scale online learning.* Dubai, IEEE, pp. 296-305.

Roos, J. M., Mela, C. F. & Shachar, R., 2020. The Effect of Links and Excerpts on Internet News Consumption. *Journal of Marketing Research,* 57(3), pp. 395-421.

Rost, A., 2002. *The concept of hypertext in digital journalism.* Barcelona, s.n.

Rowlands, I., 2000. Who can count the dust of Jacob? From bibliometrics to cybermetrics. En: *The Internet: Its impact and Evaluation.* London: Aslib/IMI, pp. 114-130.

Rudat, A. & Buder, J., 2015. Making retweeting social: The influence of content and context information on sharing news in Twitter. *Computers in Human Behavior,* Volumen 46, pp. 74-84.

Rudat, A., Buder, J. & Hesse, F. W., 2014. Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior,* Volumen 35, pp. 132-139.

Ruiz Mitjana, L., 2019. *Prueba de Kolmogórov-Smirnov: qué es y cómo se usa en estadística.* [En línea]
Available at: https://psicologiaymente.com/miscelanea/prueba-kolmogorov-smirnov
[Último acceso: 20 diciembre 2020].

Russell, F. M., Hendricks, M. A., Choi, H. & Stephens, E. C., 2015. Who Sets the News Agenda on Twitter?. *Digital Journalism,* 3(6), pp. 925-943.

Ryan, W., Hazlewood, W. R. & Makice, K., 2008. *Twitterspace: A Co-developed Display using Twitter to Enhance Community Awareness.* s.l., s.n., pp. 230-233.

Sánchez Sánchez, D. A., 2007. El periodismo digital. Una nueva etapa del periodismo moderno. *Revista Lasallista de investigación,* 4(1), pp. 67-73.

Sánchez-González, M. & Palomo-Torres, M.-B., 2014. Conocimiento y valoración del «crowd-funding» en Comunicación: La visión de profesionales y futuros periodistas. *Comunicar,* 22(43), pp. 101-110.

Sánchez-Pita, F. & Alonso-Berrocal, J. L., 2013. Los sitios Web de centros de investigación biosanitaria de Castilla y León. Un análisis cibermétrico. *Revista Latina de Comunicación Social,* Volumen 68, pp. 383-419.

Saadeghvaziri, F., Dehdashti, Z. & Reza Kheyrkhah Askarabad, M., 2013. Web advertising. Assessing beliefs, attitudes, purchase intention and behavioral responses. *Journal of Economic and Administrative Sciences,* 29(2), pp. 99-112.

Saif, H., He, Y. & Alani, H., 2012. *Semantic Sentiment Analysis of Twitter.* Berlin, Heidelberg, Springer, pp. 508-524.

Saif, H., He, Y., Fernandez, M. & Alani, H., 2015. Contextual Semantics for Sentiment Analysis of Twitter. *Information Processing & Management,* 52(1), pp. 5-19.

Salaverría, R., 2005. *Redacción periodística en internet.* Pamplona: Eunsa.

Salaverría, R., 2009. *Los medios de comunicación ante la convergencia digital.* Bilbao, Servicio Editorial de la Universidad del País Vasco, pp. 11-13.

Salminen, J. y otros, 2020. How Does Personification Impact Ad Performance and Empathy? An Experiment with Online Advertising. *International Journal of Human–Computer Interaction,* 37(2), pp. 141-155.

Salwen, M. B., Garrison, B. & Driscoll, P. D., 1995. Online newspaper market size and the use of World Wide Web technologies. En: *Online News and the Public.* New Jersey: Lawrence Erlbaum Associates, pp. 257-276.

San Millán Fernández, E., Medrano García, M. L. & Blanco Jiménez, F. J., 2008. Social media marketing, redes sociales y metaversos. *Universidad, Sociedad y Mercados*

Globales. Asociación Española de Dirección y Economía de la Empresa (AEDEM), pp. 353-366.

Sankaranarayanan, J. y otros, 2009. *TwitterStand: News in Tweets.* Seattle, ACM, pp. 42-51.

Sarlan, A., Nadam, C. & Basri, S., 2014. *Twitter Sentiment Analysis.* Putrajaya, IEEE, pp. 212-216.

Sarzosa Rivera, S. & Medina Chicaiza, P., 2017. Métricas para la difusión de noticias: un acercamiento teórico.

Saura, J. R., Palos-Sánchez, P. & Cerdá Suárez, L. M., 2017. Understanding the Digital Marketing Environment with KPIs and Web Analytics. *Future Internet,* 9(4), p. 76.

Savage, N., 2011. Twitter as medium and message. *Communications of the ACM,* 54(3), pp. 18-20.

Scharnhorst, A. & Wouters, P., 2006. Web indicators: a new generation of S&T indicators?. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics,* Volumen 10.

Schudson, M., 2016. The crisis in news: Can you whistle a happy tune?. En: *The crisis of journalism reconsidered: Democratic culture, professional codes, digital future.* New York: Cambridge Univesity Press, pp. 98-115.

Schultz, T., 1999. Interactive Options in Online Journalism: A Content Analysis of 100 U.S. Newspapers. *Journal of Computer-Mediated Communication,* 5(1).

Schumann, D. W. & Thorson, E., 2010. *Advertising and the World Wide Web.* Malwah: Lawrence Erlbaum Associates.

Schumann, J. H., von Wangenheim, F. & Groene, N., 2014. Targeted Online Advertising: Using Reciprocity Appeals to Increase Acceptance among Users of Free Web Services. *Journal of Marketing,* 78(1), pp. 59-75.

Scott, B., 2005. A Contemporary History of Digital Journalism. *Television & New Media,* 6(1), pp. 89-126.

Shanahan, J. G. & Kurra, G., 2011. *Digital Advertising: An Information Scientist's Perspective.* Berlin, Heidelberg, Springer, pp. 209-237.

Shiri, A., 1998. Cybermetrics: A New Horizon in Information Research. *Cybermetrics,* Volumen 50, p. 111.

Shyam Sundar, S., Narayan, S., Obregon, R. & Uppal, C., 1998. Does web advertising work? Memory for print vs. online media. *Journalism & Mass Communication Quarterly,* 75(4), pp. 822-835.

Sigel, A., Braun, G. & Sena, M., 2008. The impact of banner ad styles on interaction and click-through rates. *Issues in Information Systems,* 9(2), pp. 337-342.

Singer, J. B., 1997. Changes and consistencies Newspaper journalists contemplate online future. *Newspaper Research Journal,* 18(1-2).

Singer, J. B., 1997. Still Guarding the Gate? The Newspaper Journalist's Role in an On-line World. *Convergence,* 3(1).

Small, T. A., 2011. What the hashtag?. *Information, Communication & Society,* 14(6), pp. 872-895.

Smith, A. G., 2006. Google Scholar as a cybermetric tool: a comparison with the New Zealand PBRF research assessment. *Indicators,* Volumen 7.

Social BlaBla, 2020. *Social BlaBla.* [En línea] Available at: https://www.socialblabla.com/como-saber-tu-grado-de-engagement-en-twitter-y-facebook.html
[Último acceso: 9 febrero 2020].

Spivack, N., 2007. *How the WebOS Evolves?.* [En línea] Available at: http://www.novaspivack.com/technology/how-the-webos-evolves
[Último acceso: 6 febrero 2020].

Sprenger, T. O., Sandner, P. G., Tumasjan, A. & Welpe, I. M., 2014. News or Noise? Using Twitter to Identify and Understand Company-specific News Flow. *Journal of Business Finance & Accounting,* 41(7-8), pp. 791-830.

Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.-N., 2000. Web usage mining. *ACM SIGKDD Explorations Newsletter,* 1(2), p. 12.

Stieglitz, S. & Dang-Xuan, L., 2013. Emotions and Information Diffusion in Social Media — Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems,* 29(4), pp. 217-248.

Suber, P., 2006. *Google AdSense ads for open-access journals.* [En línea] Available at: https://dash.harvard.edu/bitstream/handle/1/4391163/suber_adsense.htm?sequence=1 [Último acceso: 29 septiembre 2020].

Suh, B., Hong, L., Pirolli, P. & Chi, E. H., 2010. *Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network.* s.l., IEEE, pp. 177-184.

Sukpanich, N. & Chen, L.-D., 1998. *Measuring the Effectiveness of Web Advertising.* s.l., s.n., p. 124.

Sukunesan, S., Huynh, M. & Sharp, G., 2021. Examining the Pro-Eating Disorders Community on Twitter Via the Hashtag #proana: Statistical Modeling Approach. *JMIR Ment Health,* 8(7).

Tandoc, E. C., 2014. Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society,* 16(4), pp. 559-575.

Tandoc, E. C. & Thomas, R. J., 2015. The Ethics of Web Analytics. *Digital Journalism,* 3(2), pp. 243-258.

Tankovska, H., 2021. *U.S. Twitter usage reasons 2019.* [En línea] Available at: https://www.statista.com/statistics/276393/reasons-for-us-users-to-follow-brands-on-twitter/ [Último acceso: 11 mayo 2021].

Tascón Gabella, M., 2018. *Analítica web y mediciones de audiencia: el caso de los medios nativos digitales en España (Trabajo Fin de Grado en Periodismo).* s.l.:s.n.

Teevan, J., Ramage, D. & Morris, M. R., 2011. *#TwitterSearch: A Comparison of Microblog Search and Web Search.* Hong Kong, ACM, pp. 35-44.

Thelwall, M., 2012. A history of webometrics. *Bulletin of the American Society for Information Science and Technology,* 38(6), pp. 18-23.

Thelwall, M., Buckley, K. & Paltoglou, G., 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology,* 62(2), pp. 406-418.

Thelwall, M. & Cugelman, B., 2017. Monitoring Twitter strategies to discover resonating topics: The case of the UNDP. *El Profesional de la Información,* 26(4).

Thelwall, M. & Vaughan, L., 2004. Webometrics: An introduction to the special issue. *Journal of the American Society for Information Science and Technology,* 55(14), pp. 1213-1215.

Thomas, K., Grier, C. & Paxson, V., 2011. *Suspended Accounts in Retrospect: An Analysis of Twitter Spam.* Berlin, ACM, pp. 243-258.

Thurman, N., 2013. Newspaper Consumption in the Digital Age. *Digital Journalism,* 2(2), pp. 156-178.

Tomlin, J. A., 2000. Entropy approach to unintrusive targeted advertising on the Web. *Computer Networks,* 33(1), pp. 767-774.

Tremayne, M., 2004. The Web of Context: Applying Network Theory to the Use of Hyperlinks in Journalism on the Web. *Journalism & Mass Communication Quarterly,* 81(2), pp. 237-253.

Trillo Domínguez, M., 2008. Análisis cibermétrico de la prensa digital española: ránking de calidad web y mapa de influencia mediática.

Tsugawa, S. & Ohsaki, H., 2017. On the relation between message sentiment and its virality on social media. *Social Network Analysis and Mining,* 7(1), p. 19.

Twitter Developers, s.f. *Authentication Overview.* [En línea] Available at: https://developer.twitter.com/en/docs/basics/authentication/overview [Último acceso: 29 junio 2020].

Twitter Developers, s.f. *Building search queries.* [En línea] Available at: https://developer.twitter.com/en/docs/labs/recent-search/guides/search-queries
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Entities object.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/entities-object
[Último acceso: 7 agosto 2020].

Twitter Developers, s.f. *GET followers/ids.* [En línea] Available at: https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *GET statuses/user_timeline.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Introduction to Tweet JSON.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json
[Último acceso: 7 agosto 2020].

Twitter Developers, s.f. *Metrics.* [En línea] Available at: https://developer.twitter.com/en/docs/twitter-api/metrics
[Último acceso: 17 septiembre 2020].

Twitter Developers, s.f. *Migrating to the New Twitter API v2.* [En línea] Available at: https://developer.twitter.com/en/docs/twitter-api/migrate
[Último acceso: 17 septiembre 2020].

Twitter Developers, s.f. *Overview.* [En línea] Available at: https://developer.twitter.com/en/docs/basics/authentication/overview
[Último acceso: 7 agosto 2020].

Twitter Developers, s.f. *POST statuses/filter.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Pricing.* [En línea] Available at: https://developer.twitter.com/en/pricing
[Último acceso: 23 junio 2020].

Twitter Developers, s.f. *Rate limiting.* [En línea] Available at: https://developer.twitter.com/en/docs/basics/rate-limiting
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Response codes.* [En línea] Available at: https://developer.twitter.com/en/docs/basics/response-codes
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Search pagination.* [En línea] Available at: https://developer.twitter.com/en/docs/labs/recent-search/guides/pagination [Último acceso: 7 agosto 2020].

Twitter Developers, s.f. *Standard search.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/search/overview/standard [Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Standard search API.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Things every developer should know.* [En línea] Available at: https://developer.twitter.com/en/docs/basics/things-every-developer-should-know
[Último acceso: 29 junio 2020].

Twitter Developers, s.f. *Tools and libraries.* [En línea] Available at: https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries [Último acceso: 03 septiembre 2021].

Twitter Developers, s.f. *Tweet object.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *User object.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object
[Último acceso: 30 junio 2020].

Twitter Developers, s.f. *Working with timelines.* [En línea] Available at: https://developer.twitter.com/en/docs/tweets/timelines/guides/working-with-timelines
[Último acceso: 7 agosto 2020].

Twitter, 2020. *How to use hashtags.* [En línea] Available at: https://help.twitter.com/en/using-twitter/how-to-use-hashtags [Último acceso: 11 mayo 2021].

Ungria Gaitan, J., 2014. *Marketing de Contenidos y SEO: la clave del éxito.* [En línea] Available at: http://digitalmarketingtrends.es/marketing-de-contenidos-y-seo-la-clave-del-exito/
[Último acceso: 4 febrero 2020].

van Steenburg, E., 2012. Consumer recall of brand versus product banner ads. *Journal of Product and Brand Management,* 21(6), pp. 452-464.

Veglis, A. & Pomportsis, A., 2004. New production models for newspaper organizations. *WSEAS Transactions on Communications 3.1,* 3(1), pp. 218-222.

Villuendas Solsona, O., 2018. El valor de la publicidad on-line en la prensa digital: propuesta de un modelo de análisis de su eficiencia.

Vis, F., 2013. Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots. *Digital Journalism,* 1(1), pp. 27-47.

Voorhees, C., Mccall, M. & Calantone, R., 2011. *Customer loyalty: a new look at the benefits of improving segmentation efforts with rewards programs,* Ithaca, NY, USA: Center for Hospitality Research Publications.

Vosoughi, S., Roy, D. & Aral, S., 2018. The spread of true and false news online. *Science,* 359(6380), pp. 1146-1151.

Vratonjic, N., Freudiger, J. & Hubaux, J.-P., 2010. Integrity of the Web Content: The Case of Online Advertising.

Vu, H. T., 2014. The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism: Theory, Practice & Criticism,* 15(8), pp. 1094-1110.

Wahl-Jorgensen, K., 2017. A Manifesto of Failure for Digital Journalism. En: *Remaking the News: Essays on the Future of Journalism Scholarship in the Digital Age.* Cambridge, Massachusetts, Londres: The MIT Press.

Wainwright, C., 2012. *How to Conduct Competitive Analysis to Step Up Your Content Strategy.* [En línea] Available at: https://blog.hubspot.com/blog/tabid/6307/bid/31619/How-to-Conduct-Competitive-Analysis-to-Step-Up-Your-Content-Strategy.aspx
[Último acceso: 21 mayo 2019].

Wang, A. H., 2010. *Don't follow me. Spam Detection in Twitter.* s.l., IEEE, pp. 1-10.

Wang, X. y otros, 2011. *Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach.* New York, ACM Press, pp. 1031-1040.

Wilkinson, D. & Thelwall, M., 2012. Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology,* 53(8), pp. 1631-1646.

Williams, A., s.f. *TwitterOAuth PHP Library for the Twitter REST API.* [En línea] Available at: https://twitteroauth.com/ [Último acceso: 30 junio 2020].

Wilson, D. W., 2008. Monitoring technology trends with podcasts, RSS and Twitter. *Library Hi Tech News,* 25(10), pp. 8-12.

Wojcik, S. & Hughes, A., 2019. *Sizing Up Twitter Users,* s.l.: Pew Research Center.

Wolin, L. D. & Korgaonkar, P., 2003. Web advertising: Gender differences in beliefs, attitudes and behavior. *Internet Research,* 13(5), pp. 375-385.

Wolin, L. D., Korgaonkar, P. & Lund, D., 2002. Beliefs, attitudes and behaviour towards Web advertising. *International Journal of Advertising,* 21(1), pp. 87-113.

Woodruff, A. y otros, 1996. An Investigation of Documents from the World Wide Web. *Computer Networks and ISDN Systems,* 28(7-11), pp. 963-980.

Wouters, P., Zahedi, Z. & Costas, R., 2019. Springer handbook of science and technology indicators. *Springer handbook of science and technology indicators,* pp. 687-713.

Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J., 2011. *Who Says What to Whom on Twitter.* Hyderabad, International World Wide Web Conference Committee (IW3C2), pp. 705-714.

Wu, Z. y otros, 2013. Position-wise contextual advertising: Placing relevant ads at appropriate positions of a web page. *Neurocomputing,* Volumen 120, pp. 524-535.

Yamaguchi, Y., Takahashi, T., Amagasa, T. & Kitagawa, H., 2010. *TURank: Twitter User Ranking Based on User-Tweet Graph Analysis.* Berlin, Heidelberg, Springer, pp. 240-253.

Yang, J. & Counts, S., 2010. *Predicting the Speed, Scale, and Range of Information Diffusion in Twitter.* s.l., Association for the Advancement of Artificial Intelligence.

Yan, J. y otros, 2009. *How much can Behavioral Targeting Help Online Advertising?.* Madrid, ACM, pp. 261-270.

Yoo, C., 2009. Effects beyond click-through: Incidental exposure to Web advertising. *Journal of Marketing Communications,* 15(4), pp. 227-246.

Yoo, C. Y., 2008. Unconscious processing of Web advertising: Effects on implicit memory, attitude toward the brand, and consideration set. *Journal of Interactive Marketing,* 22(2), pp. 2-18.

Yuan, S., Abidin, A. Z., Sloan, M. & Wang, J., 2012. Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users.

Yuan, Y., Caulkins, J. P. & Roehrig, S., 1998. The relationship between advertising and content provision on the Internet. *European Journal of Marketing,* 32(7-8), pp. 677-687.

Yu, C.-C. & Stotlar, D. K., 2000. Advertising Banners on Sport Web Sites. *International Journal of Applied Sports Sciences,* Volumen 12, pp. 73-94.

Yu, M. y otros, 2011. *Target-dependent Twitter Sentiment Classification.* Portland, Association for Computational Linguistics, pp. 151-160.

Zangerle, E., Gassler, W. & Specht, G., 2011. *Recommending #-Tags in Twitter.* s.l., s.n., pp. 67-78.

Zhang, P. & Kim, Y., 2008. *Web advertising: What do we know about its acceptance and impacts?-A meta-analysis of the literature.* s.l., s.n., p. 246.

Zhao, W. X. y otros, 2011. *Comparing Twitter and Traditional Media Using Topic Models.* Berlin, Springer, pp. 338-349.

Zheng, X. y otros, 2015. Detecting spammers on social networks. *Neurocomputing,* 159(1), pp. 27-34.

Zheng, Z., Alonso Berrocal, J. L. & García de Figuerola Paniagua, L. C., 2013. Análisis Cibermétrico y Visual de Twitter. pp. 177-188.

Zhu, Z. A. y otros, 2010. *A Novel Click Model and Its Applications to Online Advertising.* New York, ACM, pp. 321-330.

Zion, L. & Craig, D., 2015. *Ethics for Digital Journalists: Emerging Best Practices.* Nueva York: Routledge.

Zubiaga, A., Spina, D., Martínez, R. & Fresno, V., 2015. Real-time classification of Twitter trends. *Journal of the Association for Information Science and Technology,* 66(3), pp. 462-473.

Zumstein, D. & Kaufmann, M., 2009. *A fuzzy Web analytics model for Web mining.* Algarve, Portugal, IADIS Press, pp. 59-66.

Glosario

A

AJAX

Siglas de JavaScript Asíncrono y XML. Es una técnica utilizada para el desarrollo de aplicaciones web de carácter interactivo., 124

API

Siglas en inglés de Application Programming Interface, o interfaz de prorgamación de aplicaciones. Hace referencia a un conjunto de procesos, funciones y métodos que ofrece una biblioteca de programación para poder ser empleada por otro programa informático, a menudo externo a la entidad fuente., 106

B

Backlink

Enlaces entrantes a una página web enviados desde otras páginas., 164

Banner

Espacio en un sitio web donde se inserta publicidad., 95

Base de datos

Conjunto de informaciones organizadas y estructuradas de un modo específico., 243

Benchmarking

Técnica para establecer comparaciones y medir el rendimiento., 123

Blog

Sitio web con formato de bitácora o diario personal., 99

C

Ciberespacio

Realidad virtual, construida digitalmente con ordenadores., 85

Clic

También conocido como click, se refiere a la acción de pulsar un botón en el ratón. Hacer clic se suele referir a presionar uno de los botones de un ratón y soltarlo rápidamente., 116

Cookie

Archivo que envía un sitio web y es almacenado en el navegador del usuario., 121

CPM

Siglas en inglés de Cost Per Millar, o precio por mil impresiones., 96

CPS

Siglas en inglés de Cost Per Sale, o precio por venta., 96

CR

Siglas en inglés de Conversation Rate, o ratio de conversión., 96

CTR

Siglas en inglés de Click Through Rate, o proporción por clics., 96

D

Digital

Información representada de modo binario. Se utiliza muy a menudo como sinónimo de online o "en línea"., 95

E

ECPM

Siglas en inglés de Effective Cost Per Millar, o precio estimado por mil impresiones., 96

Email

También conocido como e-mail o correo electrónico, es un mensaje digital que se transmite a través de una red informática., 94

Engagement

En inglés significa compromiso. En el ámbito del marketing digital se refiere a la cercanía que hay entre una marca y sus seguidores., 123

Enlace

Elemento que establece un vínculo entre dos documentos digitales., 86

F

Facebook

Red social fundada en 2004 por Mark Zuckerberg y que se ha convertido en una de las más famosas de la actualidad., 109

G

Google AdWords

Sistema desarrollado por Google para que las empresas puedan incluir anuncios en los resultados de las búsquedas y otras páginas web., 148

Google Analytics

Sistema desarrollado por Google para poder visualizar la analítica web., 148

Google Scholar

Herramienta de Google para buscar literatura académica., 170

Google Website Optimizer

Herramienta de Google para facilitar la toma de decisiones de editores y diseñadores web a la hora de realizar modificaciones en una página web o campaña publicitaria., 149

Google+

Red social perteneciente a Google. Ya no existe en la actualidad., 109

H

Hardware

Conjunto de componentes que forman la parte material (física) de un ordenador., 84

Hashtag

Significa en inglés etiqueta. Consiste en una palabra o serie de palabras o caracteres alfanuméricos precedidos de "#". Sirve como una herramienta de comunicación para organizar, clasificar o agrupar las publicaciones., 104

Hipermedia

Combinación de hipertexto y multimedia. Consiste en un conjunto de elementos vinculados entre sí mediante enlaces., 153

HTML

Siglas en inglés de HyperText Markup Language, o lenguajes de marcas de hipertexto. Se trata de un lenguaje de marcas para crar documentos web., 122

I

Inbound marketing

También conocido como marketing entrante, es una metodología que utiliza técnicas de marketing y publicidad no intrusivas para entrar en contacto con el usuario cuando comienza el proceso de conversión, y acompañarlo así hasta la transacción final., 177

Internet

Es una red de redes que permite la interconexión descentalizada de diferentes dispositivos., 76

J

JavaScript

Lenguaje de programación que sirve sobre todo para producir recursos interactivos en una página web., 124

K

KPI

Siglas en inglés de Key Performance Indicator, o indicador clave de rendimiento. Hace referencia a una medida de evaluación del rendimiento de un proceso., 124

M

Machine learning

En inglés significa aprendizaje automático, y es el subcampo de la informática que desarrolla técnicas para que las computadoras aprendan., 121

Multidispositivo

Hace referencia a aquello que se aplica a varios dispositivos de manera indistinta., 126

Multimedia

Lo que utiliza varios medios de manera simultánea mientras transmite información., 85

O

Offline

En inglés significa desconectado. Hace referencia a aquello que no está en línea., 141

Online

Palabra inglesa que significa "en línea". Designa algo que está conectado o que hace uso de una red, usualmente Internet., 76

P

PageRank
Marca creada por Google para englobar una familia de algoritmos utilizados para asignar una valoración numérica a la relevancia de los documentos indexados por su motor de búsqueda., 157
Podcast
Acrónimo en inglés de iPod, un reproductor de música, y broadcast, transmisión o emisión en español, se trata de un archivo de audio descargado y/o escuchado vía Internet a través de un dispositivo., 228
Polimedia
Sistema de creación de materiales multimedia como apoyo a la docencia y a la investigación., 153
Porcentaje de rebote
Mide cuántos usuarios visitan una página web y la abandonan directamente sin visitar ninguna otra página de la misma web., 131
PPA
Siglas en inglés de Pay Per Action, o pago por acción., 96
PPC
Siglas en inglés de Pay Per Click, o pago por clic., 96
PPD
Siglas en inglés de Pay Per Download, o pago por descarga., 96
PPI
Siglas en inglés de Pay Per Installation, o pago por instalación., 96
PPS
Siglas en inglés de Pay Per Sale, o pago por venta., 96
PPV
Siglas en inglés de Pay Per Visual, o pago por visualización., 96

R

Red social
Sitio web que favorece la creación de comunidades virtuales., 101
Retuit
También conocido como retweet, es la acción de tuitear un mensaje de otro usuario. Se suele utilizar RT como abreviatura de este., 104
RSS
Es una tecnología o formato de datos que sirve para el envío de contenidos a aquellos usuarios que estén suscritos a la web., 215

S

SEO
Siglas en inglés de Search Engine Optimization, técnica de optimización de un sitio web para alcanzar el mejor posicionamiento posible en los buscadores., 146

Software
Conjunto de programas, instrucciones y reglas informáticas que ejecutan tareas en un ordenador., 84
Spammer
Usuario que publica spam o contenido publicitario no solicitado., 191

T

TT
Siglas en inglés de Trending Topic, o tema que es tendencia en el momento actual., 104
Tuit
También conocido como tuit, se trata de una publicación o actualización de estado realizada en Twitter., 102
Twitter
Red de microblogging que permite escribir y leer mensajes conocidos como tuits., 100

U

URL
Siglas en inglés de Uniform Resource Locator, o localizador uniforme de recursos. Es una secuencia de caracteres que, mediante un estándar, permite denominar recursos en Internet para que puedan ser localizados., 115

V

Videocast
Técnica multimedia que permite emitir información en vídeo mediante una transmisión digital., 228

W

Wiki
Página web con contenido que puede ser editado por múltiples usuarios, pudiendo todos ellos añadir, modificar o eliminar información., 200
World Wide Web
También conocido sucintamente como Internet o Web (en mayúscula), se utiliza en el ámbito tecnológico para referirse a una red informática y, en general, a Internet., 221

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Programa de doctorado en Industrias de la Comunicación y Culturales

Departamento de Comunicación Audiovisual, Documentación e Historia del Arte

TESIS DOCTORAL

Diseño de una metodología cibermétrica de cálculo del éxito para la optimización de contenidos web

Víctor Manuel Yeste Moreno

Directores

Dr. Jorge Ignacio Serrano Cobos

Dra. María de los Ángeles Calduch Losa

Valencia, septiembre de 2021

TOMO II

6. Anexos

6.1. Análisis de los subconjuntos de artículos

Debido a que se ha aplicado la metodología cinco veces, a continuación, se puede encontrar todos los análisis realizados que llevaron a las regresiones en la fase 1 de análisis de los datos de entrenamiento.

Es necesario recordar que el conjunto total de los datos se puede descargar desde figshare[44], para hacer posible así la réplica del estudio en el futuro. El conjunto de datos se ha publicado con el título "Dataset from PhD Thesis" en la categoría Social and Community Informatics, con licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) y las palabras clave: *digital journalism*, *machine learning*, *cyberjournalism*, *altmetrics*, *web analytics*, *Twitter analytics*, *trend analysis*, *trend prediction*, *digital content optimization*, *cybermetrics* y *web advertising*.

6.1.1. Análisis de los artículos de la categoría Cine

6.1.1.1. Variables de éxito

El objetivo de esta fase es la predicción de los valores de éxito mediante el resto de los indicadores. Por ello, es importante comenzar por analizar estos escenarios por separado para ver tanto sus características como tratar de describir su comportamiento anómalo, si lo tuvieran.

a) Páginas vistas únicas (total)

Este escenario de éxito se ve identificado por la columna uniquepageviews_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 187 valores con un rango de entre 1 y 348.

Presenta un sesgo estandarizado de 30,2482 y una curtosis estandarizada de 108,433. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

[44] https://figshare.com/articles/dataset/Dataset_from_the_PhD_Thesis/16553061

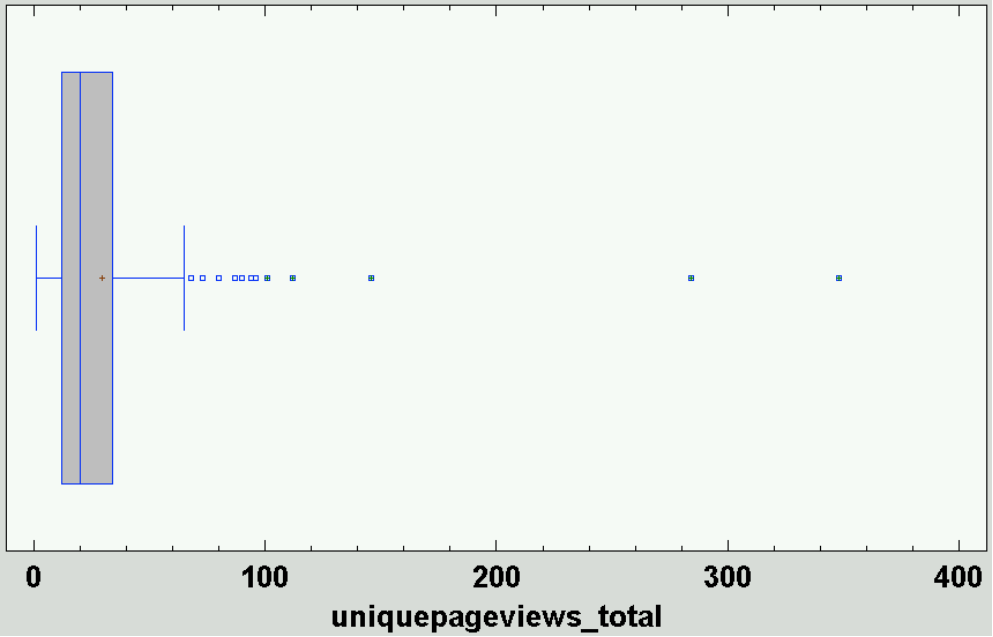


Figura 198. Cine: Gráfico de Caja y Bigotes para el valor uniquepageviews_total

En la Figura 1 se puede comprobar que existen valores anómalos de tipo extremo de 101 páginas vistas o más.

b) AdSense eCPM (promedio)

Este escenario de éxito se identifica con la columna adsense_ecpm_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 2,65.

Presenta un sesgo estandarizado de 53,4658 y una curtosis estandarizada de 298,47. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

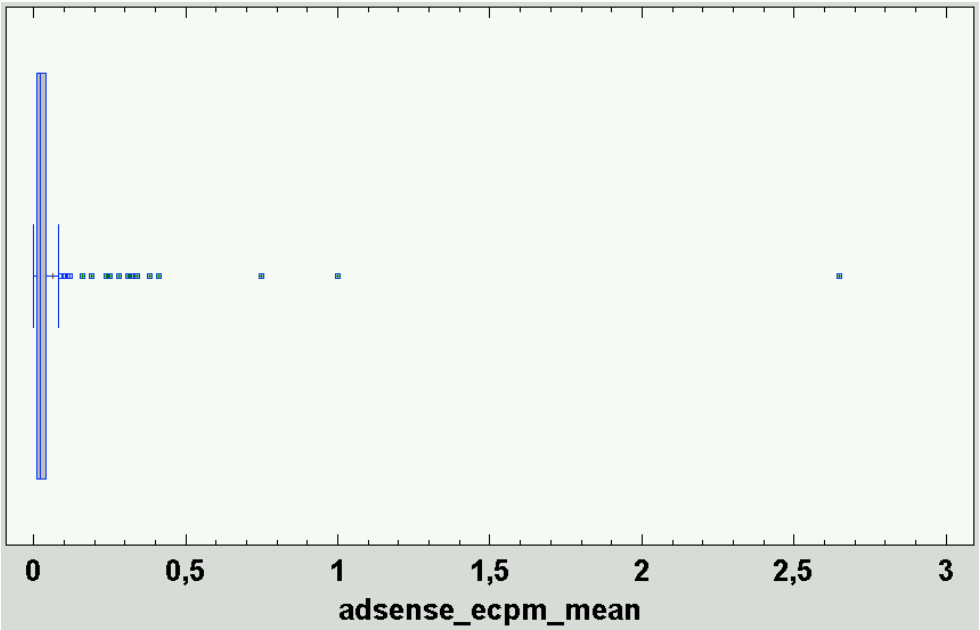


Figura 199. Cine: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean

En la Figura 199 se puede observar que existen valores anómalos de tipo extremo de 0,16 o más.

c) Duración de la visita (promedio)

Este escenario de éxito se identifica con la columna avgtimeonpage_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 884.

Presenta un sesgo estandarizado de 13,625 y una curtosis estandarizada de 24,5875. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

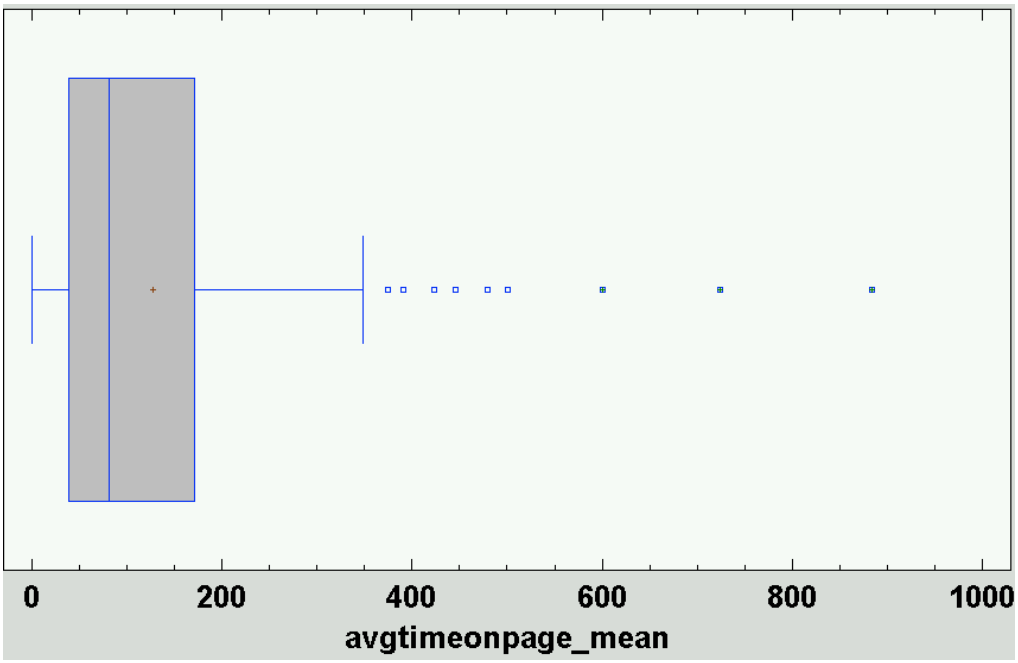


Figura 200. Cine: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean

En la Figura 200 se puede observar que existen valores anómalos de tipo extremo de 600,32 o más.

d) Páginas vistas por sesión (promedio)

Este escenario de éxito se identifica con la columna pageviewspersession_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0,39 y 9.

Presenta un sesgo estandarizado de 25,4653 y una curtosis estandarizada de 101,468 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

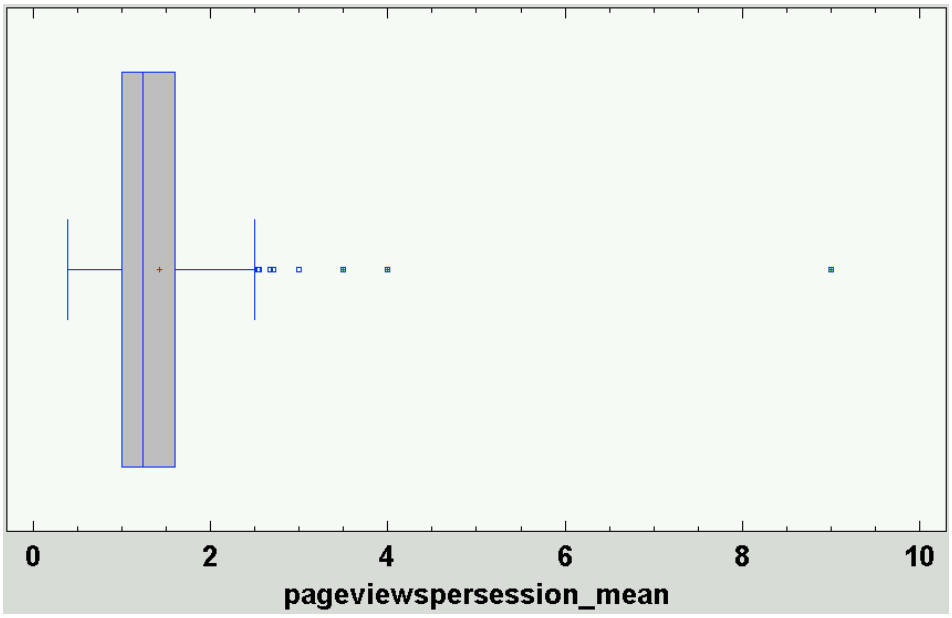


Figura 201. Cine: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean

En la Figura 201 se puede observar que existen valores anómalos de tipo extremo de 3,5 o más.

e) Nº de retuits en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 4.

Presenta un sesgo estandarizado de 5,96759 y una curtosis estandarizada de 21,3049 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

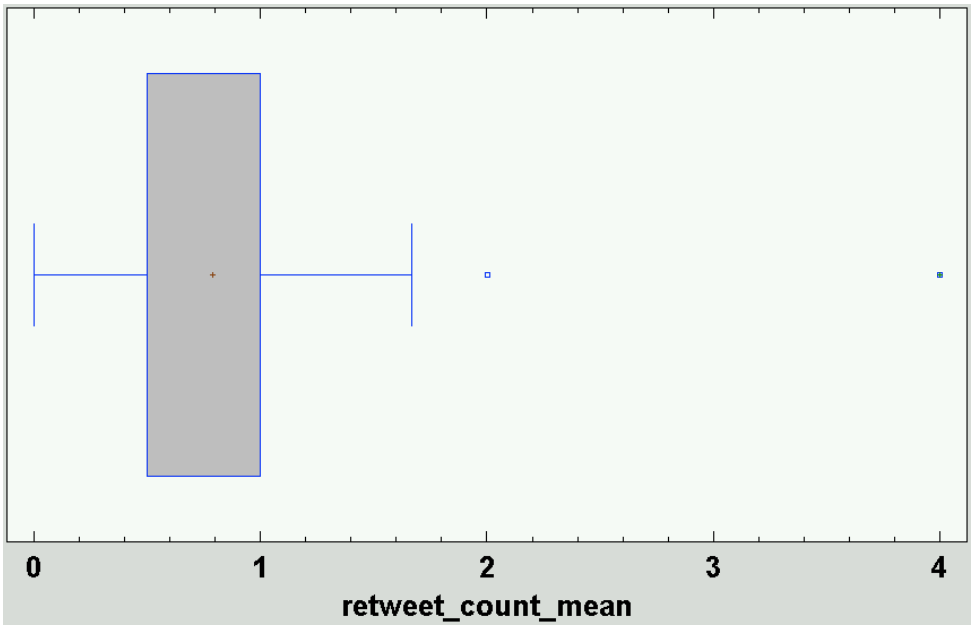


Figura 202. Cine: Gráfico de Caja y Bigotes para el valor retweet_count_mean

En la Figura 202 solo se observa un valor anómalo de tipo extremo con 4 retuits de promedio.

f) Nº de favoritos en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 11.

Presenta un sesgo estandarizado de 23,2589 y una curtosis estandarizada de 98,9982 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

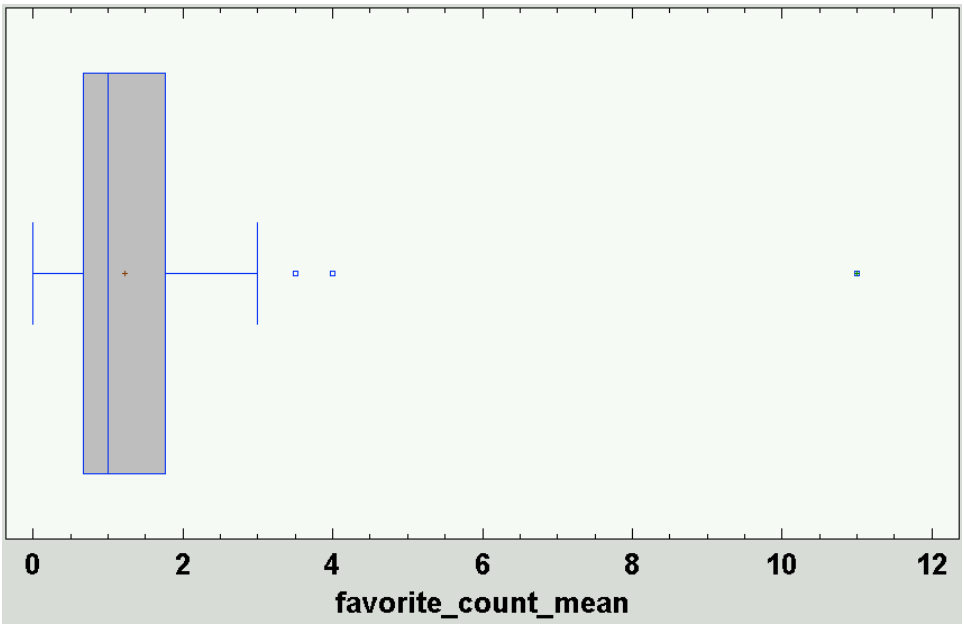


Figura 203. Cine: Gráfico de Caja y Bigotes para el valor favorite_count_mean

En la Figura 203 solo se observa un valor anómalo de tipo extremo con 11 favoritos de promedio.

g) Nº de tuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 187 valores con un rango de entre 0 y 13.983.

Presenta un sesgo estandarizado de 10,4403 y una curtosis estandarizada de 13,9138 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

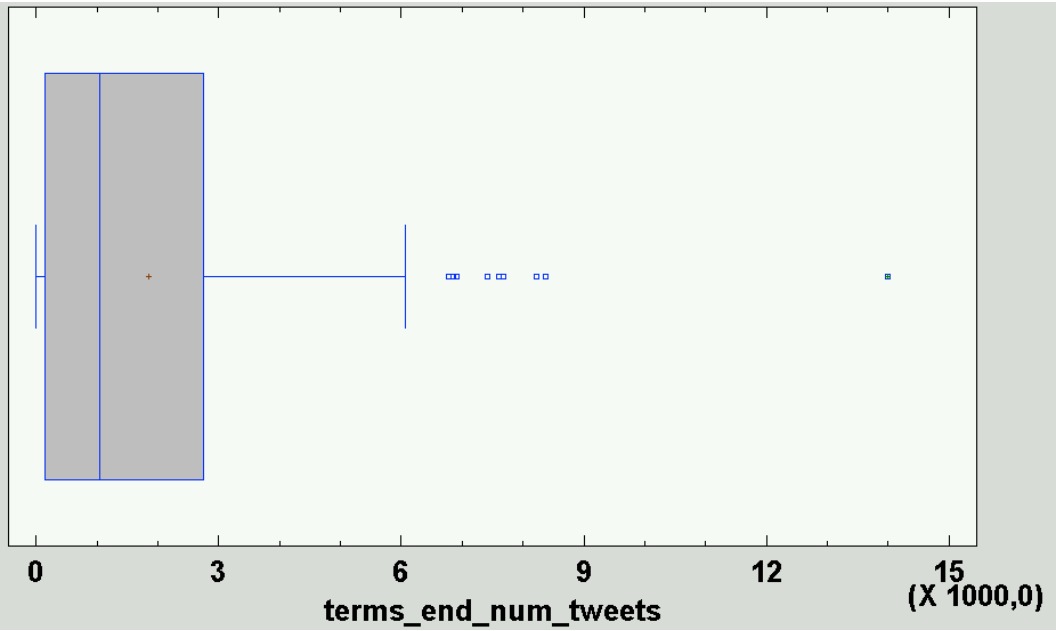


Figura 204. Cine: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets

En la Figura 204 se puede observar que existe un valor anómalo de tipo extremo de 13,983.

h) Nº de retuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 45.998.500.

Presenta un sesgo estandarizado de 64,3579 y una curtosis estandarizada de 400,756 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

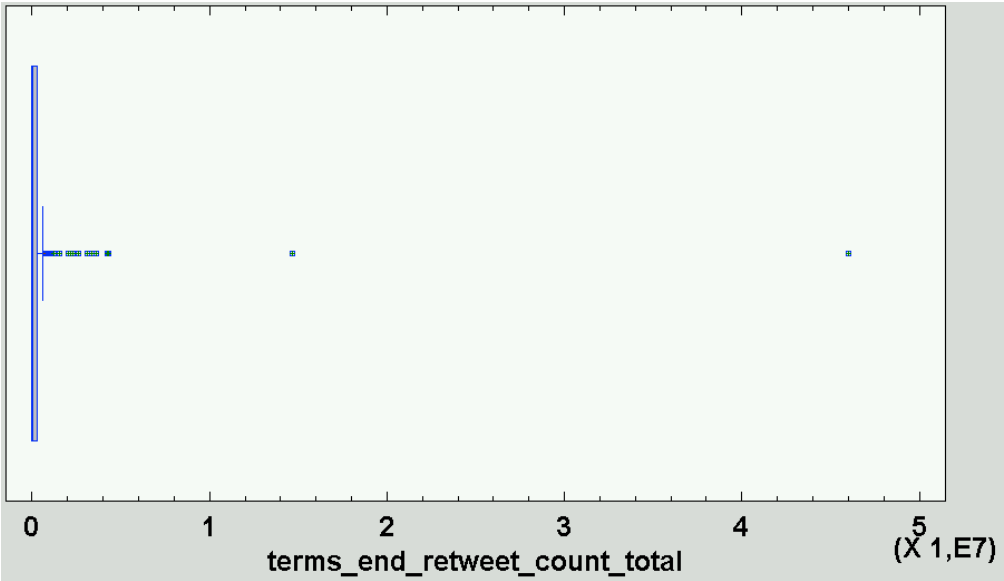


Figura 205. Cine: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total

En la Figura 205 se puede observar que existen valores anómalos de tipo extremo de 1.209.810 o más.

i) Nº de retuits de la tendencia 14 días después (promedio)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 2705.

Presenta un sesgo estandarizado de 43,3667 y una curtosis estandarizada de 189,337. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

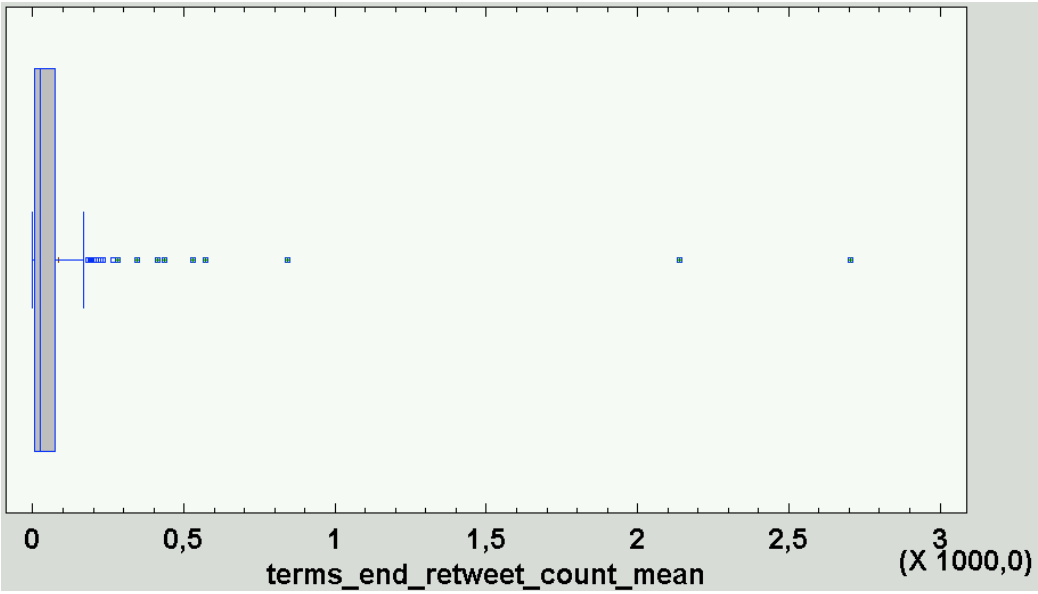


Figura 206. Cine: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean

En la Figura 206 se puede observar que existen valores anómalos de tipo extremo de 281,41 o más.

j) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de las variables de éxito. Se pueden observar los siguientes datos de estas:

Tabla 52

Cine: Resumen estadístico de las variables de éxito

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
uniquepageviews_total	1405,29	30,2482	108,433
adsense_ecpm_mean	0,0480546	53,4658	298,47
avgtimeonpage_mean	16608,4	13,625	24,5875
pageviewspersession_mean	0,694959	25,4653	101,468

retweet_count_mean	0,258539	5,96759	21,3049
favorite_count_mean	1,16344	23,2589	98,9982
terms_end_num_tweets	4848180	10,4403	13,9138
terms_end_retweet_count_total	12729800000000	64,3579	400,756
terms_end_retweet_count_mean	71619,7	43,3667	189,337

Se puede observar en la Tabla 52 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 53

Cine: Resumen estadístico de las variables de éxito con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(uniquepageviews_total)	0,747983	0,0311532	2,91533
log(adsense_ecpm_mean)	1,10957	6,80015	5,98526
log(avgtimeonpage_mean)	0,962044	-0,632425	-1,69228
log(pageviewspersession_mean)	0,19726	2,75629	4,22674

log(retweet_count_mean)	0,14093	-1,96382	5,14455
log(favorite_count_mean)	0,275152	0,455725	3,43429
log(terms_end_num_tweets)	4,38427	-5,31741	0,222075
log(terms_end_retweet_count_total)	13,7825	-3,7089	-0,630952
log(terms_end_retweet_count_mean)	3,92887	-2,33145	0,178614

Todas las variables salvo log(avgtimeonpage_mean) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. Sin embargo, mantienen valores mucho menores, próximos al rango de -2 a +2 y con una dispersión muy parecida, por lo que, debido a la naturaleza de los datos, en este estudio se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

Nos quedamos, por tanto, con las variables que tengan menores sesgo y curtosis estandarizados de su forma original o transformación logarítmica.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 54

Cine: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(adsense_ecpm_mean)	0,00105349
log(retweet_count_mean)	0
log(terms_end_num_tweets)	0,00262223

En la Tabla 54 se puede ver que ninguna variable tiene el valor-p necesario (mayor o igual que 0,05) para confirmar que sigue una distribución normal, por lo que no se añade ninguna de estas variables al modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 55

Cine: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(avgtimeonpage_mean)
terms_ini_retweet_count_mean	log(pageviewspersession_mean)
terms_ini_favorite_count_total	log(favorite_count_mean)
terms_ini_favorite_count_mean	log(terms_end_retweet_count_total)
terms_ini_followers_talking_rate	log(terms_end_retweet_count_mean)
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, por tanto, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio, el número total de retuits de la tendencia 14 días después y el promedio de retuits de la tendencia 14 días después.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

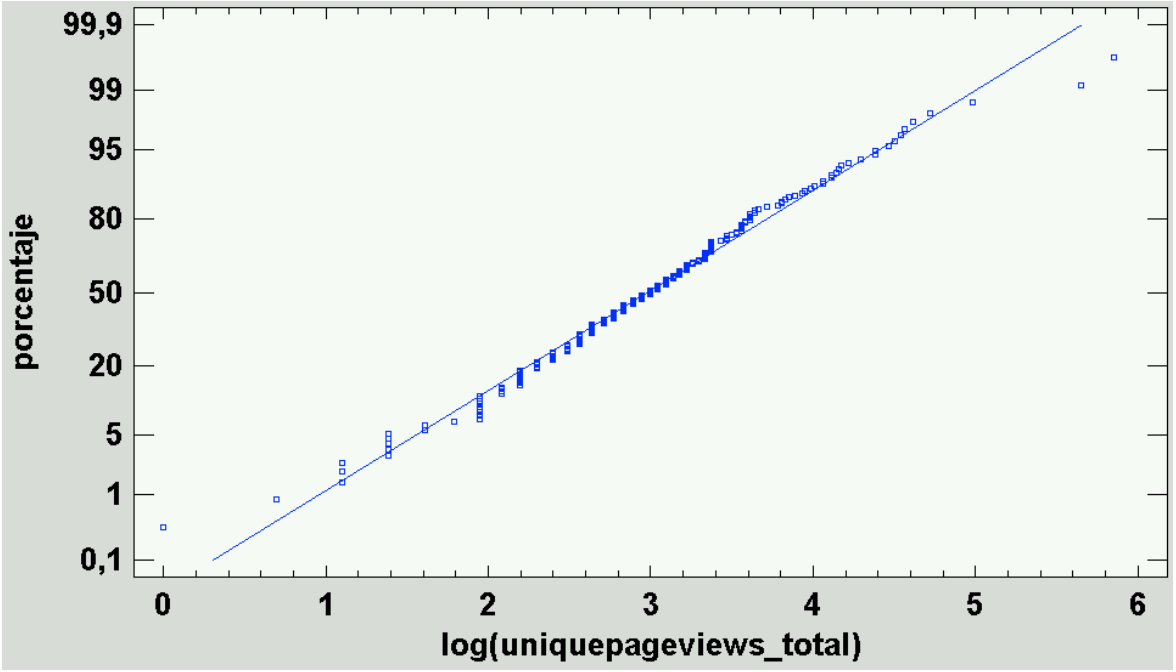


Figura 207. Cine: Gráfico de probabilidad normal de la variable log(uniquepageviews_total)

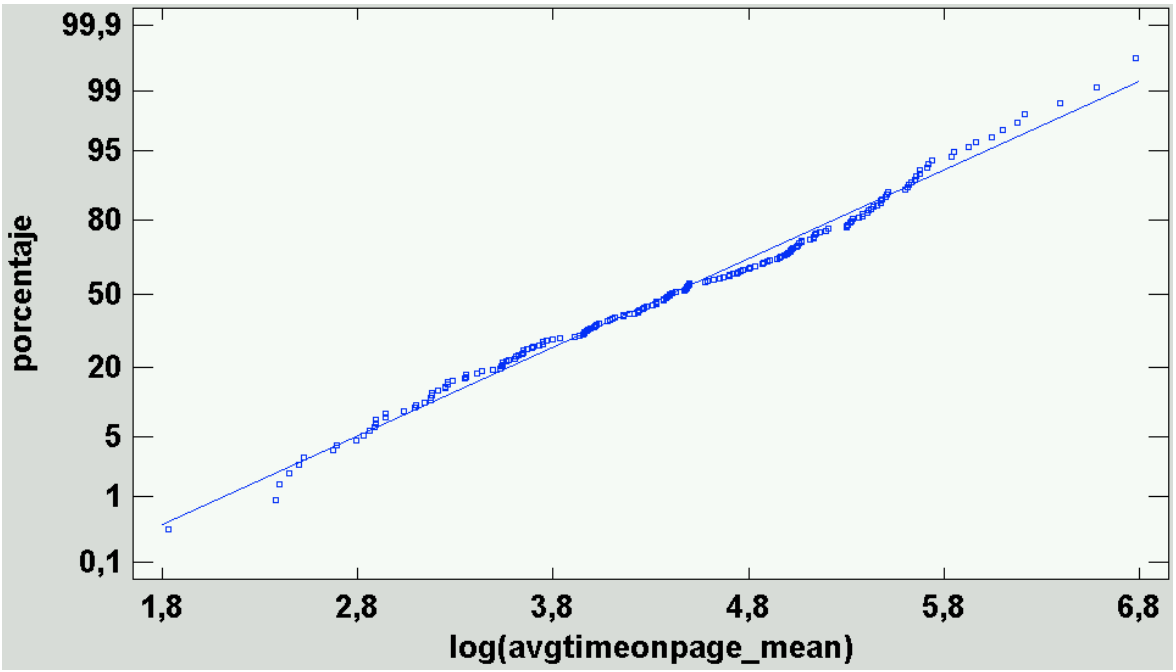


Figura 208. Cine: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean)

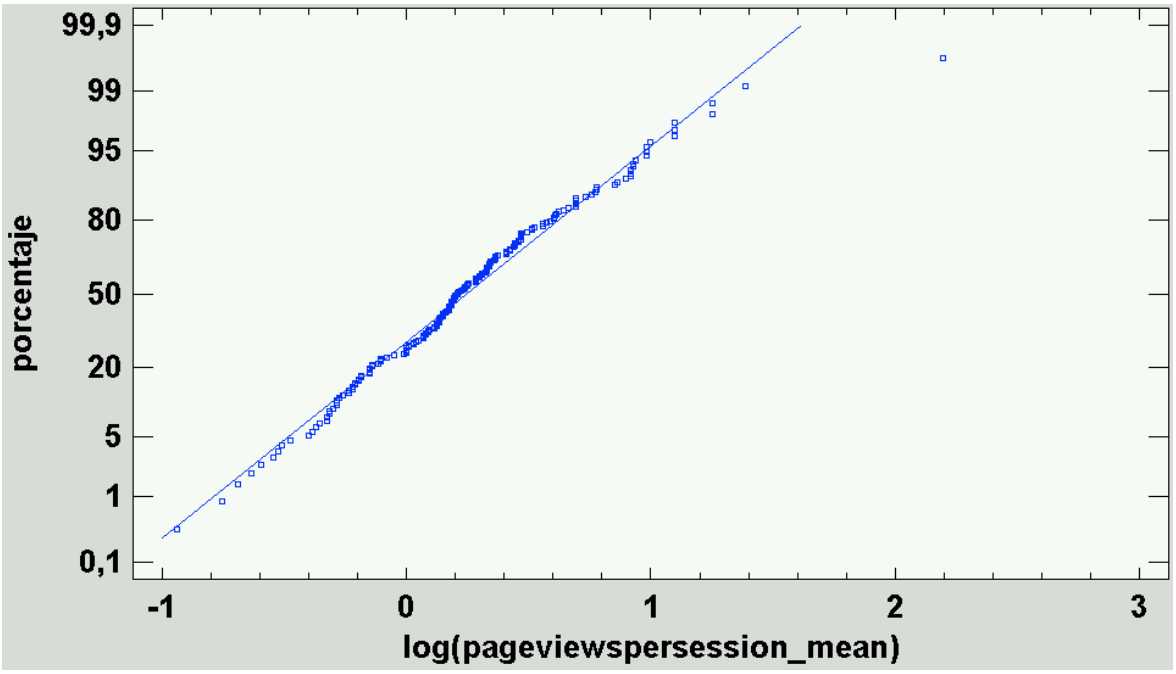


Figura 209. Cine: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean)

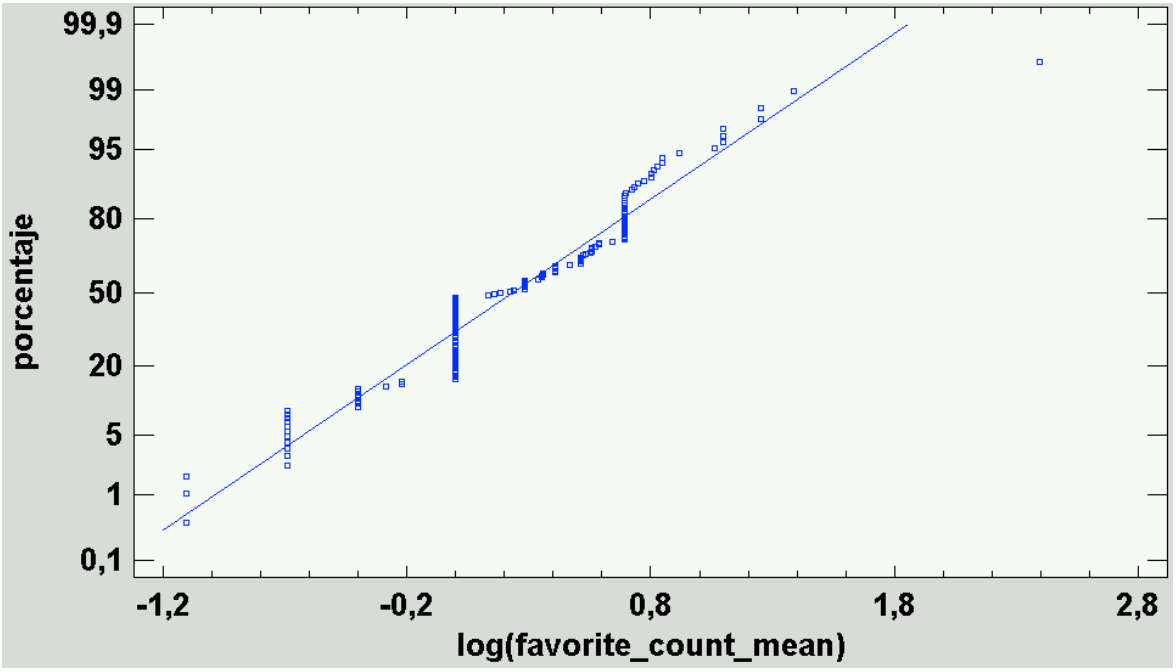


Figura 210. Cine: Gráfico de probabilidad normal de la variable log(favorite_count_mean)

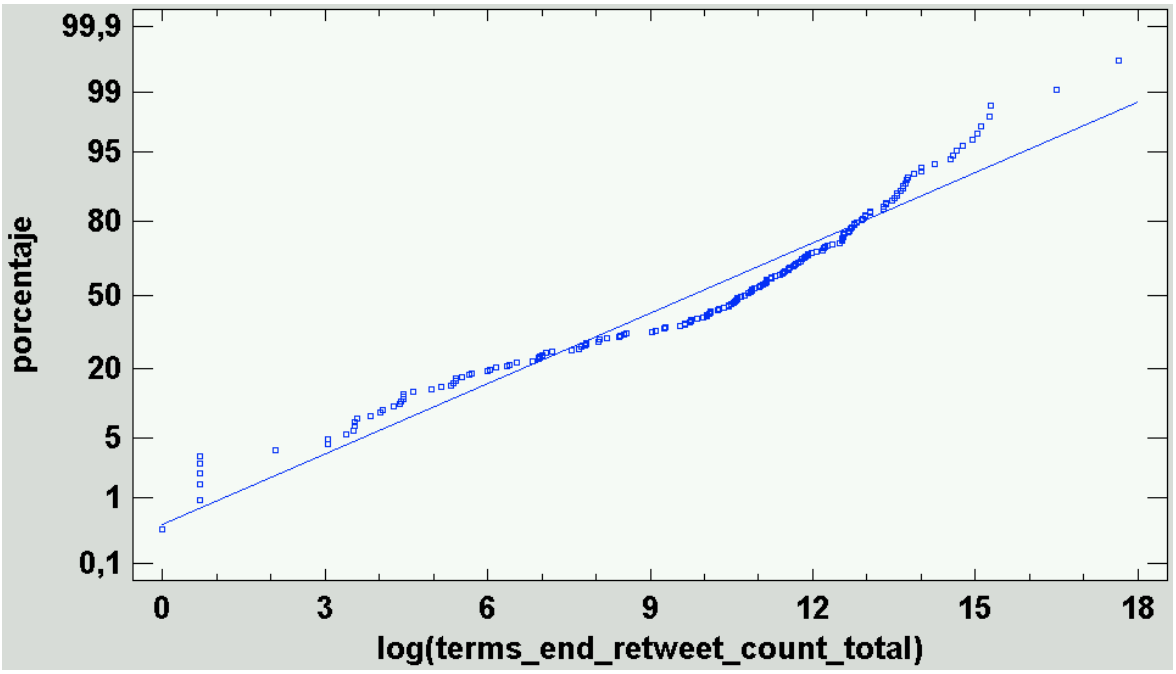


Figura 211. Cine: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total)

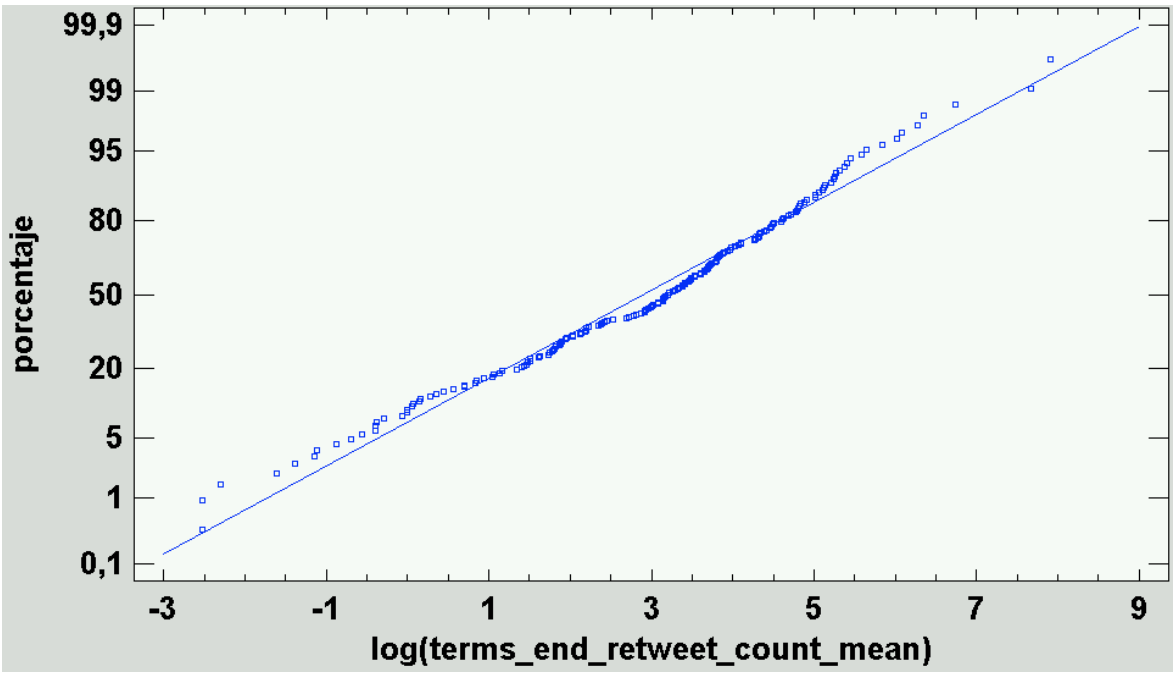


Figura 212. Cine: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

k) Filtro de alta correlación

Puesto que se han propuesto más de una variable de éxito, es conveniente evaluar si están asociadas mediante una fuerte correlación. Si fuera ese el caso, el análisis de este estudio

se simplificaría ya que las conclusiones de una variable servirían para las que estén fuertemente correlacionadas con ella, lo que permitiría dedicar menos recursos a la predicción y toma de decisiones.

Para ello, es necesario realizar un análisis multivariado de las variables de éxito con su correspondiente transformación logarítmica, cuya matriz de correlaciones Pearson se puede observar en la Figura 213:

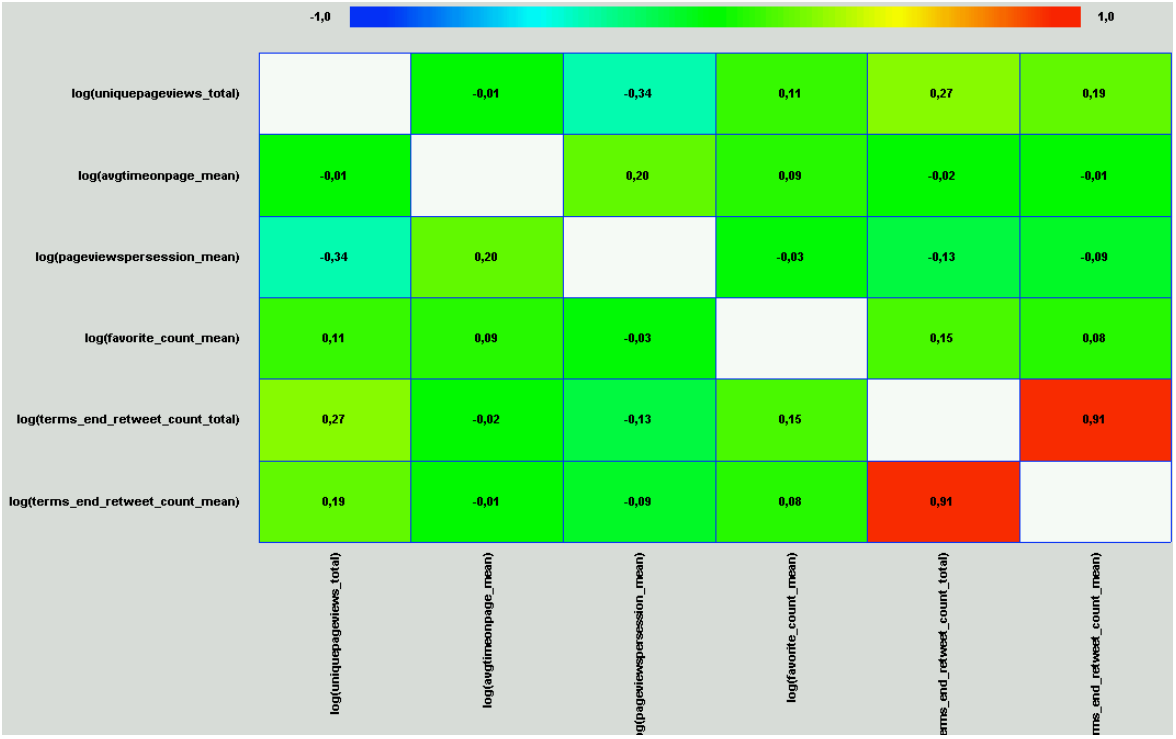


Figura 213. Cine: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- log(terms_end_retweet_count_total) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,9086 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto log(terms_end_retweet_count_mean) por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que log(terms_end_retweet_count_total).

La tabla de variables quedaría como sigue:

Tabla 56

Cine: Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(avgtimeonpage_mean)
terms_ini_retweet_count_mean	log(pageviewspersession_mean)
terms_ini_favorite_count_total	log(favorite_count_mean)
terms_ini_favorite_count_mean	log(terms_end_retweet_count_mean)
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, finalmente, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

6.1.1.2. Variables de predicción

Se parte de un conjunto de datos con diez características, por lo que es conveniente tratar de reducir la dimensión del conjunto, restableciendo la varianza sin modificar la información relevante de los datos en sí. Esto posibilitará que se reduzca el tiempo y el coste de la computación y facilita la visualización y el análisis de los datos. Además, es una condición necesaria para aplicar la regresión lineal múltiple (Anon., 2017).

a) Número de tuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 187 valores con un rango de entre 0 y 12.945.

Presenta un sesgo estandarizado de 8,63547 y una curtosis estandarizada de 7,03048. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

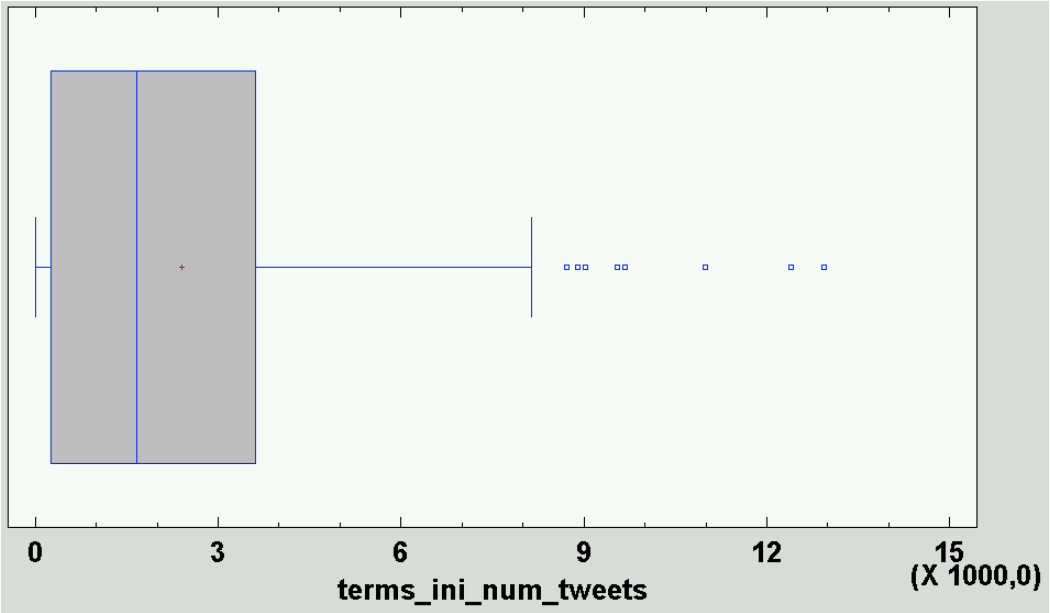


Figura 214. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets

En la Figura 214 se puede observar que no existen valores anómalos de tipo extremo.

b) Número de retuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 187 valores con un rango de entre 0 y 15.756.500.

Presenta un sesgo estandarizado de 31,2503 y una curtosis estandarizada de 112,491. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

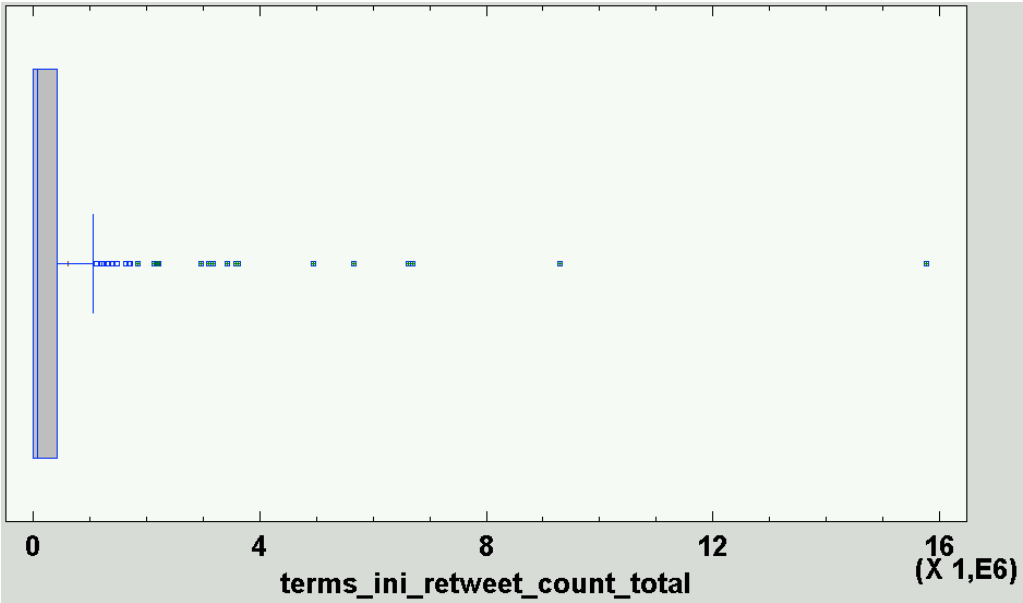


Figura 215. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total

En la Figura 215 se puede observar que existen valores anómalos de tipo extremo de 1.854.670 o más.

c) Número de retuits de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 6.260,75.

Presenta un sesgo estandarizado de 61,247 y una curtosis estandarizada de 376,719. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

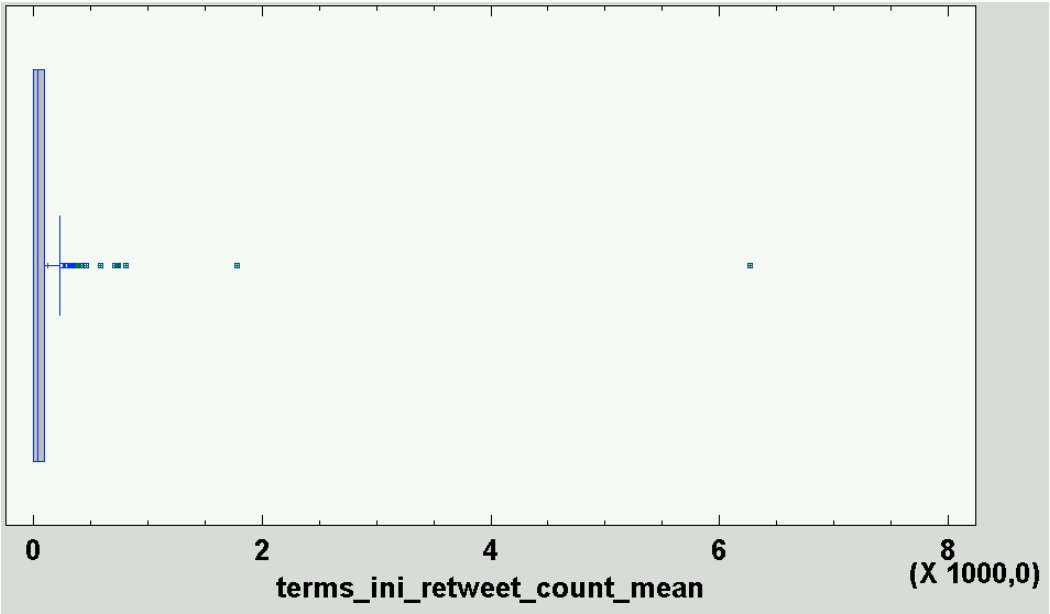


Figura 216. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean

En la Figura 216 se puede observar que existen valores anómalos de tipo extremo de 373,96 o más.

d) Número de favoritos de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 187 valores con un rango de entre 0 y 56.430.

Presenta un sesgo estandarizado de 10,9887 y una curtosis estandarizada de 11,4942. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

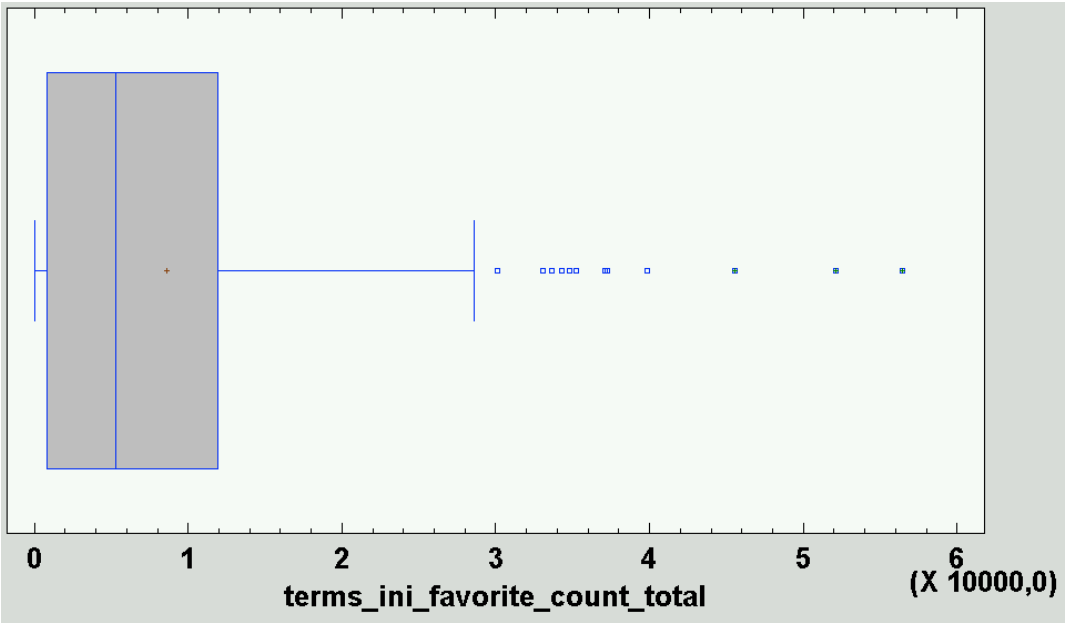


Figura 217. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total

En la Figura 217 se puede observar que existen valores anómalos de tipo extremo de 45.567 o más.

e) Número de favoritos de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 12,88.

Presenta un sesgo estandarizado de 9,31009 y una curtosis estandarizada de 14,5243. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

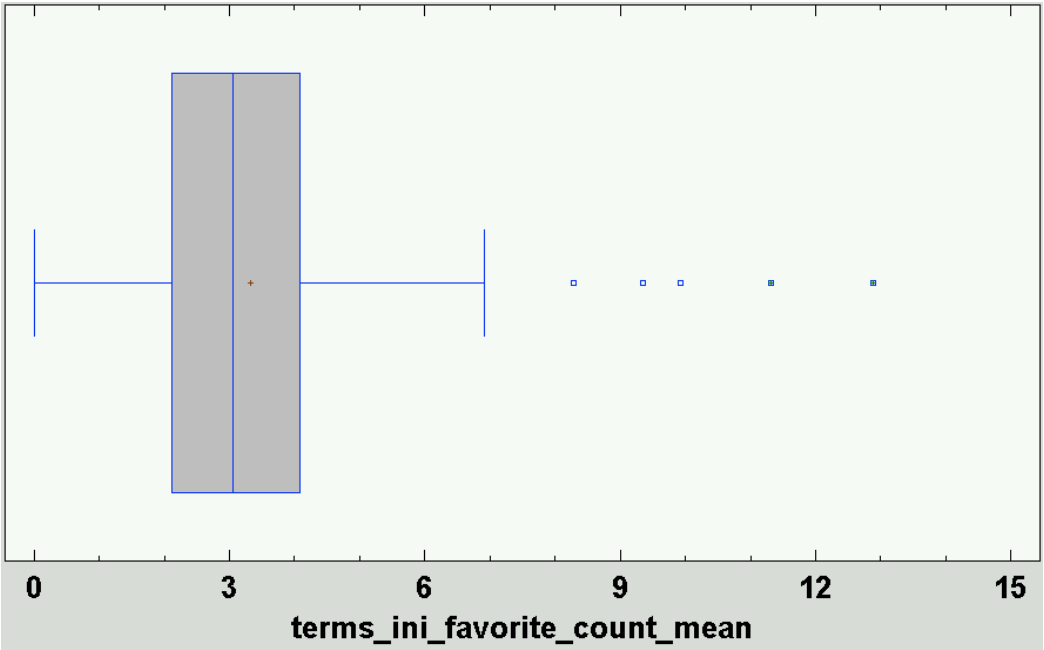


Figura 218. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean

En la Figura 218 se puede observar que existen valores anómalos de tipo extremo de 11,31 o más.

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_followers_talking_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 0,5.

Presenta un sesgo estandarizado de 32,1512 y una curtosis estandarizada de 139,658. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

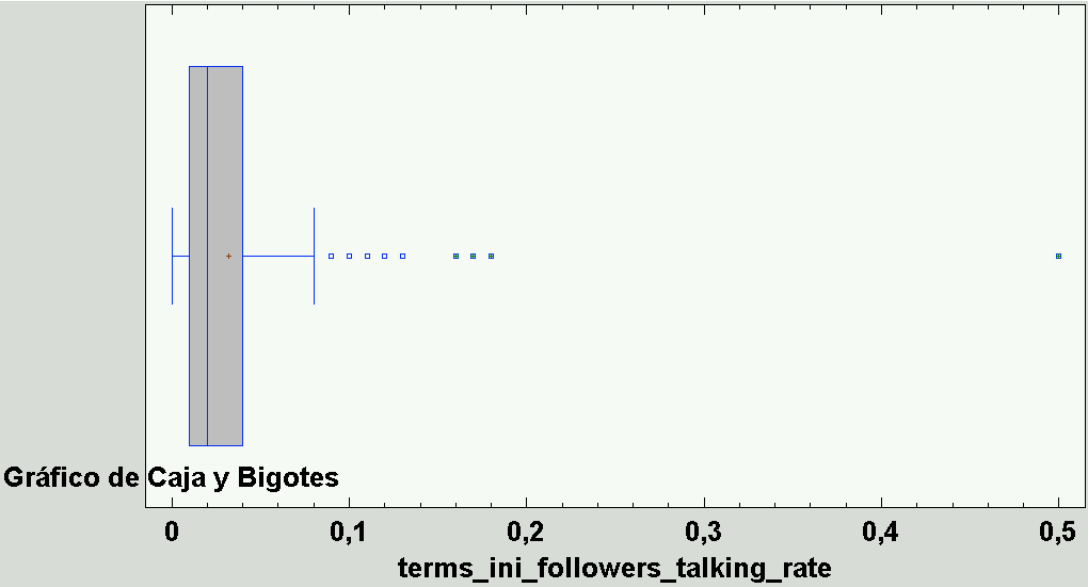


Figura 219. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate

En la Figura 219 se puede observar que existen valores anómalos de tipo extremo de 0,16 o más.

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_followers_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 269,72 y 373831.

Presenta un sesgo estandarizado de 28,0918 y una curtosis estandarizada de 88,2075. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

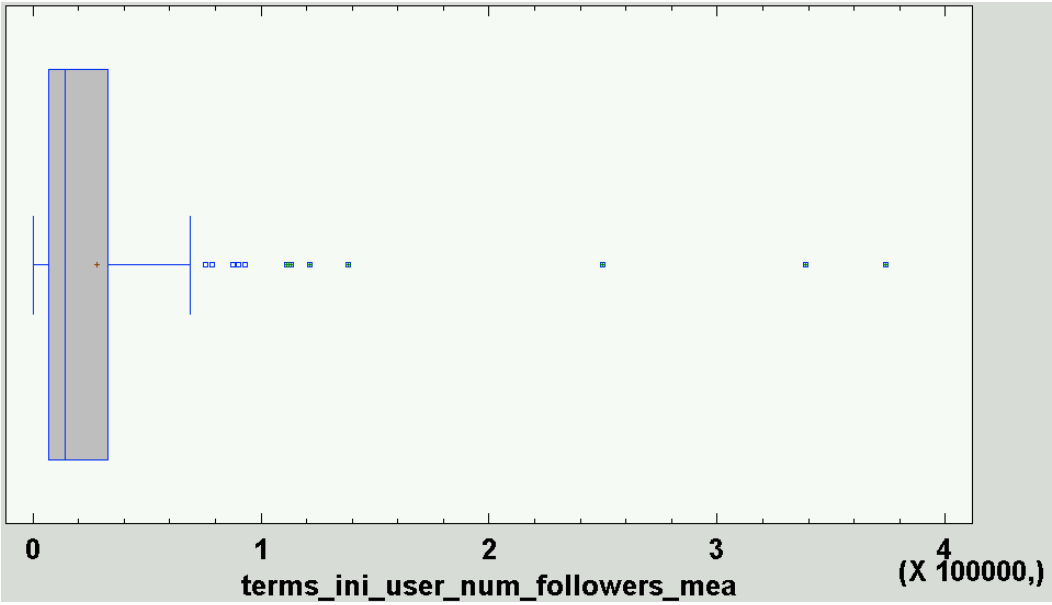


Figura 220. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean

En la Figura 220 se puede observar que existen valores anómalos de tipo extremo de 111.363 o más.

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_tweets_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 10.576,3 y 457.769.

Presenta un sesgo estandarizado de 28,273 y una curtosis estandarizada de 99,5191. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

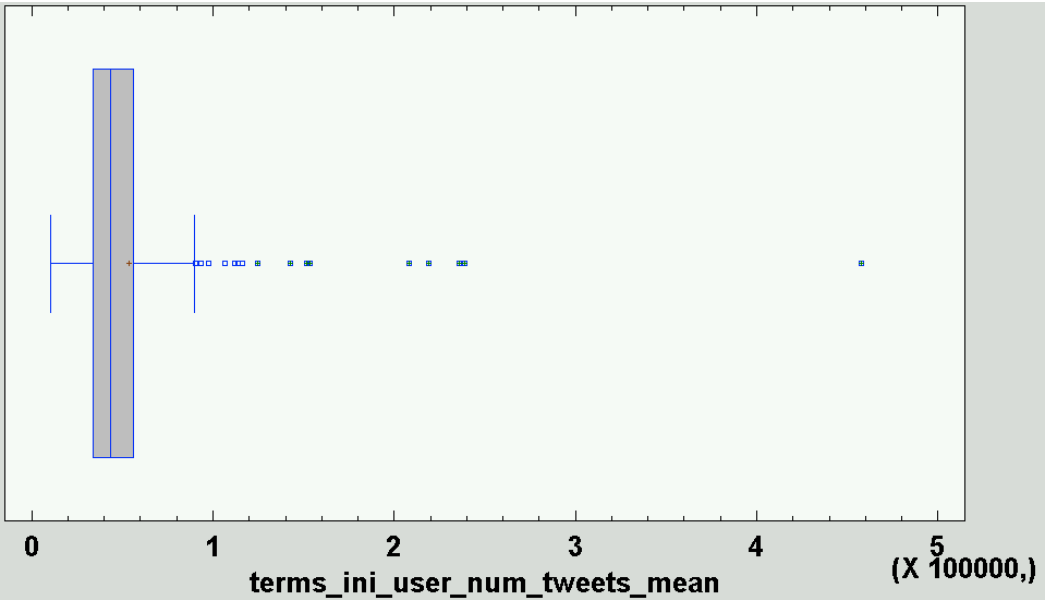


Figura 221. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean

En la Figura 221 se puede observar que existen valores anómalos de tipo extremo de 124.743 o más.

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_age_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 923,06 y 9.264.

Presenta un sesgo estandarizado de 32,6712 y una curtosis estandarizada de 166,275. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

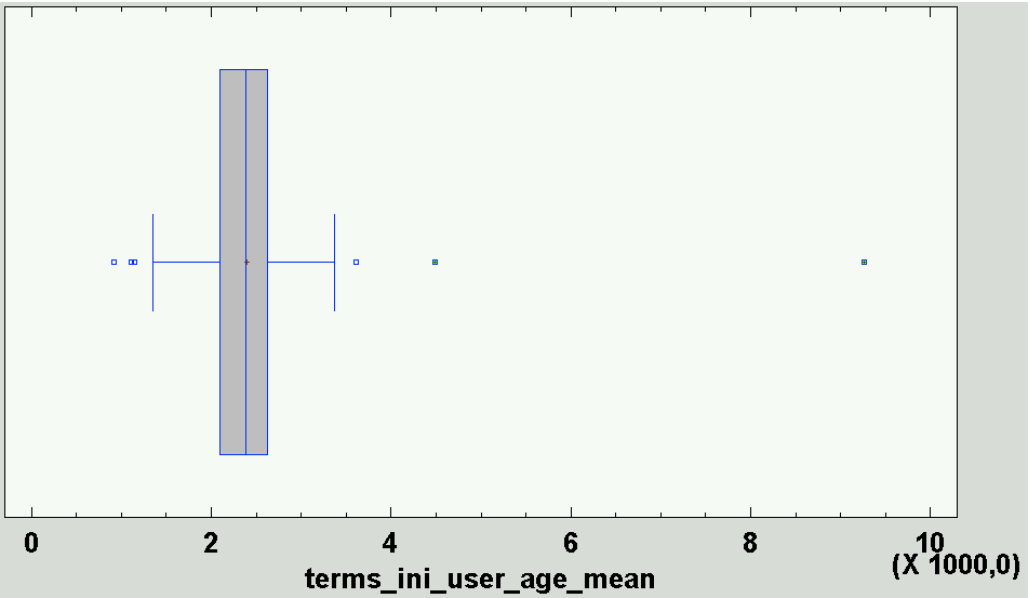


Figura 222. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean

En la Figura 222 se puede observar que existen valores anómalos de tipo extremo de 4487,3 o más.

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial

Esta variable de predicción se identifica con la columna terms_ini_url_inclusion_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 187 valores con un rango de entre 0 y 1.

Presenta un sesgo estandarizado de 4,30376 y una curtosis estandarizada de 2,77035. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

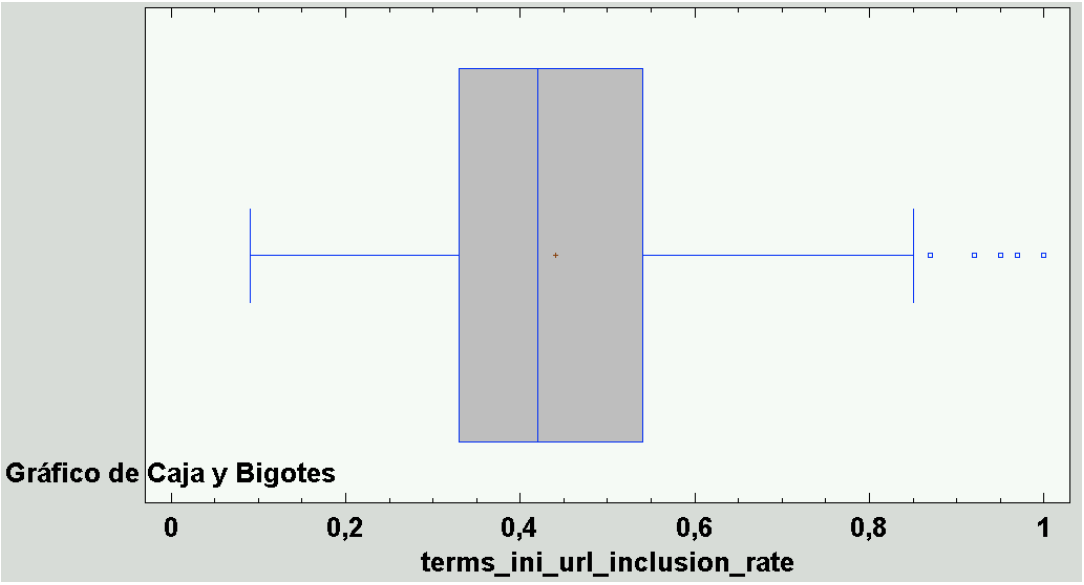


Figura 223. Cine: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate

En la Figura 223 se puede observar que no existen valores anómalos de tipo extremo.

k) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de todas las variables de predicción y éxito. De esta manera se obtienen todos los datos de las variables del modelo en un mismo análisis. Se pueden observar los siguientes datos de estas:

Tabla 57

Cine: Resumen estadístico de las variables de predicción

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
terms_ini_num_tweets	6.851.550	8,63547	7,03048
terms_ini_retweet_count_total	2.818.240.000.000	31,2503	112,491
terms_ini_retweet_count_mean	239.071	61,247	376,719

terms_ini_favorite_count_total	114.956.000	10,9887	11,4942
terms_ini_favorite_count_mean	3,52751	9,31009	14,5243
terms_ini_followers_talking_rate	0,00227635	32,1512	139,658
terms_ini_user_num_followers_mean	2.060.090.000	28,0918	88,2075
terms_ini_user_num_tweets_mean	2.097.070.000	28,273	99,5191
terms_ini_user_age_mean	449.067	32,6712	166,275
terms_ini_url_inclusion_rate	0,0310818	4,30376	2,77035

Se puede observar en la Tabla 57 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple, es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 58

Cine: Resumen estadístico de las variables de predicción con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(terms_ini_num_tweets)	3,40075	-5,46556	0,97641

log(terms_ini_retweet_count_total)	11,7499	-3,83464	-0,816051
log(terms_ini_retweet_count_mean)	3,59824	-1,35145	-0,678558
log(terms_ini_favorite_count_total)	4,16389	-6,18929	2,68984
log(terms_ini_favorite_count_mean)	0,365273	-5,72733	8,4499
log(terms_ini_followers_talking_rate)	0,690276	3,47704	0,109614
log(terms_ini_user_num_followers_mean)	1,24168	0,102952	1,47404
log(terms_ini_user_num_tweets_mean)	0,278312	5,26212	8,14203
log(terms_ini_user_age_mean)	0,0505013	2,29153	25,7952
log(terms_ini_url_inclusion_rate)	0,185085	-3,72898	2,85169

Todas las variables, salvo log(terms_ini_retweet_count_mean) y log(terms_ini_user_num_followers_mean), mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. En este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

También se puede apreciar que los valores de varianza son bastante parecidos salvo en el caso de log(terms_ini_retweet_count_total), por lo que se cumple la condición de homocedasticidad menos en esa variable. log(terms_ini_retweet_count_total) se elimina del modelo actual para que dicho requisito se cumpla.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la

prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 59

Cine: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(terms_ini_num_tweets)	0,0010425
log(terms_ini_favorite_count_total)	0,00154635
log(terms_ini_favorite_count_mean)	0,119846
log(terms_ini_user_num_tweets_mean)	0,074392
log(terms_ini_user_age_mean)	0,0322829

En la Tabla 59 se puede ver que las variables que superan el valor-p necesario (mayor o igual que 0,05) para confirmar que siguen una distribución normal son log(terms_ini_favorite_count_mean) y log(terms_ini_user_num_tweets_mean), por lo que dichas variables también son tenidas en cuenta en el modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 60

Cine: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)

log(terms_ini_user_num_followers_mean)	log(favorite_count_mean)
log(terms_ini_user_num_tweets_mean	log(terms_end_retweet_count_mean)
log(terms_ini_url_inclusion_rate)	

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

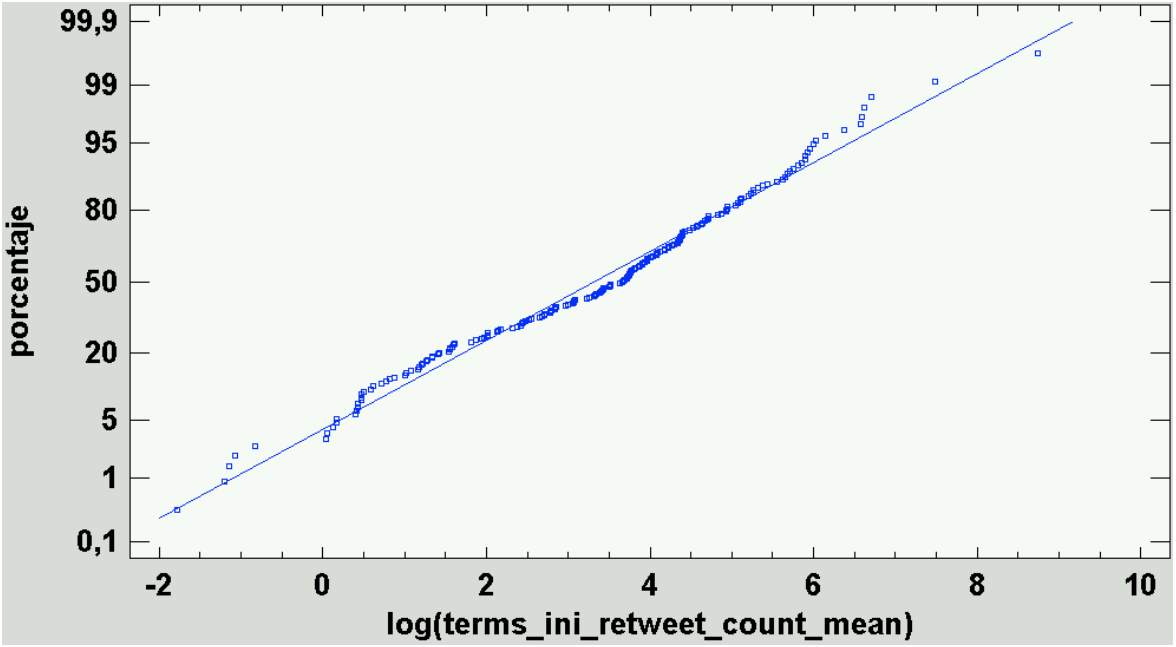


Figura 224. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean)

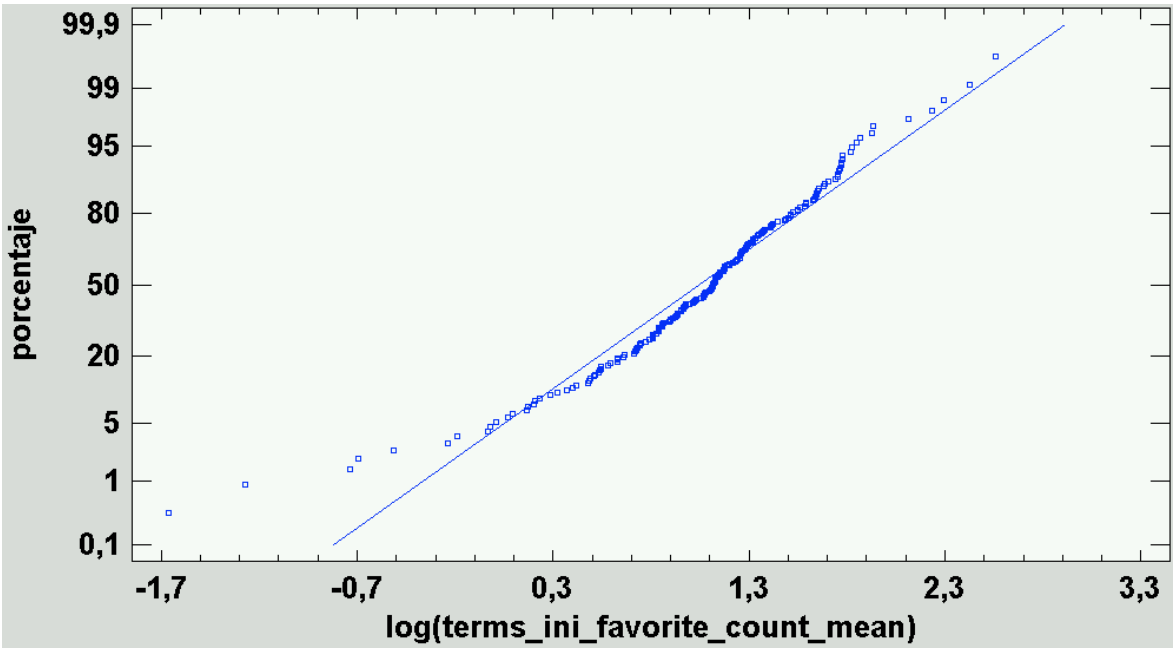


Figura 225. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean)

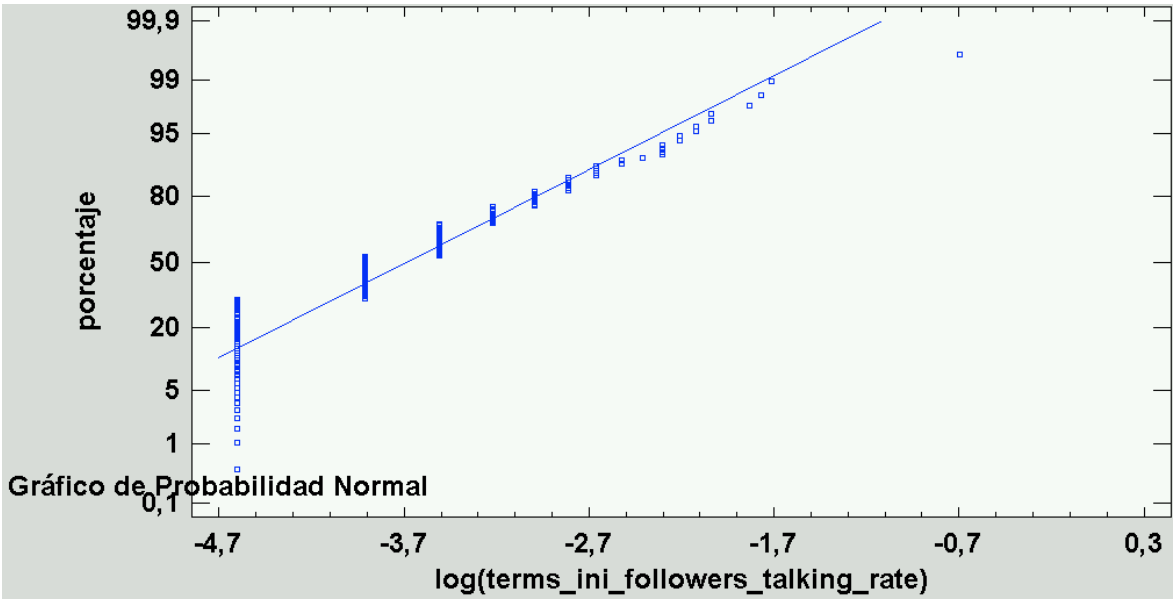


Figura 226. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate)

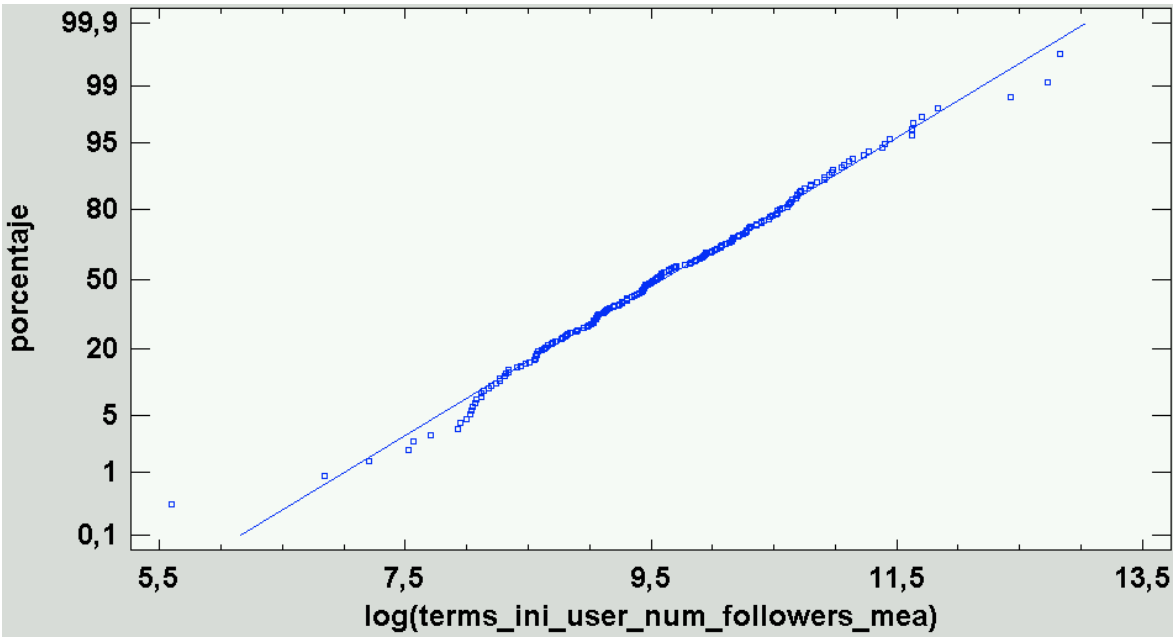


Figura 227. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean)

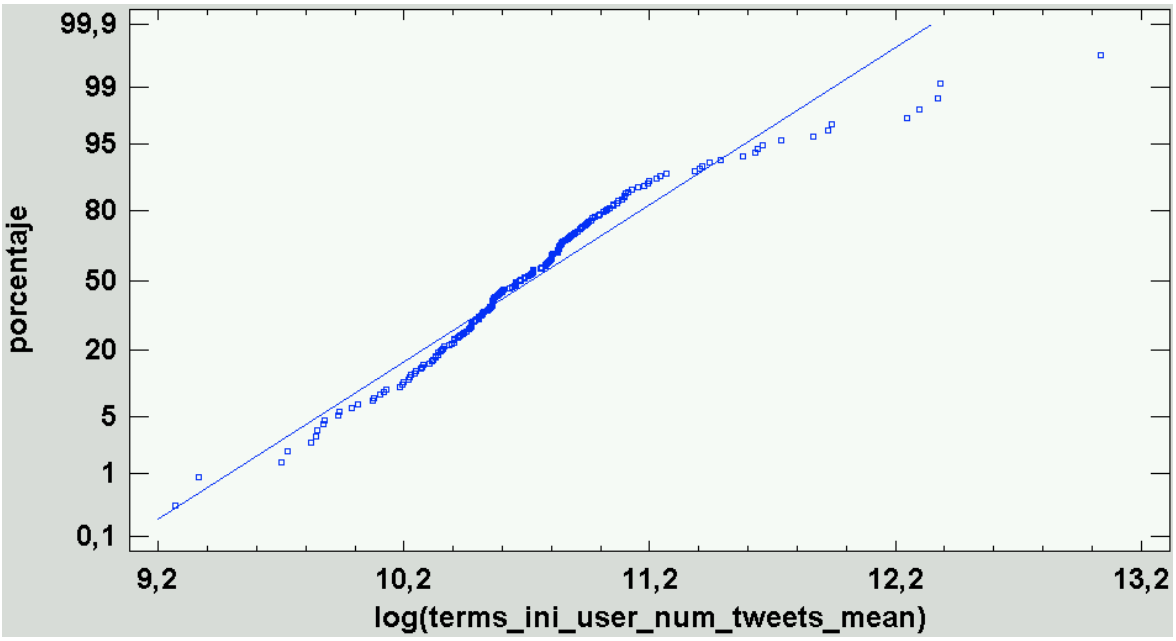


Figura 228. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean)

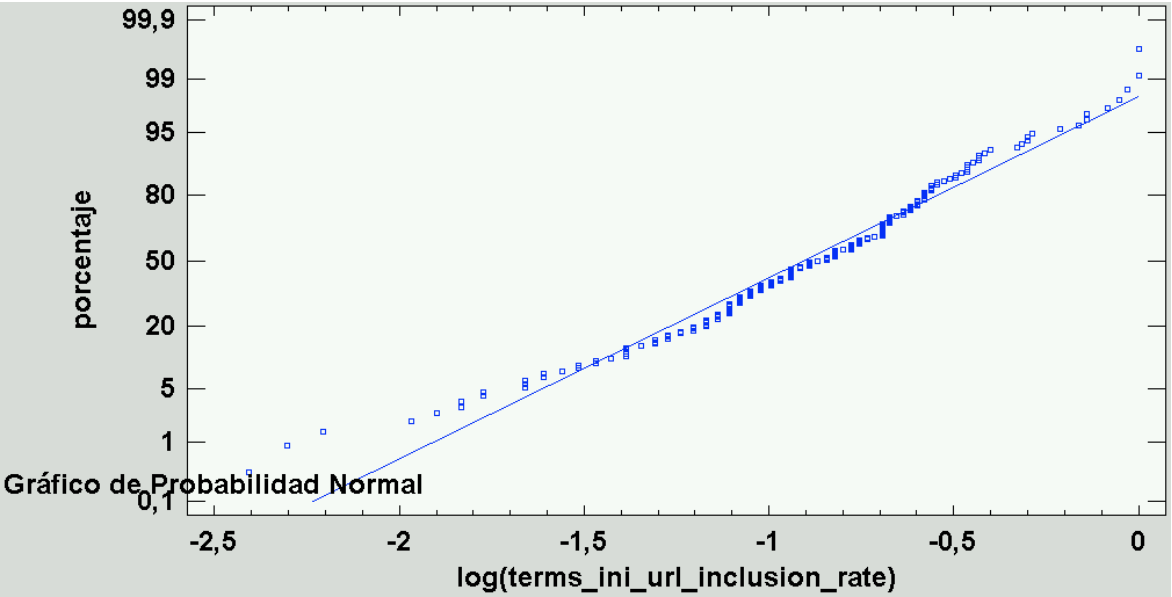


Figura 229. Cine: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

l) Filtro de alta correlación (colinealidad)

Antes de proceder, es conveniente analizar la correlación que existe entre las variables con las que contamos para el modelo. Puesto que estas han sido normalizadas en el apartado anterior, dicho análisis de correlación Pearson se realizará con su transformación logarítmica.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

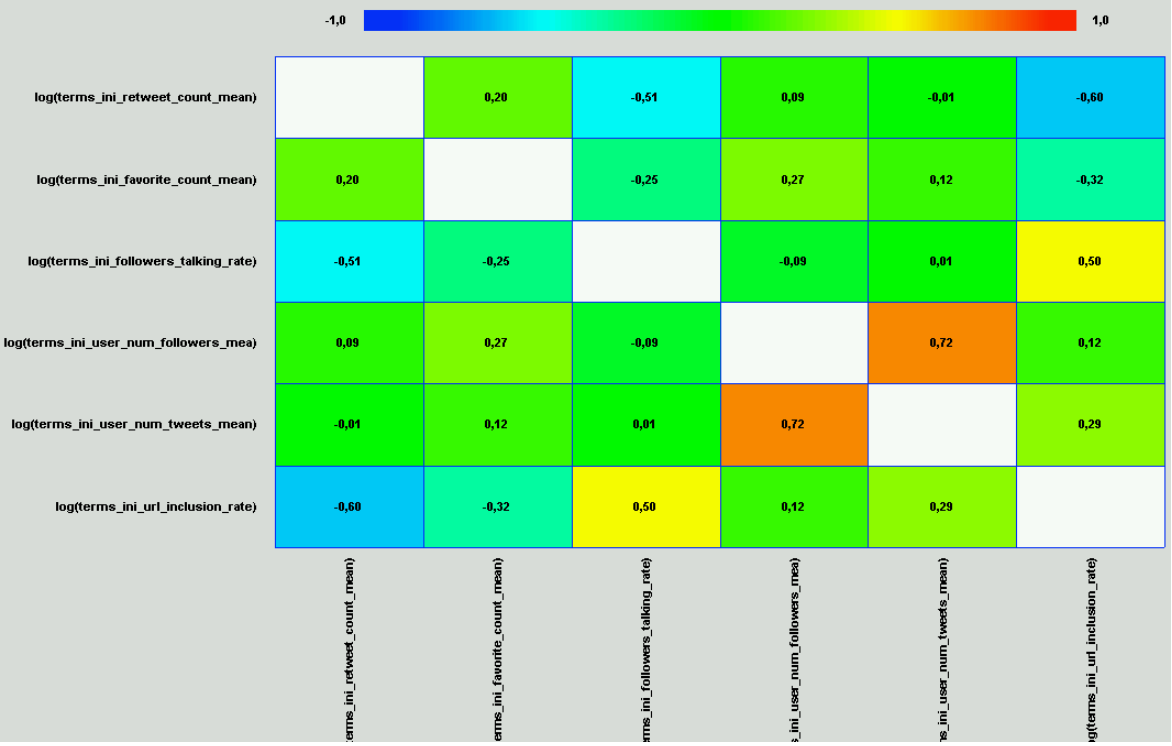


Figura 230. Cine: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente

Al hacerlo, se han obtenido la siguiente conclusión:

- log(terms_ini_user_num_followers_mean) y log(terms_ini_user_num_tweets_mean) tienen un coeficiente de correlación de 0,7205 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto log(terms_ini_user_num_followers_mean) por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que log(terms_ini_user_num_tweets_mean).

La tabla de variables quedaría como sigue:

Tabla 61

Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)

log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	log(favorite_count_mean)
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

m) Análisis de componentes principales (ACP)

A continuación, se aplica el análisis de componentes principales (ACP, o PCA en inglés), una técnica que sirve par describir un conjunto de datos según nuevas variables no correlacionadas llamadas componentes (Dunteman, 1989).

El objetivo es representar los datos de la mejor manera posible a través de mínimos cuadrados, construyendo una transformación lineal según un nuevo sistema de coordenadas para los datos originales. Es decir, se plantea representar la variabilidad de los datos con el menor número de componentes o fórmulas posible, las cuales son combinaciones lineales de las variables originales (Dunteman, 1989).

Si las variables originales están muy correlacionadas entre sí, la mayor parte de la variabilidad se podrá expresar en pocas componentes. Si están totalmente incorrelacionadas, el número de componentes será igual al de las variables y este análisis carecerá de interés.

Las componentes se ordenan según la varianza original siendo la primer componente la que tenga la varianza de mayor tamaño. Cuanto mayor sea su varianza, mayor será la información que aporta esa componente (Amat Rodrigo, 2017).

Al construir la matriz de coeficientes de correlación, es posible una base de vectores propios, cuya transformación lineal es necesaria para mejora la simplicidad e interpretación que permita tratar de reducir la dimensionalidad de los datos (Dunteman, 1989). Esta reducción se efectuaría seleccionando las componentes principales que más aportan a la varianza e ignorando el resto. Esta selección se produce ordenando las componentes de

mayor a menor aportación a la explicación de la variabilidad, y seleccionando tantas como sean necesarias hasta alcanzar un valor propio mayor o igual a 1.

De esta manera, el método ACP condensa la información de múltiples características en unas pocas, ya que se pretende explicar aproximadamente la información con menos valores que los originales.

Realizando el análisis de componentes principales se ha obtenido un total de dos componentes, explicando así el 68,628% de los datos con un valor propio de 1,17944. Las dos componentes tienen la Tabla 62 de pesos, siendo cada peso un valor de entre -1 y 1.

Tabla 62

Cine: Tabla de pesos de las componentes

	Componente	*Componente*
	1	2
log(terms_ini_retweet_count_mean)	0,541123	-0,132032
log(terms_ini_favorite_count_mean)	0,353349	0,487547
log(terms_ini_followers_talking_rate)	-0,520195	0,0547434
log(terms_ini_user_num_followers_mea)	0,0975612	0,81577
log(terms_ini_url_inclusion_rate)	-0,549735	0,276385

De esta manera, por ejemplo, la primer componente principal tiene la fórmula siguiente, en donde los valores de las variables se han estandarizado restándoles su promedio y dividiéndolos entre su desviación estándar:

0,541123 * log(terms_ini_retweet_count_mean) + 0,353349 * log(terms_ini_favorite_count_mean) – 0,520195 * log(terms_ini_followers_talking_rate) + 0,0975612 * log(terms_ini_user_num_followers_mean) – 0,549735 * log(terms_ini_url_inclusion_rate)

Se puede representar la mayor parte de esa variabilidad con solo dos componentes principales, de los cuales:

- Componente 1: está en buena medida explicada por el promedio de retuits de usuarios que no siguen al medio y la no inclusión de una URL en los tuits. Se podría interpretar que consta de aquellas tendencias con un gran nivel de participación conversacional de usuarios que no son seguidores de la cuenta del medio.
- Componente 2: está en buena medida explicada por el número medio de favoritos y usuarios con un gran número de seguidores. Se puede comprender como aquellas tendencias con recepción positiva de usuarios con muchos seguidores.

También se observa que las dos variables que aportan positivamente a las dos componentes principales son: el promedio de favoritos y el promedio de seguidores de los usuarios que participan en la tendencia. El resto sí que aportan un valor negativo en alguna de ellas.

La relación entre las variables y las dos componentes principales se puede ver en la siguiente gráfica, ya que las variables se muestran en dos dimensiones formadas por estas componentes:

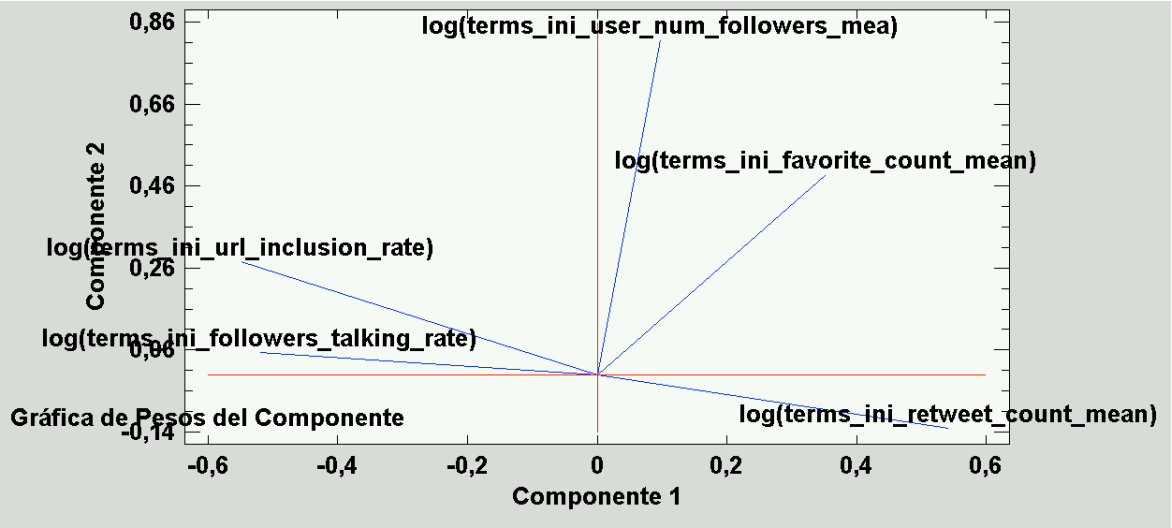


Figura 231. Cine: Gráfica de pesos de cada componente principal

En la Tabla 62 se puede comprobar que todas las variables tienen una presencia significativa en alguno de las principales componentes, por lo que no se puede eliminar ninguna de las variables originales mediante esta técnica.

A continuación, se van a analizar todos los artículos de la categoría Cine, de manera que se pueda comprobar si las características de los datos y las ecuaciones de predicción varían según la sección a la que pertenezca el artículo.

6.1.1.3. Regresión lineal múltiple

Para estudiar la posible relación entre las variables independientes de predicción de que disponemos y cada variable dependiente de éxito o, dicho de otro modo, para tratar de predecir el cálculo de estas, vamos a realizar un modelo de regresión múltiple.

Para realizar las regresiones múltiples, se cuenta con la Tabla *63* de variables resultante de todos los análisis anteriores:

Tabla 63

Cine: Lista final de variables de predicción y de éxito para la regresión lineal múltiple

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	log(favorite_count_mean)
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

Las variables de predicción responden a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos aplicados a la tendencia el día de la publicación del artículo.

La lista de variables de éxito está formada por la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

a) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión múltiple con la variable dependiente log(uniquepageviews_total). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 64

Cine: Valor-P de las variables de la regresión múltiple de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	2,84968	0,0002
log(terms_ini_retweet_count_mean)	-0,0197352	0,6924

log(terms_ini_favorite_count_mean)	0,100637	0,4534
log(terms_ini_followers_talking_rate)	-0,155448	0,1335
log(terms_ini_user_num_followers_mean)	-0,0699937	0,3415
log(terms_ini_url_inclusion_rate)	-0,208117	0,3948
Modelo		0,149

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 65

Cine: Valor-P de las variables de la regresión múltiple simplificada de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	2,25485	0
log(terms_ini_followers_talking_rate)	-0,193301	0,0223
Modelo		0,0223

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(2,25485 - 0,193301 * log(terms_ini_followers_talking_rate))

Para realizar el cálculo de uniquepageviews_total será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

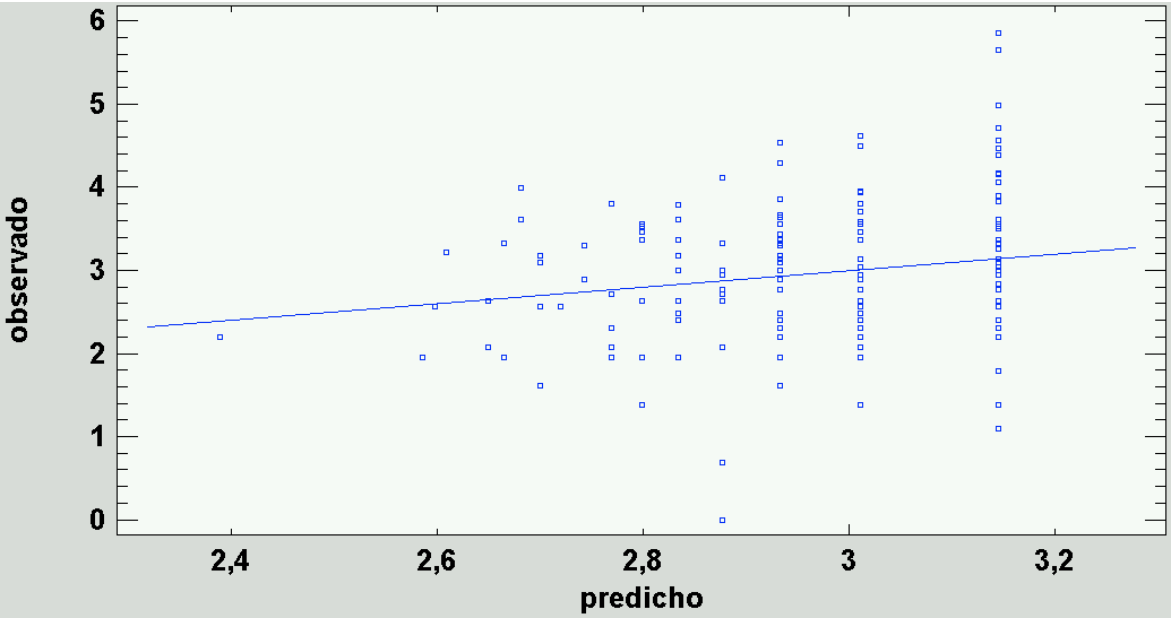


Figura 232. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 3,26236% de la variabilidad de log(uniquepageviews_total), mientras que el R-Cuadrado ajustado indica un 2,65009%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

b) Duración de la visita (promedio)

Para tratar de predecir el valor de la duración de la visita (promedio) es necesario realizar la regresión múltiple con la variable dependiente log(avgtimeonpage_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 66

Cine: Valor-P de las variables de la regresión múltiple de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	5,77747	0
log(terms_ini_retweet_count_mean)	0,0236996	0,679

log(terms_ini_favorite_count_mean)	-0,0406008	0,7875
log(terms_ini_followers_talking_rate)	0,0517285	0,6595
log(terms_ini_user_num_followers_mean)	-0,147131	0,077
log(terms_ini_url_inclusion_rate)	-0,25415	0,3549
Modelo		0,2421

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 67

Cine: Valor-P de las variables de la regresión múltiple simplificada de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	6,10286	0
log(terms_ini_user_num_followers_mean)	-0,176932	0,006
Modelo		0,006

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

avgtimeonpage_mean = exp(6,10286 - 0,176932 * log(terms_ini_user_num_followers_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

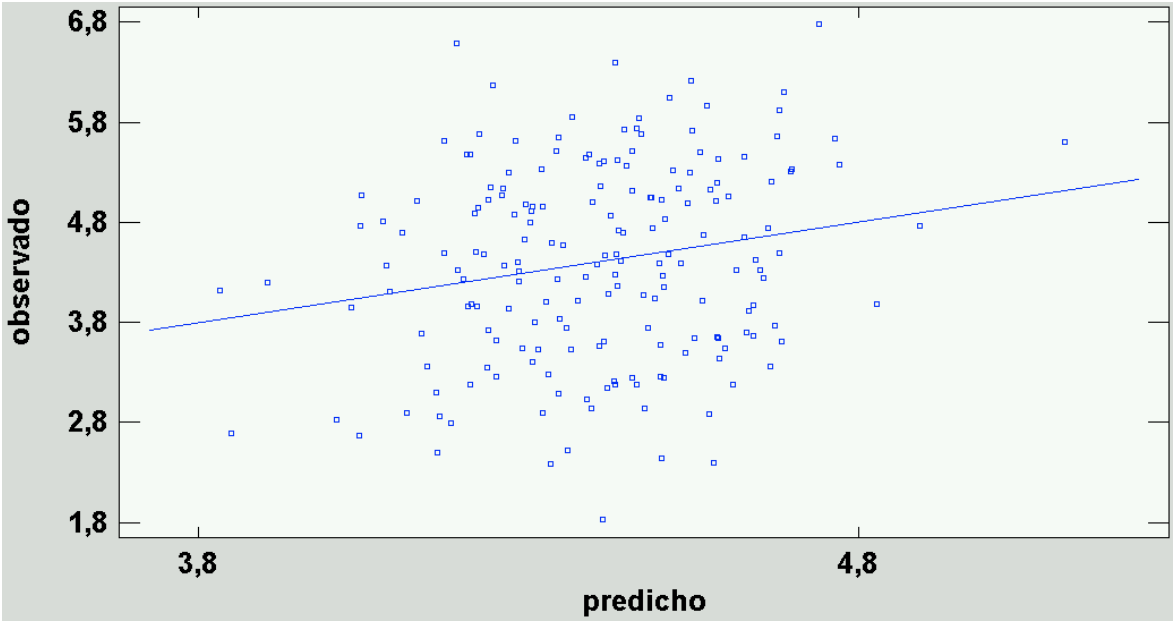


Figura 233. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 4,02429% de la variabilidad de log(avgtimeonpage_mean), mientras que el R-Cuadrado ajustado indica un 3,50268%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Páginas vistas por sesión (promedio)

Para tratar de predecir el valor de las páginas vistas por sesión (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(pageviewspersession_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 68

Cine: Valor-P de las variables de la regresión múltiple de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	1,26813	0,0012
log(terms_ini_retweet_count_mean)	0,0247658	0,3308
log(terms_ini_favorite_count_mean)	-0,0997969	0,1458

log(terms_ini_followers_talking_rate)	0,0940163	0,0755
log(terms_ini_user_num_followers_mean)	-0,0621391	0,0987
log(terms_ini_url_inclusion_rate)	0,0431623	0,729
Modelo		0,0248

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 69

Cine: Valor-P de las variables de la regresión múltiple simplificada de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	1,28579	0,0005
log(terms_ini_followers_talking_rate	0,0925889	0,0331
log(terms_ini_user_num_followers_mean)	-0,0710515	0,0402
Modelo		0,0087

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

pageviewspersession_mean = exp(1,28579 + 0,0925889 * log(terms_ini_followers_talking_rate) - 0,0710515 * log(terms_ini_user_num_followers_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

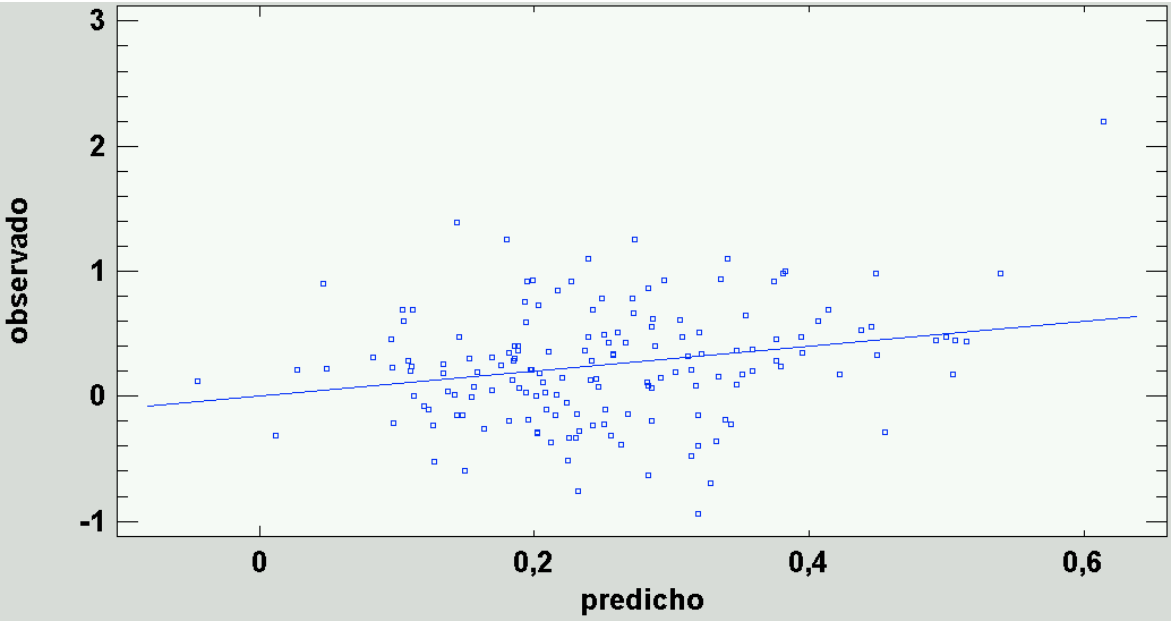


Figura 234. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 5,86202% de la variabilidad de log(pageviewspersession_mean), mientras que el R-Cuadrado ajustado indica un 4,66281%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

d) Nº de favoritos en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de favoritos en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(favorite_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 70

Cine: Valor-P de las variables de la regresión múltiple de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	-0,0227536	0,9607
log(terms_ini_retweet_count_mean)	-0,0370613	0,247

log(terms_ini_favorite_count_mean)	0,000262135	0,9976
log(terms_ini_followers_talking_rate)	-0,17636	0,0104
log(terms_ini_user_num_followers_mean)	-0,0266116	0,5459
log(terms_ini_url_inclusion_rate)	0,0278405	0,8587
Modelo		0,1717

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 71

Cine: Valor-P de las variables de la regresión múltiple simplificada de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	-0,224108	0,2603
log(terms_ini_followers_talking_rate)	-0,120241	0,0223
Modelo		0,0223

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

favorite_count_mean = exp(-0,224108 - 0,120241 * log(terms_ini_followers_talking_rate))

Para realizar el cálculo de favorite_count_mean será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

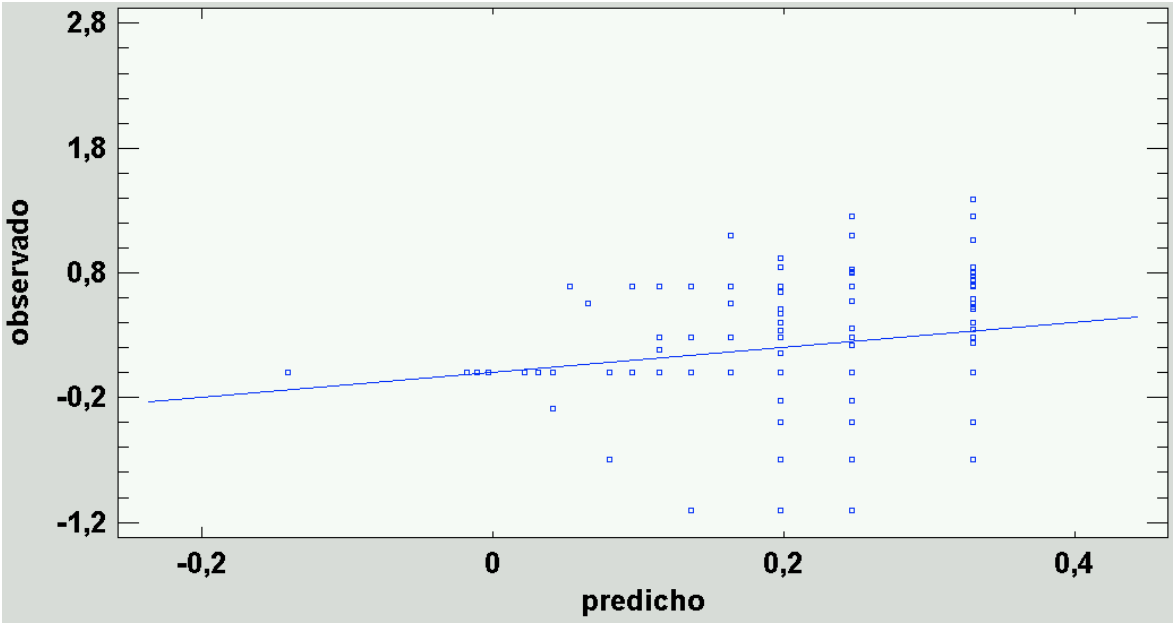


Figura 235. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 3,89283% de la variabilidad de log(favorite_count_mean), mientras que el R-Cuadrado ajustado indica un 3,16475%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Nº de retuits de la tendencia 14 días después (promedio)

Para tratar de predecir el valor de el número de retuits de la tendencia 14 días después (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 72

Cine: Valor-P de las variables de la regresión múltiple de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	-3,44216	0,0145
log(terms_ini_retweet_count_mean)	0,392242	0,0002

log(terms_ini_favorite_count_mean)	-0,574456	0,0225
log(terms_ini_followers_talking_rate)	-0,481299	0,0136
log(terms_ini_user_num_followers_mean)	0,314192	0,0192
log(terms_ini_url_inclusion_rate)	-0,802128	0,0807
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 73

Cine: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	-0,677982	0,2792
log(terms_ini_retweet_count_mean)	0,470587	0
log(terms_ini_followers_talking_rate)	-0,520687	0,0079
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_mean = exp(-0,677982 + 0,470587 * log(terms_ini_retweet_count_mean) - 0,520687 * log(terms_ini_followers_talking_rate))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

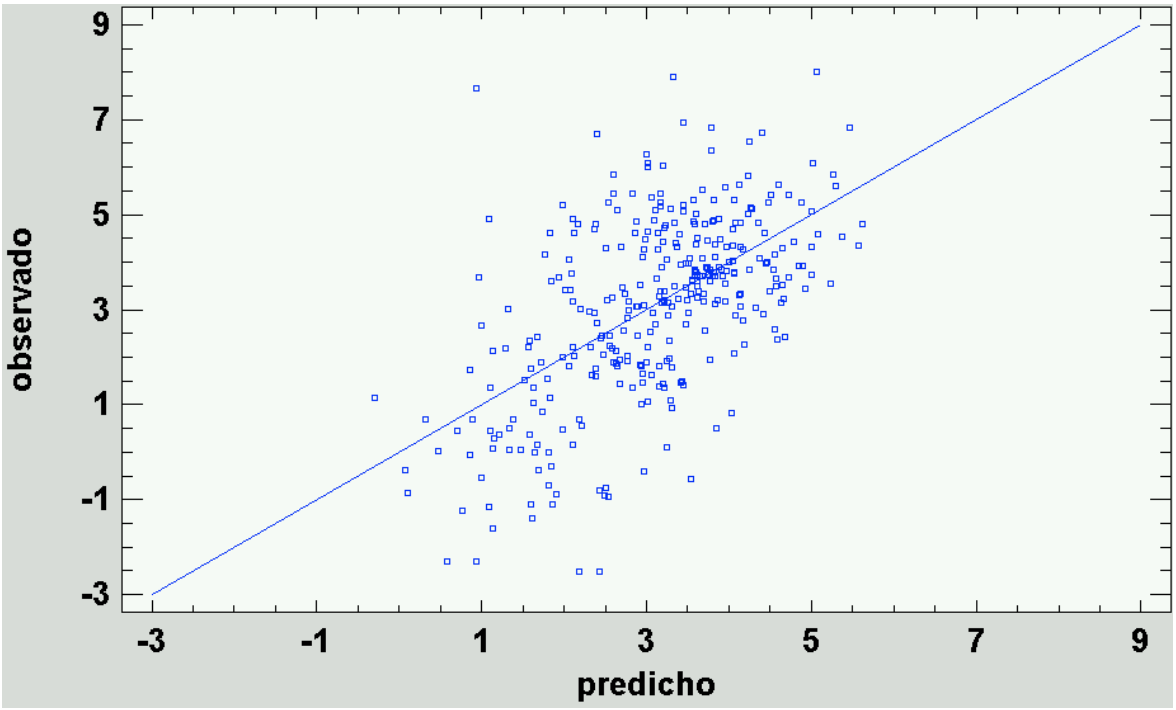


Figura 236. Cine: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 33,3411% de la variabilidad de log(terms_end_retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 32,4523%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen más la línea que en las regresiones lineales múltiples anteriores.

f) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones lineales múltiples de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 74

Cine: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones lineales múltiples

Variables de predicción	Variables de éxito
terms_ini_retweet_count_mean	uniquepageviews_total
terms_ini_favorite_count_mean	avgtimeonpage_mean
terms_ini_followers_talking_rate	pageviewspersession_mean
terms_ini_user_num_followers_mean	favorite_count_mean
terms_ini_url_inclusion_rate	terms_end_retweet_count_mean

Se puede observar en la Tabla *74* que todas las variables de predicción salvo terms_ini_url_inclusion_rate participan en alguna de las ecuaciones de predicción, por lo que todas menos la previamente mencionada son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- La ratio de seguidores del medio que hablan de la tendencia explica parte de los datos de páginas vistas únicas.
- La ratio de seguidores del medio que hablan de la tendencia explica parte de los datos de la duración de la visita.
- La ratio de seguidores del medio que hablan de la tendencia y el número de seguidores de los usuarios que participan en la tendencia explican parte de los datos de la media de páginas vistas por sesión.
- La ratio de seguidores del medio que hablan de la tendencia explica parte de los datos del número medio de favoritos en la cuenta del medio.
- El promedio de retuits y la ratio de seguidores del medio que hablan de la tendencia explican en parte el promedio de retuits 14 días después.
- La ratio de seguidores del medio que hablan de la tendencia, por tanto, participa en la predicción de todas las variables de éxito salvo la duración promedio de la visita.

6.1.1.4. Regresión binomial negativa o de Poisson

A continuación, se tratará de predecir todas las variables de éxito que sean de conteo (enteros y sin números negativos) a partir de todas las variables de predicción que sean independientes entre sí según la regresión binomial negativa o la regresión de Poisson.

a) Filtro de alta correlación (colinealidad)

Las variables que aporten información para tratar de realizar la regresión deben ser independientes, motivo por el cual es necesario hacer un filtro de alta correlación de manera que se asegure que todas aportan información diferente.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

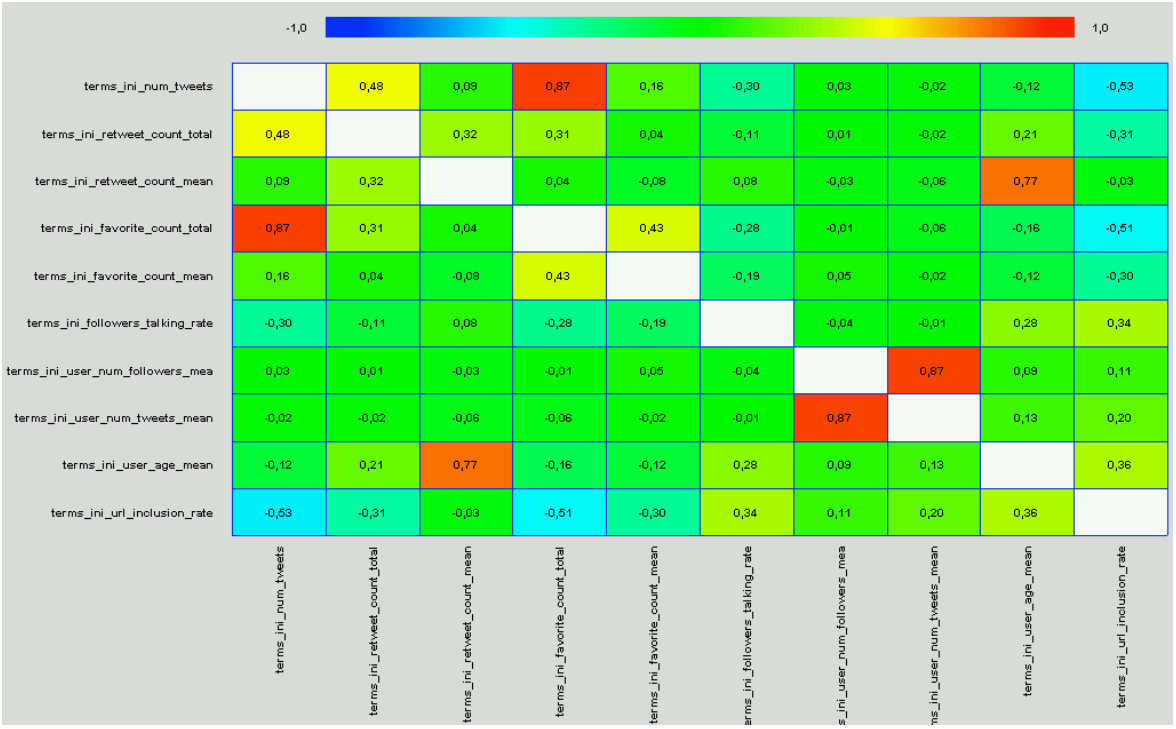


Figura 237. Cine: Matriz de correlaciones Pearson entre las variables de predicción

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_ini_num_tweets y terms_ini_favorite_count_total tienen un coeficiente de correlación de 0,8695 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

- terms_ini_retweet_count_mean y terms_ini_user_age_mean tienen un coeficiente de correlación de 0,7705 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- tems_ini_user_num_followers_mean y terms_ini_user_num_tweets_mean tienen un coeficiente de correlación de 0,8658 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige terms_ini_num_tweets, terms_ini_user_age_mean y tems_ini_user_num_followers_mean por tener un sesgo y una curtosis estandarizados menores, como se puede comprobar en el anexo 6.1.1.2.

La tabla de variables quedaría como sigue:

Tabla 75

Cine: Lista de variables de predicción y de éxito para la regresión binomial negativa o de Poisson tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a: el número total de tuits, el número total de retuits, el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de edad en días de la cuenta de los usuarios que participan y la

ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

b) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión con la variable dependiente uniquepageviews_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable uniquepageviews_total, el Chi-cuadrado calculado es 1.387.800 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 76

Cine: Valor-P de las variables de la regresión binomial negativa de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	1,58892	0
terms_ini_num_tweets	0,000533817	0
terms_ini_retweet_count_total	-0,000000509408	1
terms_ini_favorite_count_mean	0,0894829	1
terms_ini_followers_talking_rate	2,60869	0
terms_ini_user_num_followers_mean	-0,00000599681	0
terms_ini_user_age_mean	0,000194736	1
terms_ini_url_inclusion_rate	-0,188429	1
Modelo		1

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que

0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 77

Cine: Valor-P de las variables de la regresión binomial negativa simplificada de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,27857	0
terms_ini_num_tweets	0,000176761	0
terms_ini_user_num_followers_mean	-0,00000157529	0,0014
terms_ini_user_age_mean	-0,000170714	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(3,27857 + 0,000176761 * terms_ini_num_tweets - 0,00000157529 * terms_ini_user_num_followers_mean - 0,000170714 * terms_ini_user_age_mean)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

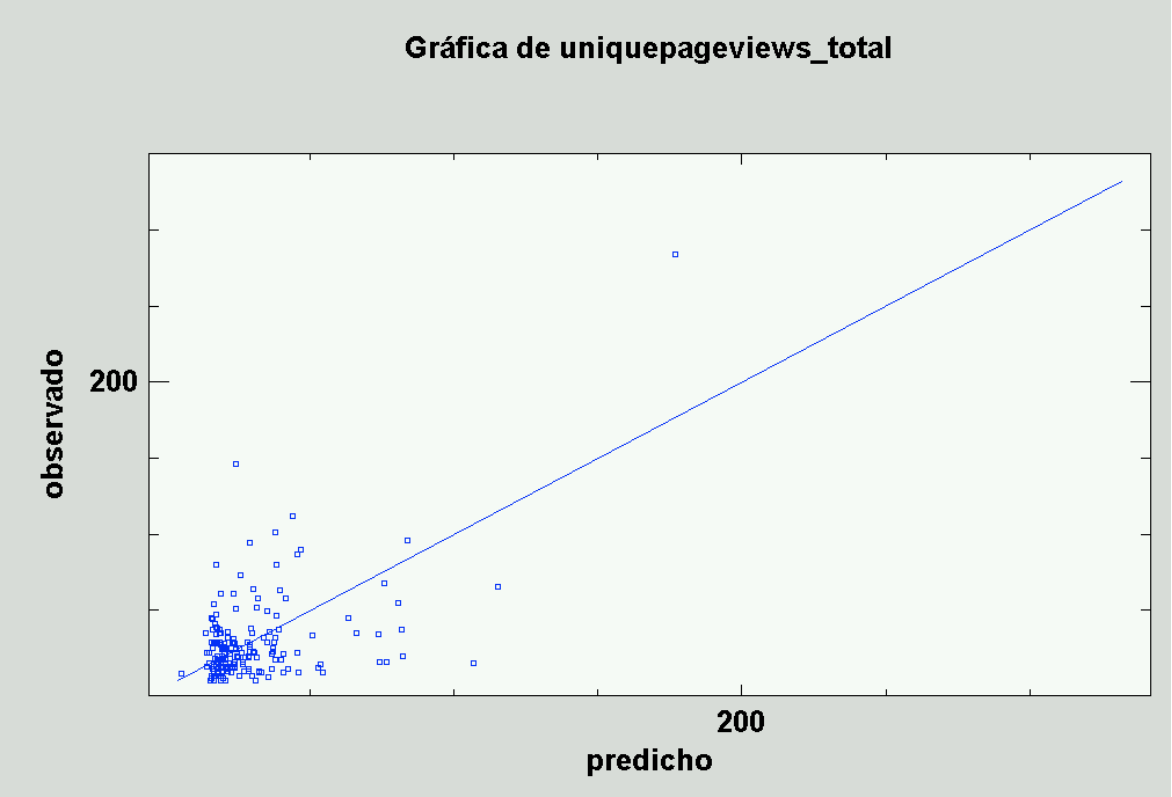


Figura 238. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 6,54364% de la variabilidad de uniquepageviews_total, mientras que el R-Cuadrado ajustado indica un 6,13461%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_num_tweets, el Chi-cuadrado calculado es 2.367.740.000.000.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 78

Cine: Valor-P de las variables de la regresión binomial negativa de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	18,8005	0
terms_ini_num_tweets	0,000248719	0,0003
terms_ini_retweet_count_total	-0,0000000455589	1
terms_ini_favorite_count_mean	-0,00315402	0,8823
terms_ini_followers_talking_rate	-1,12477	0,6697
terms_ini_user_num_followers_mean	-0,0000027329	1
terms_ini_user_age_mean	-0,0000913075	0,6841
terms_ini_url_inclusion_rate	-1,10704	0,2694
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 79

Cine: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	17,967	0
terms_ini_num_tweets	0,000268319	0

Modelo	0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(17,967 + 0,000268319 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

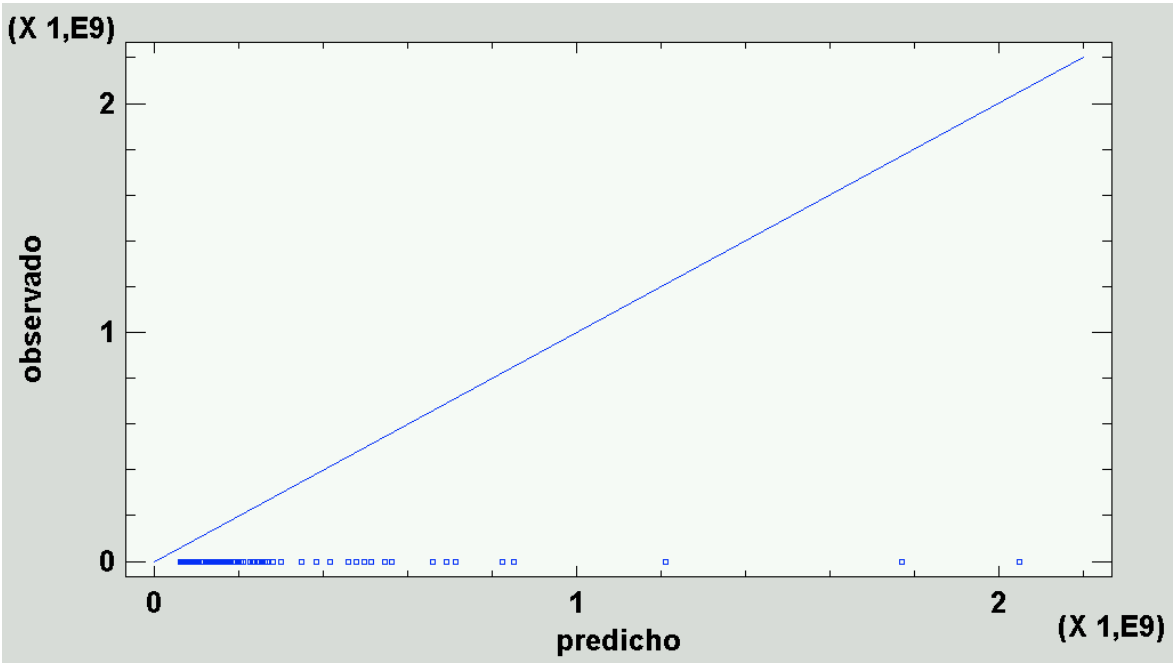


Figura 239. Cine: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 9,05712% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 7,97882%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

d) Nº de retuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de retuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_retweet_count_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_retweet_count_total, el Chi-cuadrado calculado es 2.367.740.000.000.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 80

Cine: Valor-P de las variables de la regresión binomial negativa de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	34,406	0
terms_ini_num_tweets	0,000767186	0
terms_ini_retweet_count_total	-0,00000029225	0
terms_ini_favorite_count_mean	-0,775262	0
terms_ini_followers_talking_rate	-26,1508	1
terms_ini_user_num_followers_mean	-0,0000827386	1
terms_ini_user_age_mean	-0,00144072	1
terms_ini_url_inclusion_rate	11,4398	0
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 81

Cine: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	22,9283	0
terms_ini_num_tweets	0,00126057	0
terms_ini_favorite_count_mean	-0,813487	0,0001
terms_ini_url_inclusion_rate	21,9201	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_total = exp(22,9283 + 0,00126057 * terms_ini_num_tweets - 0,813487 * terms_ini_favorite_count_mean + 21,9201 * terms_ini_url_inclusion_rate)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

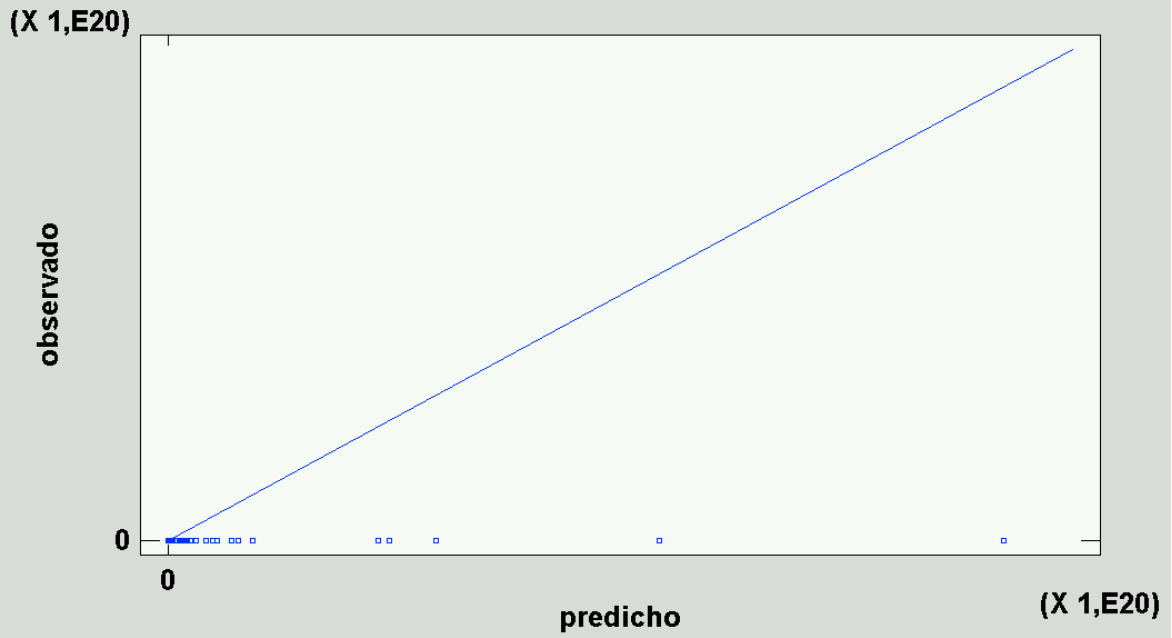


Figura 240. Cine: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 20,0829% de la variabilidad de terms_end_retweet_count_total, mientras que el R-Cuadrado ajustado indica un 17,6913%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones binomiales negativas de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 82

Cine: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones binomiales negativas

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

Se puede observar en la

Tabla *82* que no todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que no todas son necesarias para las variables de éxito elegidas y que se pueden estudiar. Es el caso del total de retuits y la ratio de seguidores de la cuenta del medio que participan en la tendencia.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits, el promedio de seguidores y el promedio de la edad en días de los usuarios que participan en la tendencia explican parte de los datos de páginas vistas únicas.
- El número de tuits explica parte de los datos del número de tuits 14 días después.

El número de tuits, el número de retuits, el promedio de favoritos y la ratio de inclusión de URL en los tuits explican en parte el número de retuits 14 días después.

6.1.2. Análisis de los artículos de la categoría Series

6.1.2.1. Variables de éxito

El objetivo de esta fase es la predicción de los valores de éxito mediante el resto de los indicadores. Por ello, es importante comenzar por analizar estos escenarios por separado para ver tanto sus características como tratar de describir su comportamiento anómalo, si lo tuvieran.

a) Páginas vistas únicas (total)

Este escenario de éxito se ve identificado por la columna uniquepageviews_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 104 valores con un rango de entre 3 y 704.

Presenta un sesgo estandarizado de 21,9251 y una curtosis estandarizada de 67,0581. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

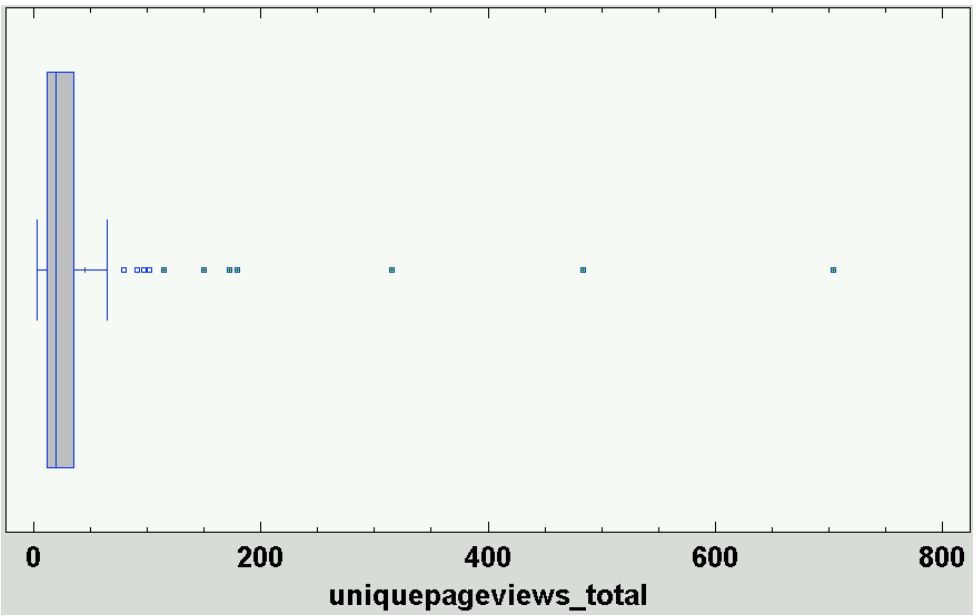


Figura 241. Series: Gráfico de Caja y Bigotes para el valor uniquepageviews_total

En la Figura 1 se puede comprobar que existen valores anómalos de tipo extremo de 115 páginas vistas o más.

b) AdSense eCPM (promedio)

Este escenario de éxito se identifica con la columna adsense_ecpm_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 0,47.

Presenta un sesgo estandarizado de 14,4467 y una curtosis estandarizada de 30,1185. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

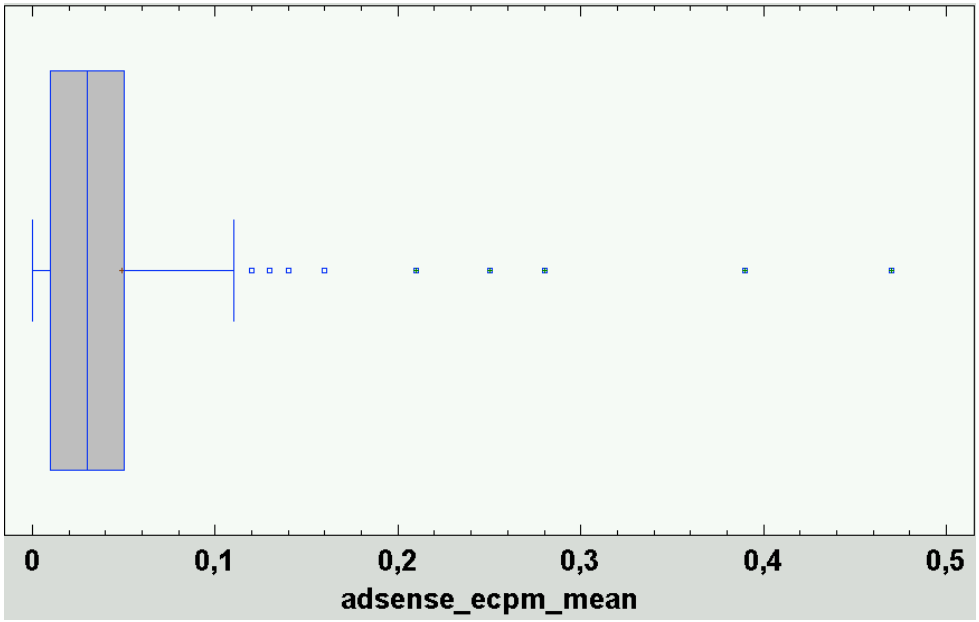


Figura 242. Series: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean

En la Figura 242 se puede observar que existen valores anómalos de tipo extremo de 0,21 o más.

c) Duración de la visita (promedio)

Este escenario de éxito se identifica con la columna avgtimeonpage_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 673.

Presenta un sesgo estandarizado de 5,9847 y una curtosis estandarizada de 3,63637. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

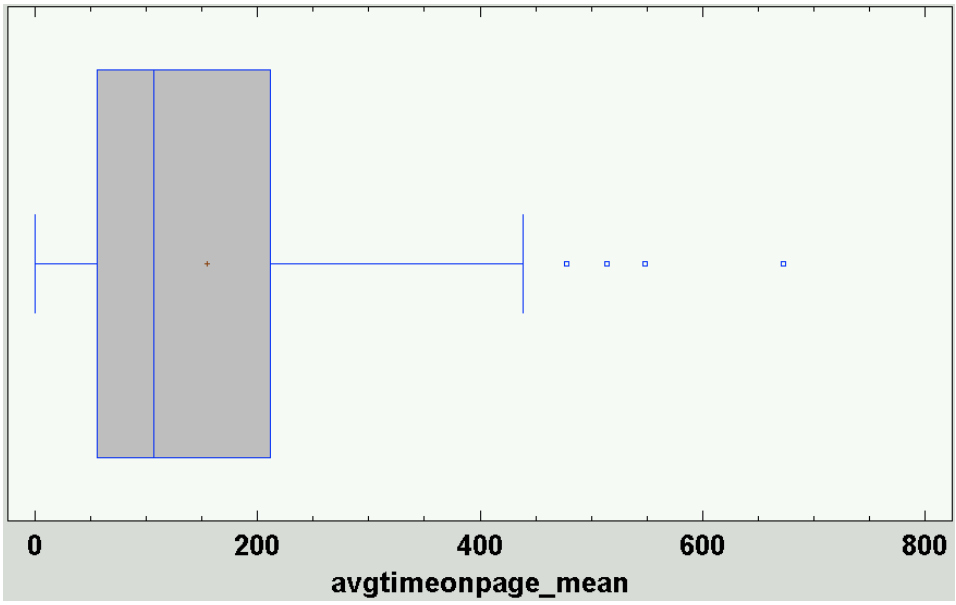


Figura 243. Series: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean

En la Figura 243 se puede observar que no existen valores anómalos de tipo extremo.

d) Páginas vistas por sesión (promedio)

Este escenario de éxito se identifica con la columna pageviewspersession_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0,39 y 6.

Presenta un sesgo estandarizado de 14,6631 y una curtosis estandarizada de 35,2043 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

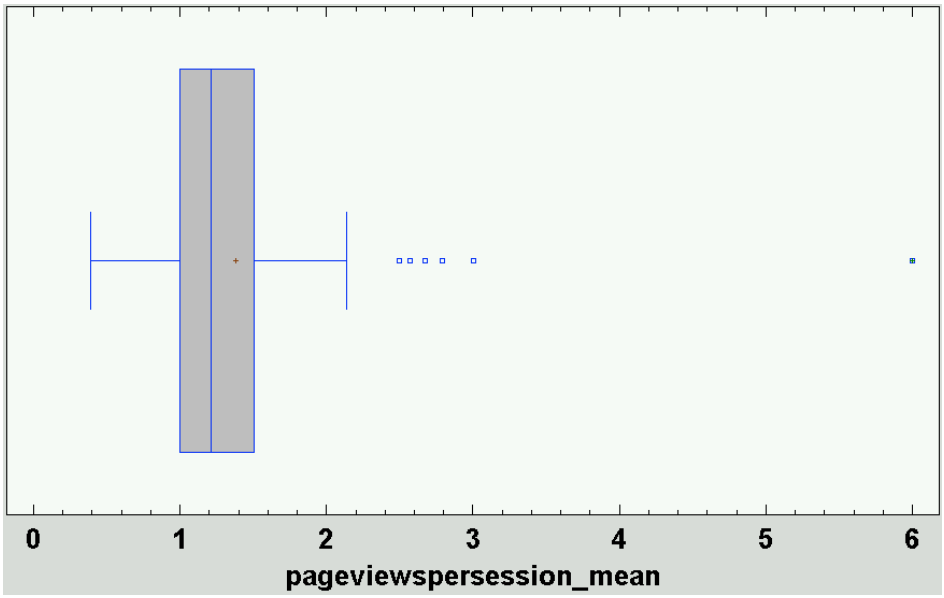


Figura 244. Series: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean

En la Figura 244 se puede observar que existe un valor anómalo de tipo extremo de 6.

e) Nº de retuits en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 2.

Presenta un sesgo estandarizado de -0,408634 y una curtosis estandarizada de 0,695636 Puesto que ambos valores estadísticos están dentro del rango de -2 a +2, indican que siguen una normalidad, es decir, que ambos se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

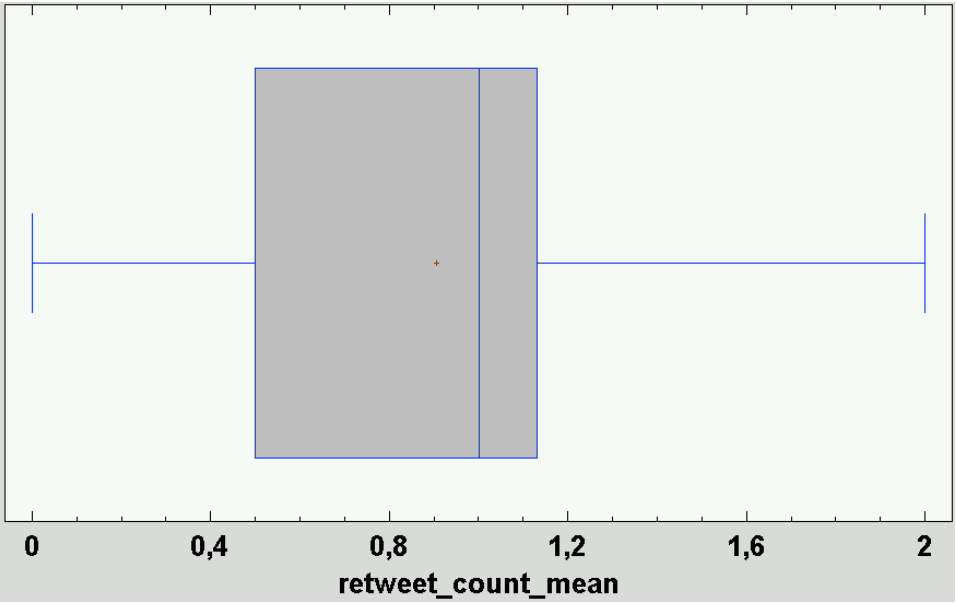


Figura 245. Series: Gráfico de Caja y Bigotes para el valor retweet_count_mean

En la Figura 245 no se observa ningún valor anómalo de tipo extremo.

f) Nº de favoritos en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 5.

Presenta un sesgo estandarizado de 3,40176 y una curtosis estandarizada de 3,49466. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

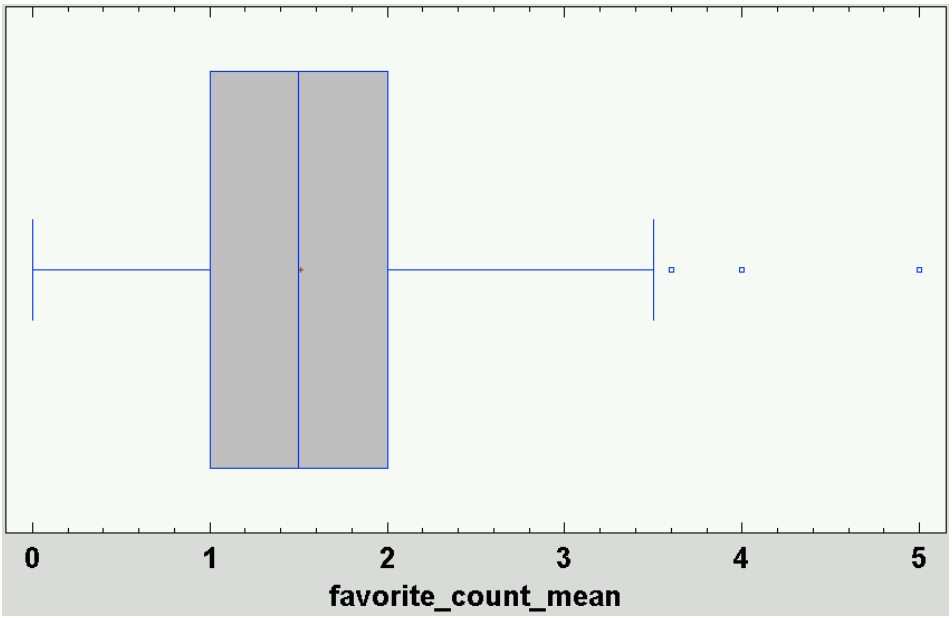


Figura 246. Series: Gráfico de Caja y Bigotes para el valor favorite_count_mean

En la Figura 246 no se observa ningún valor anómalo de tipo extremo.

g) Nº de tuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 104 valores con un rango de entre 0 y 8.667.

Presenta un sesgo estandarizado de 4,02492 y una curtosis estandarizada de 0,324223 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

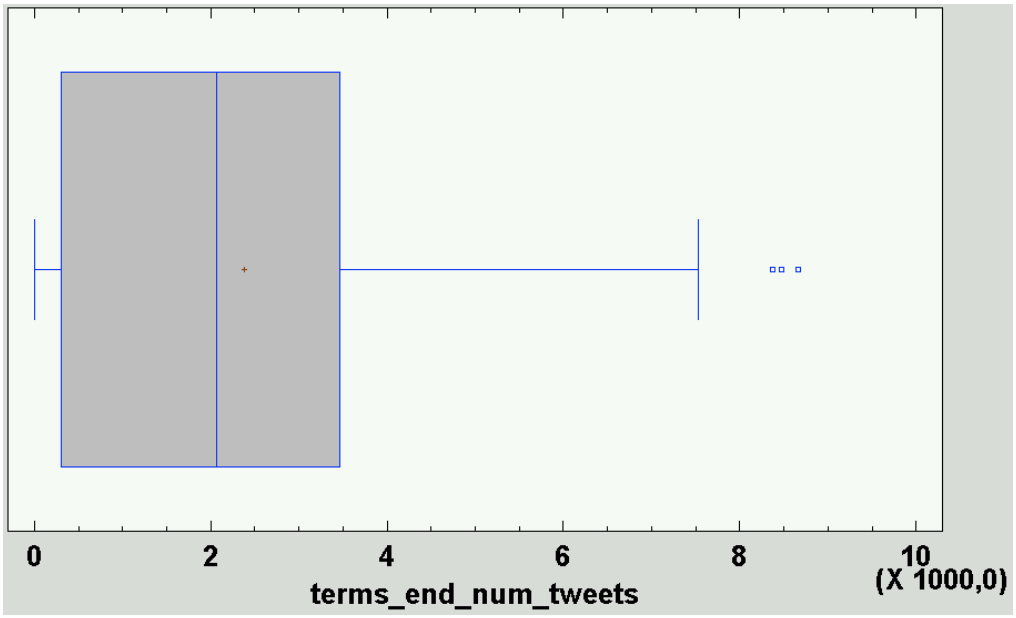


Figura 247. Series: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets

En la Figura 247 no se puede observar ningún valor anómalo de tipo extremo.

h) Nº de retuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 4.941.200.

Presenta un sesgo estandarizado de 14,8088 y una curtosis estandarizada de 31,4162. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

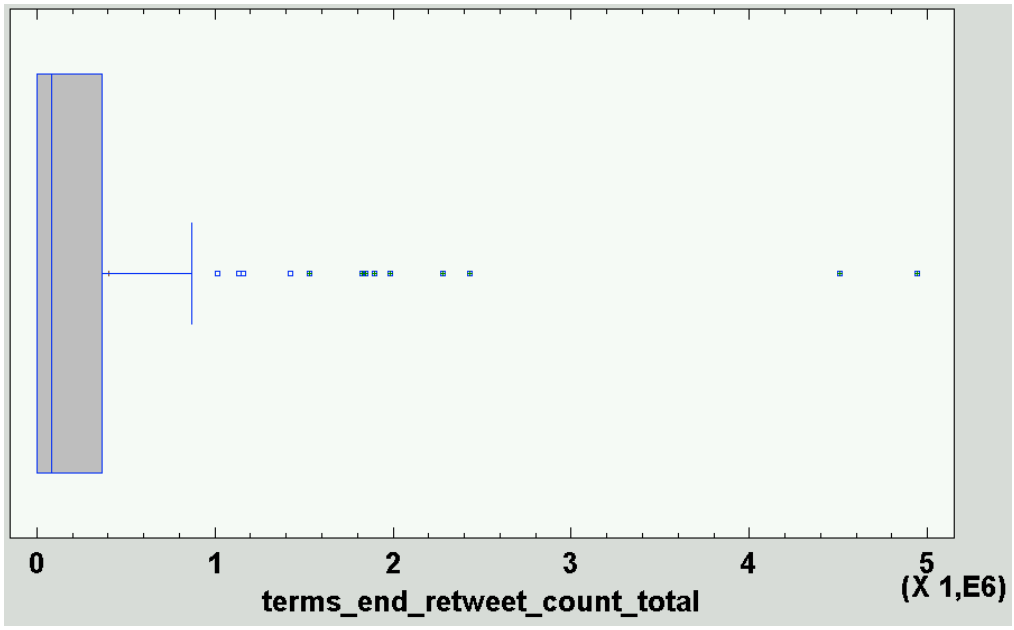


Figura 248. Series: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total

En la Figura 248 se puede observar que existen valores anómalos de tipo extremo de 1.531.660 o más.

i) Nº de retuits de la tendencia 14 días después (promedio)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 1.041,09.

Presenta un sesgo estandarizado de 18,8859 y una curtosis estandarizada de 45,8353. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

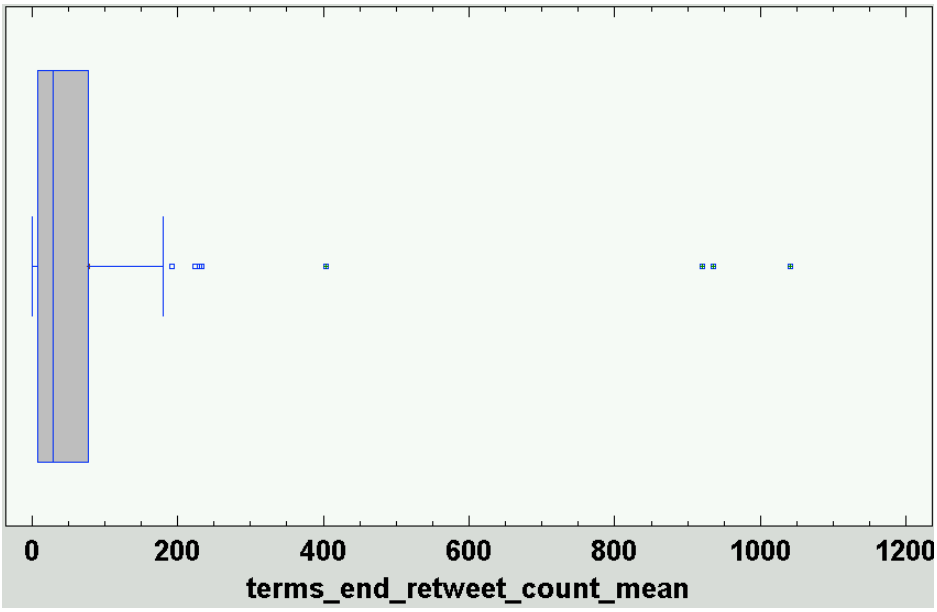


Figura 249. Series: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean

En la Figura 249 se puede observar que existen valores anómalos de tipo extremo de 403,56 o más.

j) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de las variables de éxito. Se pueden observar los siguientes datos de estas:

Tabla 83

Series: Resumen estadístico de las variables de éxito

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
uniquepageviews_total	8.220,08	21,9251	67,0581
adsense_ecpm_mean	0,00546153	14,4467	30,1185
avgtimeonpage_mean	18.970,3	5,9847	3,63637
pageviewspersession_mean	0,695373	14,6631	35,2043

retweet_count_mean	0,240243	-0,408634	0,695636
favorite_count_mean	0,91253	3,40176	3,49466
terms_end_num_tweets	5.283.050	4,02492	0,324223
terms_end_retweet_count_total	656.324.000.000	14,8088	31,4162
terms_end_retweet_count_mean	28.000,1	18,8859	45,8353

Se puede observar en la Tabla 83 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2 salvo retweet_count_mean, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 84

Series: Resumen estadístico de las variables de éxito con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(uniquepageviews_total)	1,0562	3,19631	2,49584
log(adsense_ecpm_mean)	1,00054	2,11288	-0,643645
log(avgtimeonpage_mean)	0,890564	-1,1642	-0,889792
log(pageviewspersession_mean)	0,205642	2,48803	4,1638

retweet_count_mean	0,240243	-0,408634	0,695636
log(favorite_count_mean)	0,220074	-0,531057	0,886999
log(terms_end_num_tweets)	3,12363	-5,03233	2,710004
log(terms_end_retweet_count_total)	11,5146	-3,54906	-0,127349
log(terms_end_retweet_count_mean)	3,5879	-2,71949	0,395108

Todas las variables salvo log(avgtimeonpage_mean), retweet_count_mean y log(favorite_count_mean) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. Sin embargo, mantienen valores mucho menores, próximos al rango de -2 a +2 y con una dispersión muy parecida, por lo que en este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

Nos quedamos, por tanto, con las variables que tengan menores sesgo y curtosis estandarizados de su forma original o transformación logarítmica.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5, la tabla queda así:

Tabla 85

Series: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	retweet_count_mean
terms_ini_followers_talking_rate	log(favorite_count_mean)

terms_ini_user_num_followers_mean	terms_end_num_tweets
terms_ini_user_num_tweets_mean	log(terms_end_retweet_count_total)
terms_ini_user_age_mean	log(terms_end_retweet_count_mean)
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, por tanto, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio, el número total de retuits de la tendencia 14 días después y el promedio de retuits de la tendencia 14 días después. También incluye la versión original del promedio de retuits en la cuenta del medio y el número de tuits de la tendencia 14 días después.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

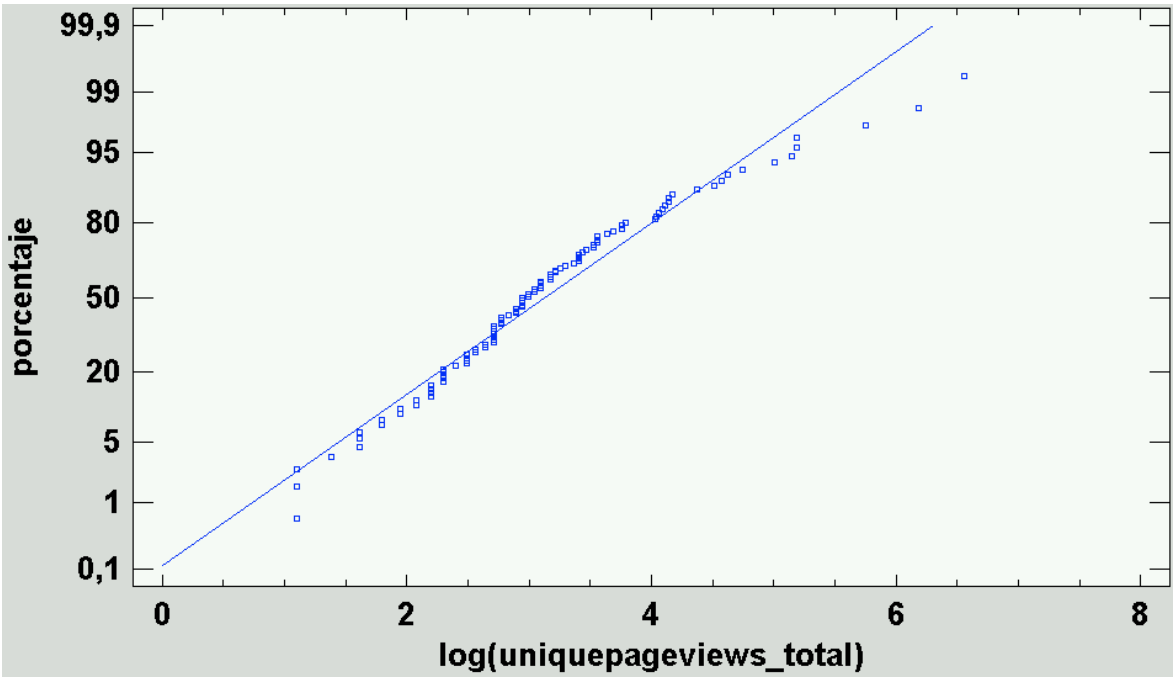


Figura 250. Series: Gráfico de probabilidad normal de la variable log(uniquepageviews_total)

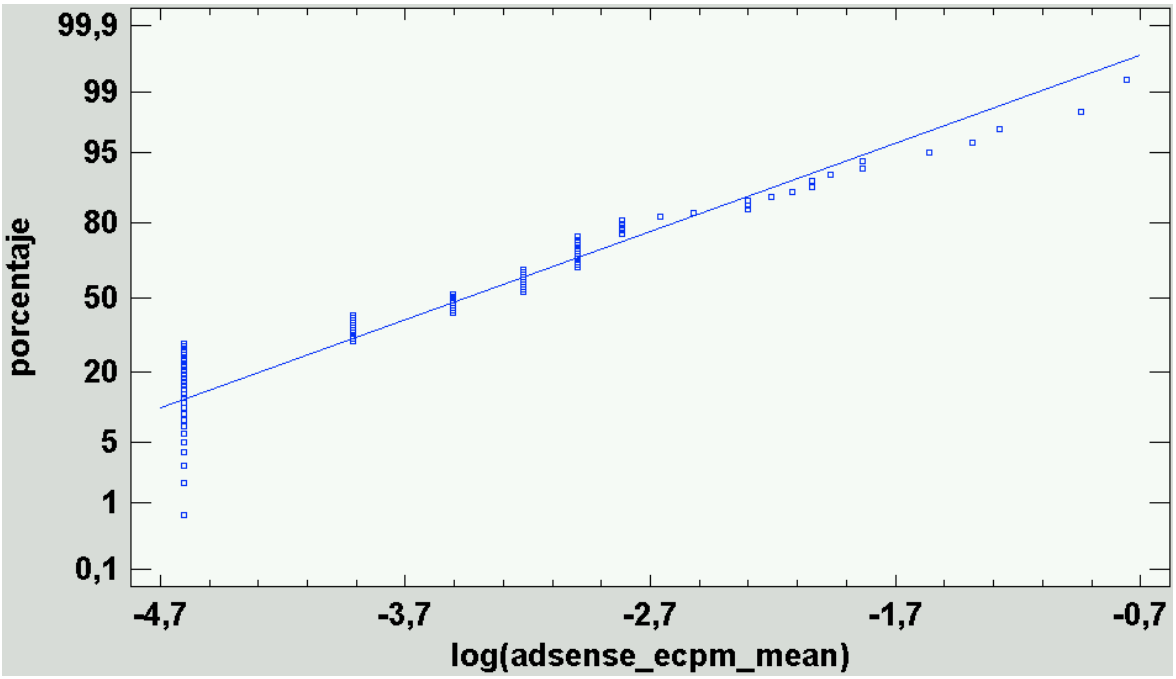


Figura 251. Series: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean)

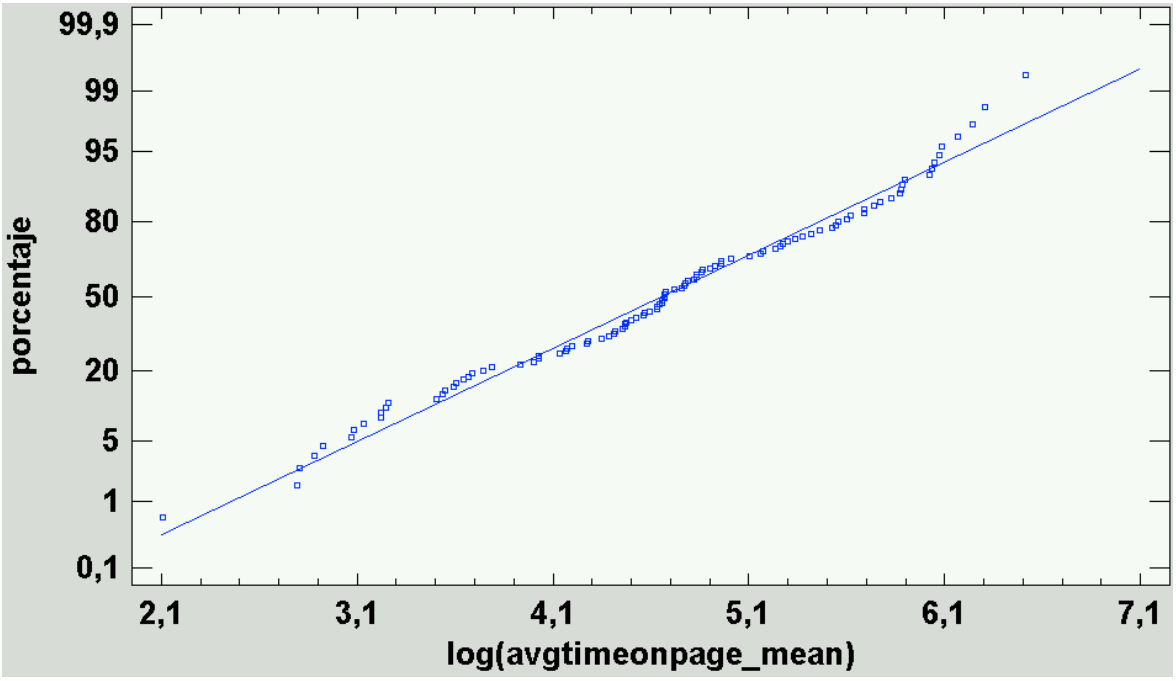


Figura 252. Series: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean)

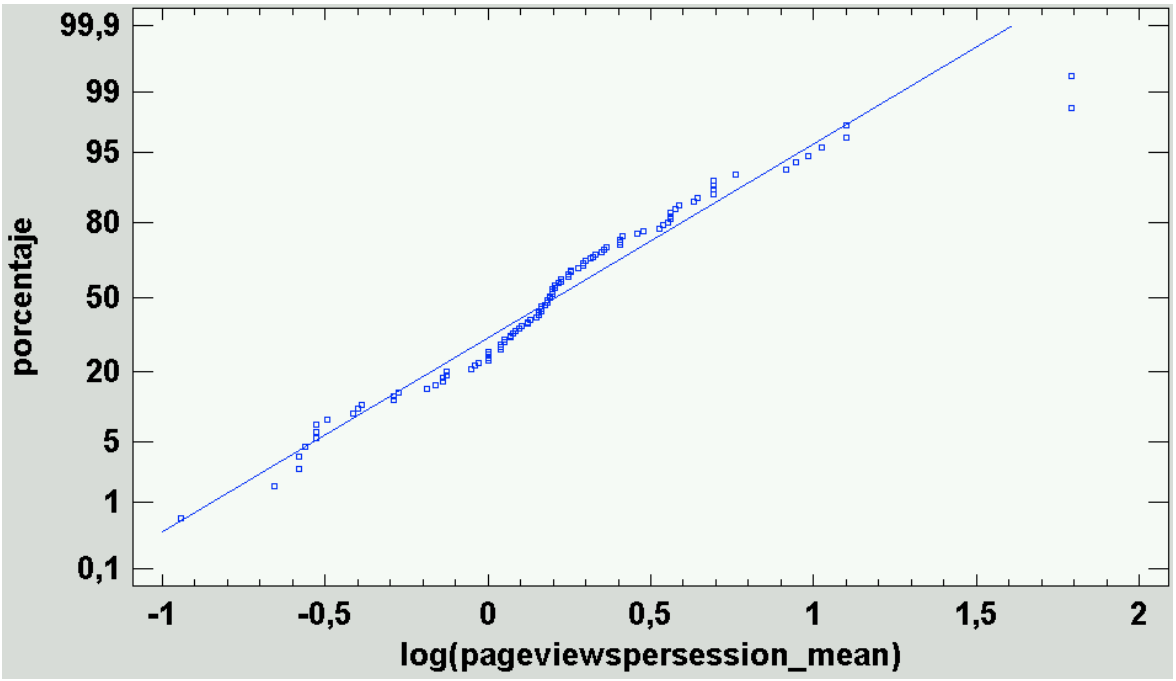


Figura 253. Series: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean)

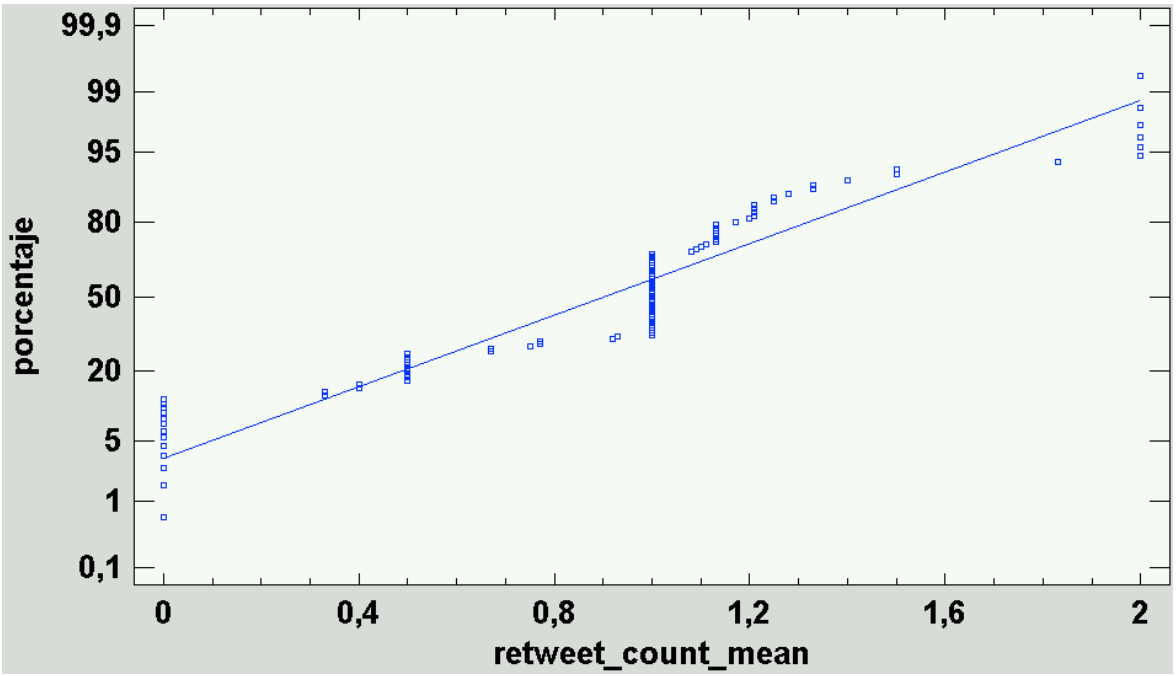


Figura 254. Series: Gráfico de probabilidad normal de la variable retweet_count_mean

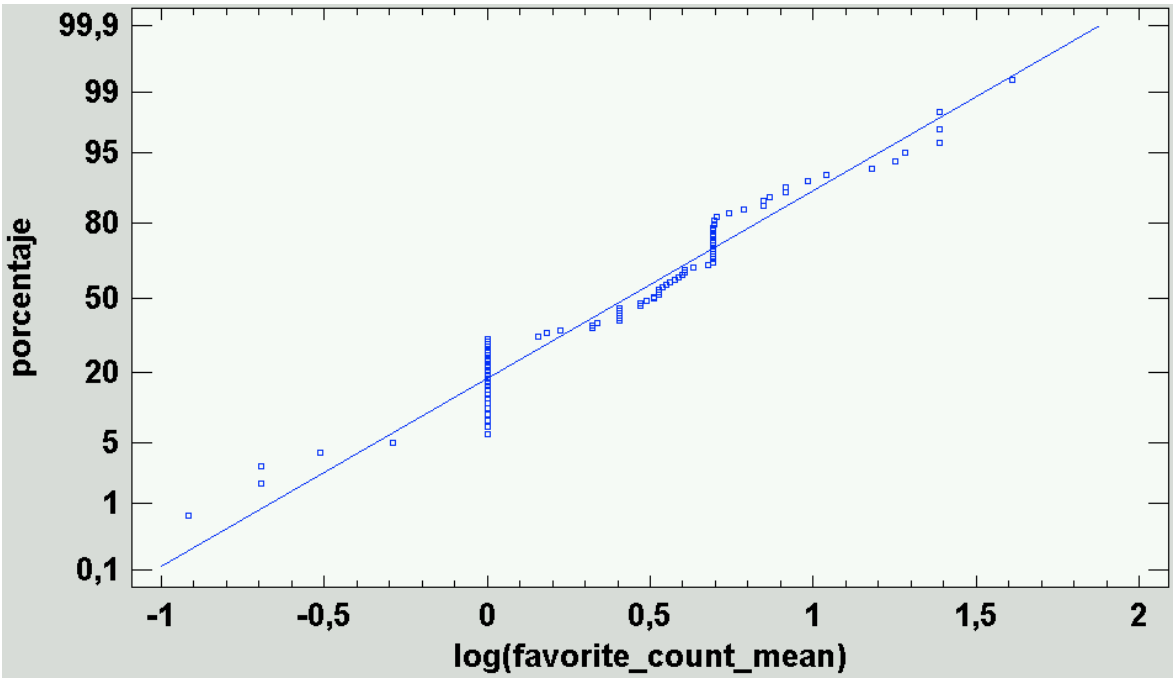


Figura 255. Series: Gráfico de probabilidad normal de la variable log(favorite_count_mean)

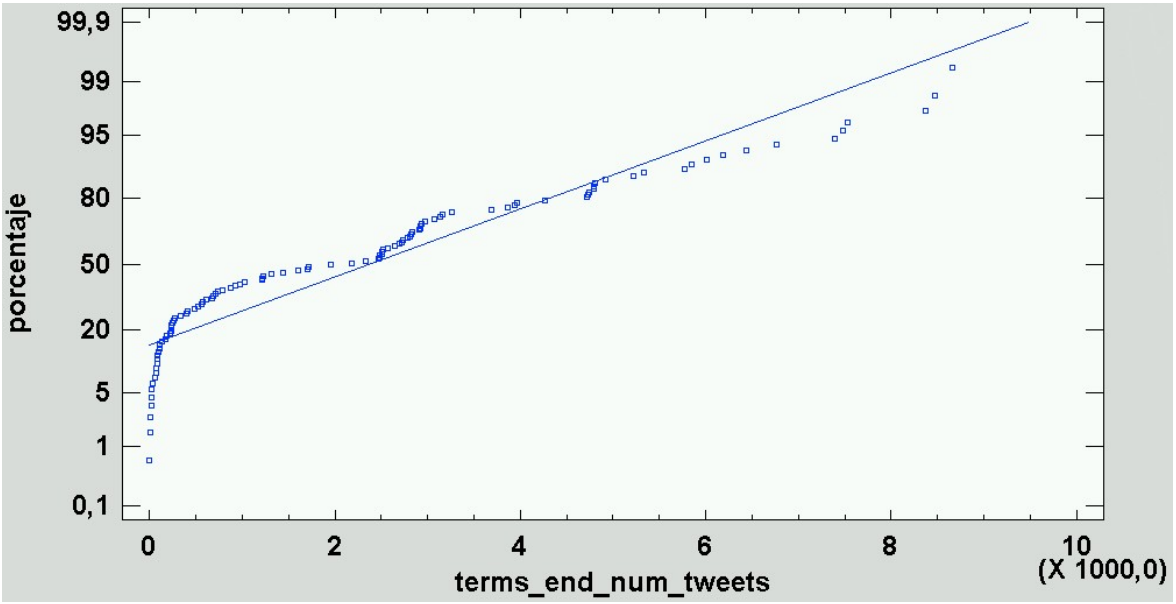


Figura 256. Series: Gráfico de probabilidad normal de la variable terms_end_num_tweets

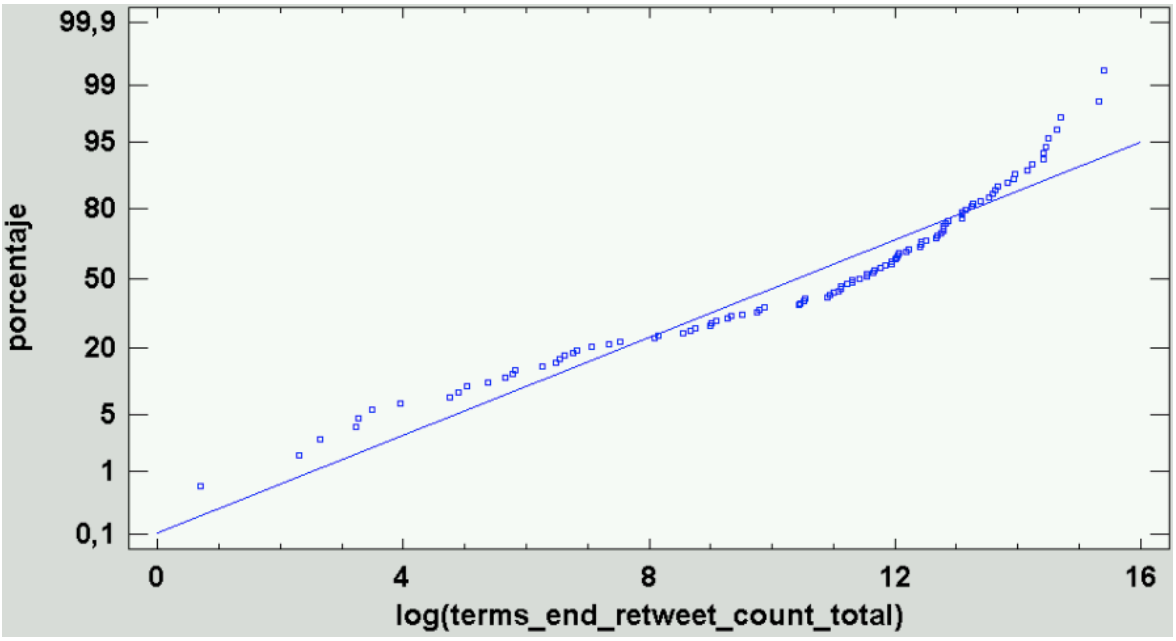


Figura 257. Series: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total)

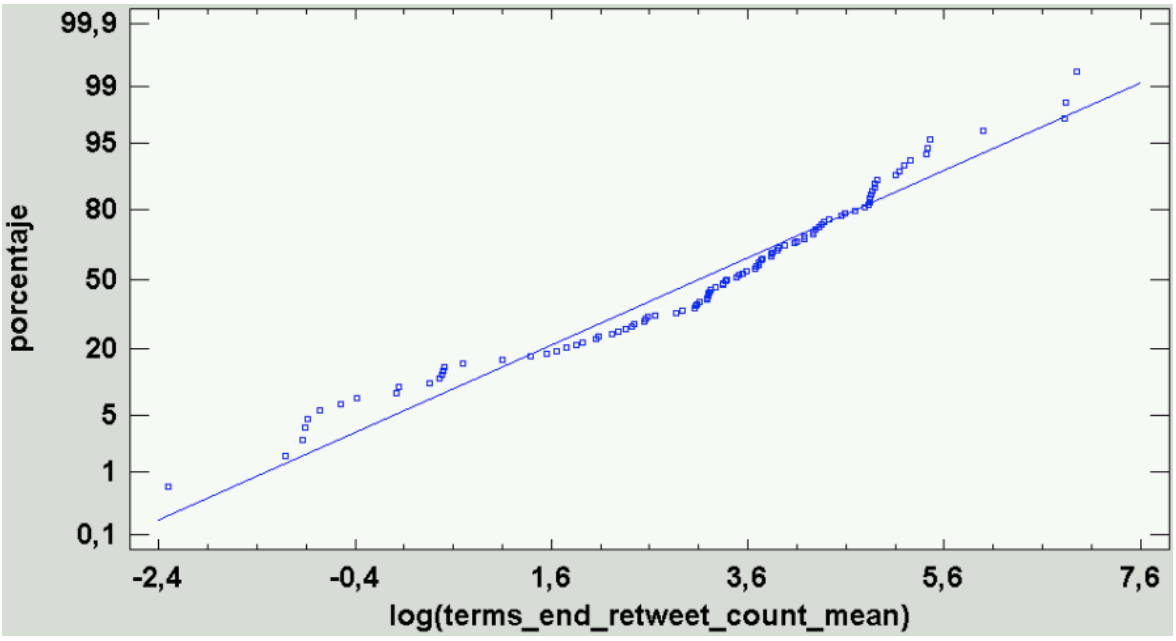


Figura 258. Series: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

k) Filtro de alta correlación

Puesto que se han propuesto más de una variable de éxito, es conveniente evaluar si están asociadas mediante una fuerte correlación. Si fuera ese el caso, el análisis de este estudio se simplificaría ya que las conclusiones de una variable servirían para las que estén

fuertemente correlacionadas con ella, lo que permitiría dedicar menos recursos a la predicción y toma de decisiones.

Para ello, es necesario realizar un análisis multivariado de las variables de éxito con su correspondiente transformación logarítmica, cuya matriz de correlaciones Pearson se puede observar en la Figura 259:

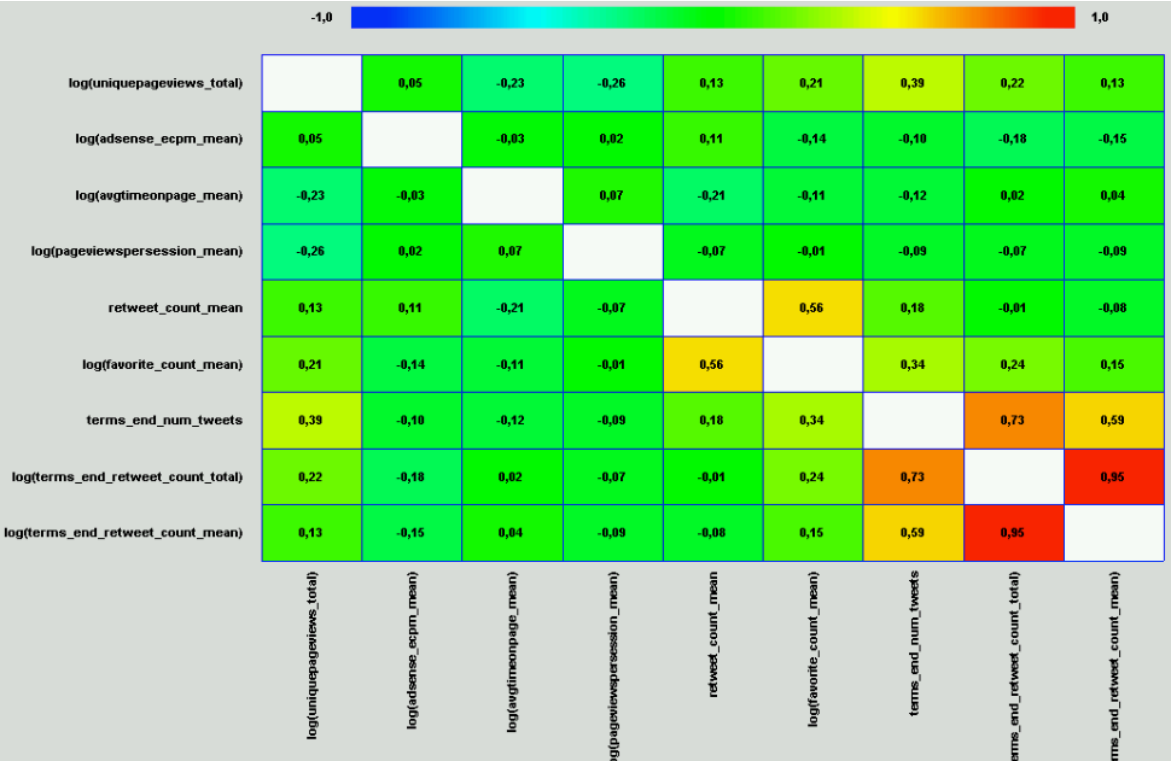


Figura 259. Series: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_end_num_tweets y log(terms_end_retweet_count_total) tienen un coeficiente de correlación de 0,7328 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_end_retweet_count_total) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,9504 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- terms_end_num_tweets y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,5854 y un valor-P cercano a 0, por lo que no existe

una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto terms_end_num_tweets y log(terms_end_retweet_count_mean), porque log(terms_end_retweet_count_total) está fuertemente correlacionada con estas dos, pero entre ellas no lo están.

La tabla de variables quedaría como sigue:

Tabla 86

Series: Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	retweet_count_mean
terms_ini_followers_talking_rate	log(favorite_count_mean)
terms_ini_user_num_followers_mean	terms_end_num_tweets
terms_ini_user_num_tweets_mean	log(terms_end_retweet_count_mean)
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, finalmente, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después. También incluye la versión original del promedio de retuits en la cuenta del medio y el número de tuits de la tendencia 14 días después.

6.1.2.2. Variables de predicción

Se parte de un conjunto de datos con diez características, por lo que es conveniente tratar de reducir la dimensión del conjunto, restableciendo la varianza sin modificar la información relevante de los datos en sí. Esto posibilitará que se reduzca el tiempo y el coste de la computación y facilita la visualización y el análisis de los datos. Además, es una condición necesaria para aplicar la regresión lineal múltiple (Anon., 2017).

a) Número de tuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 104 valores con un rango de entre 0 y 9.849.

Presenta un sesgo estandarizado de 3,45941 y una curtosis estandarizada de 0,0195947. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

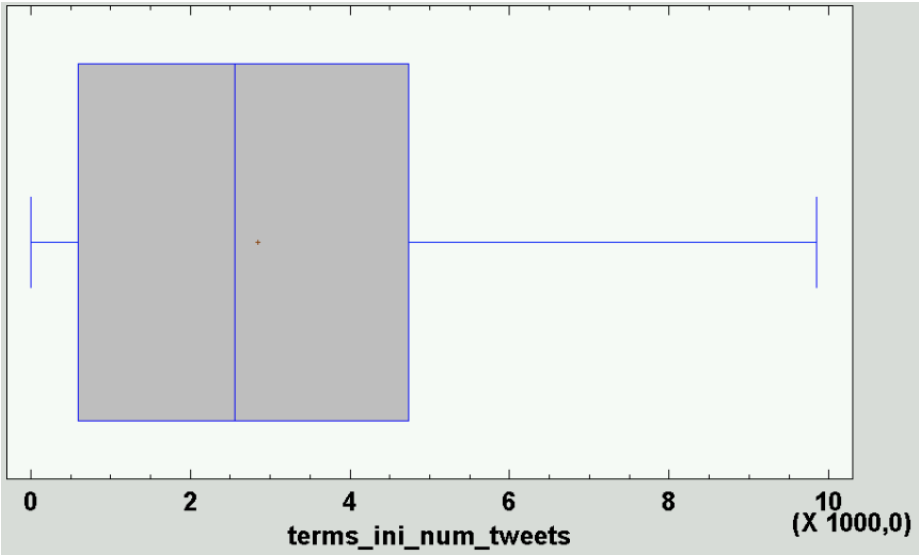


Figura 260. Series: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets

En la Figura 260 se puede observar que no existen valores anómalos de tipo extremo.

b) Número de retuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 104 valores con un rango de entre 0 y 9.805.270.

Presenta un sesgo estandarizado de 19,6336 y una curtosis estandarizada de 57,4145. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

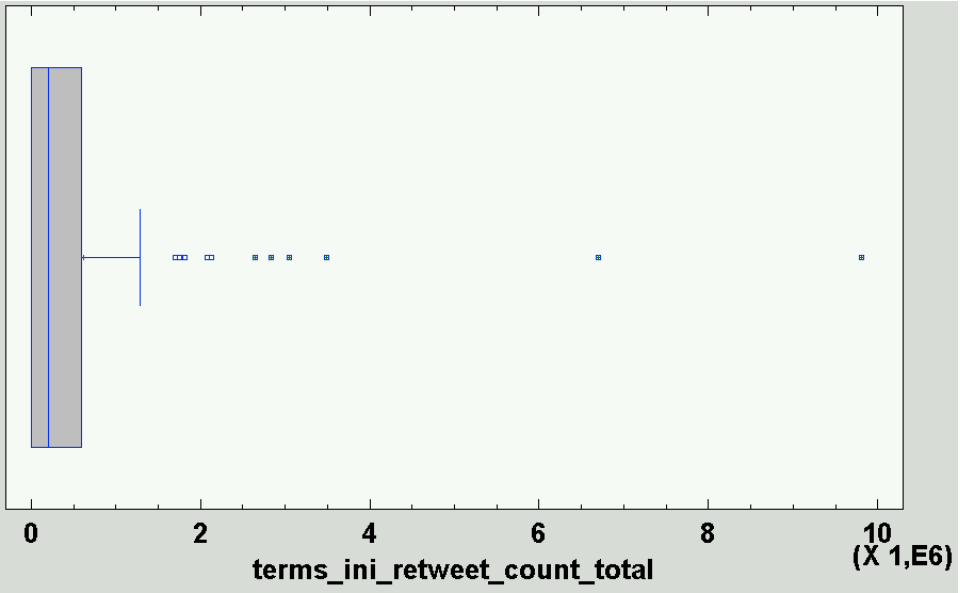


Figura 261. Series: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total

En la Figura 261 se puede observar que existen valores anómalos de tipo extremo de 2.652.370 o más.

c) Número de retuits de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 1.507,58.

Presenta un sesgo estandarizado de 16,9292 y una curtosis estandarizada de 39,4869. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

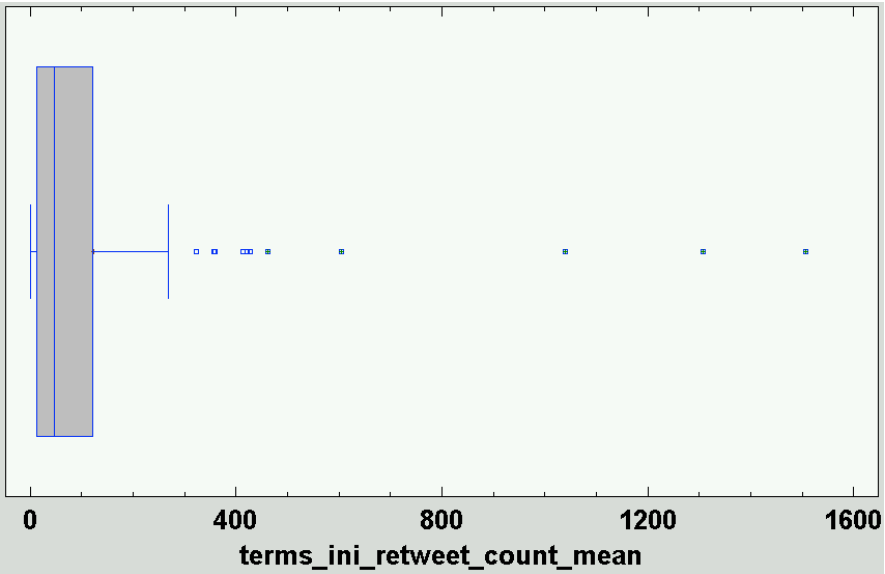


Figura 262. Series: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean

En la Figura 262 se puede observar que existen valores anómalos de tipo extremo de 460,65 o más.

d) Número de favoritos de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 104 valores con un rango de entre 0 y 39.389.

Presenta un sesgo estandarizado de 6,12904 y una curtosis estandarizada de 4,54134. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

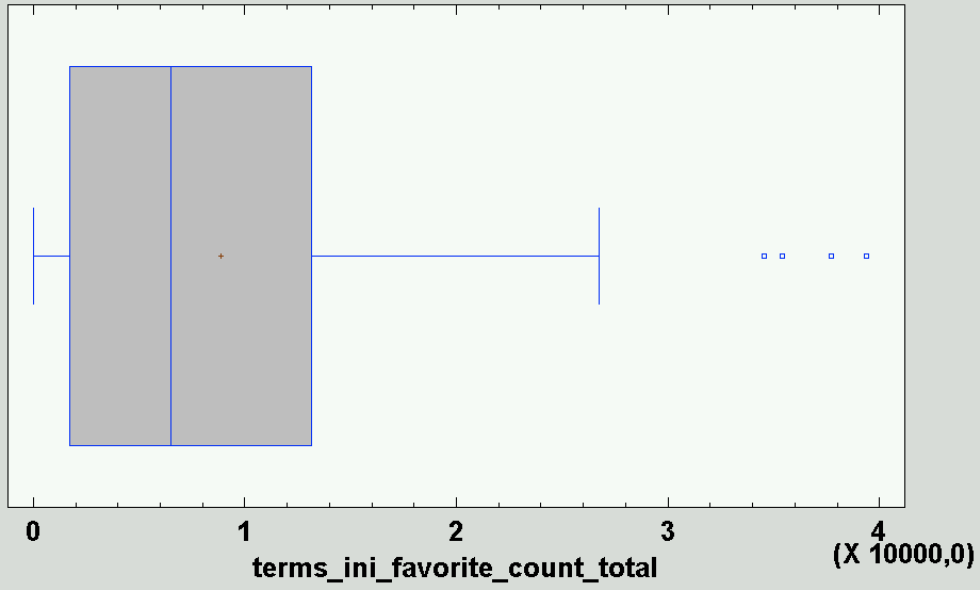


Figura 263. Series: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total

En la Figura 263 se puede observar que no existen valores anómalos de tipo extremo.

e) Número de favoritos de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 9,74.

Presenta un sesgo estandarizado de 4,77846 y una curtosis estandarizada de 3,37139. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

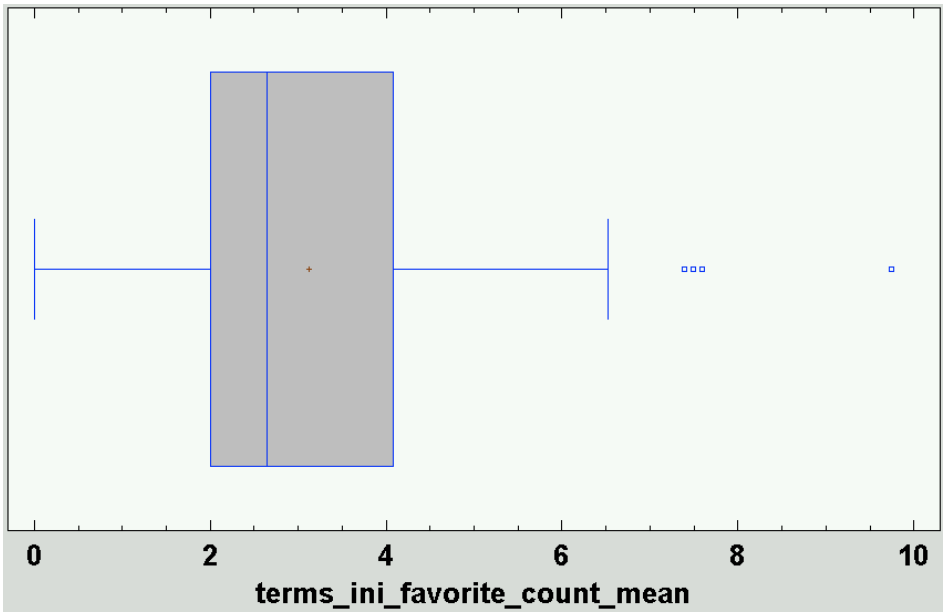


Figura 264. Series: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean

En la Figura 264 se puede observar que no existen valores anómalos de tipo extremo.

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_followers_talking_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 0,3.

Presenta un sesgo estandarizado de 19,0747 y una curtosis estandarizada de 59,7385. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

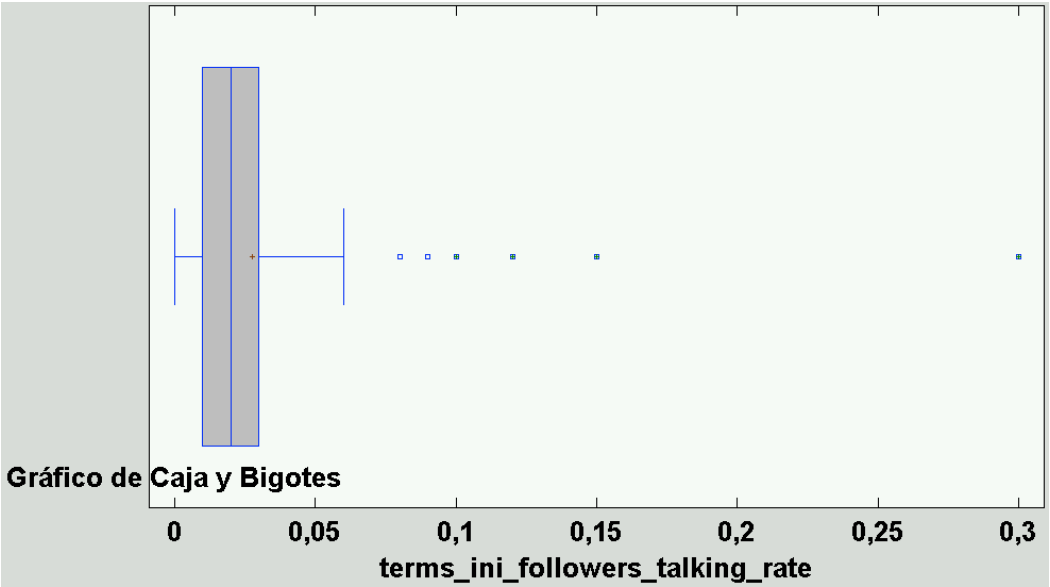


Figura 265. Series: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate

En la Figura 265 se puede observar que existen valores anómalos de tipo extremo de 0,1 o más.

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_followers_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 124.117.

Presenta un sesgo estandarizado de 10,8474 y una curtosis estandarizada de 18,9205. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

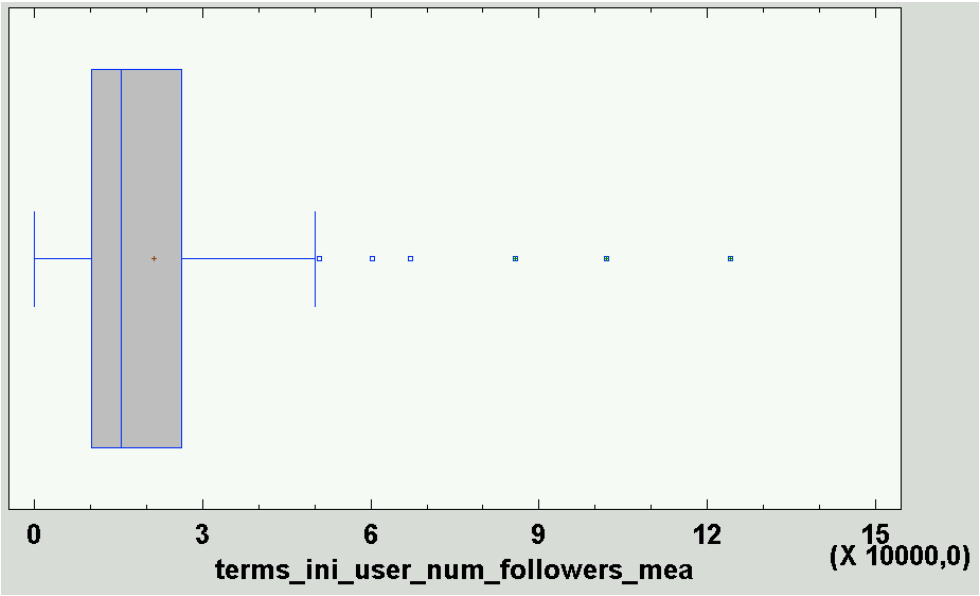


Figura 266. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean

En la Figura 266 se puede observar que existen valores anómalos de tipo extremo de 85.721,9 o más.

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_tweets_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 119.772.

Presenta un sesgo estandarizado de 6,55225 y una curtosis estandarizada de 13,4133. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

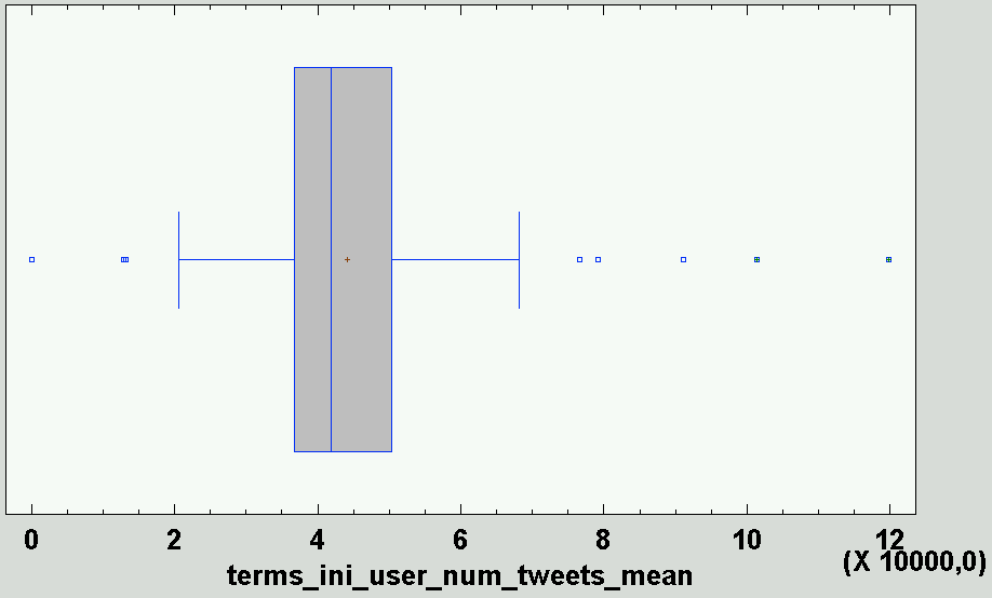


Figura 267. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean

En la Figura 267 se puede observar que existen valores anómalos de tipo extremo de 101.422 o más.

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_age_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 4.036,62.

Presenta un sesgo estandarizado de -3,54567 y una curtosis estandarizada de 11,1391. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

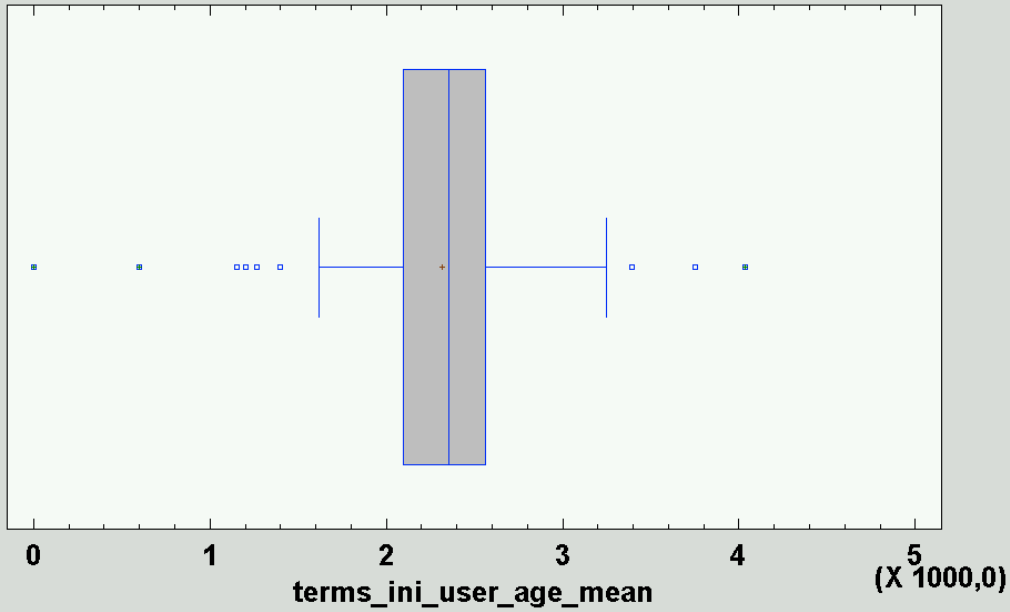


Figura 268. Series: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean

En la Figura 268 se puede observar que existen valores anómalos de tipo extremo de 600,58 o menos, y 4.036,62.

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial

Esta variable de predicción se identifica con la columna terms_ini_url_inclusion_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 104 valores con un rango de entre 0 y 0,87.

Presenta un sesgo estandarizado de 3,19443 y una curtosis estandarizada de 1,9391. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

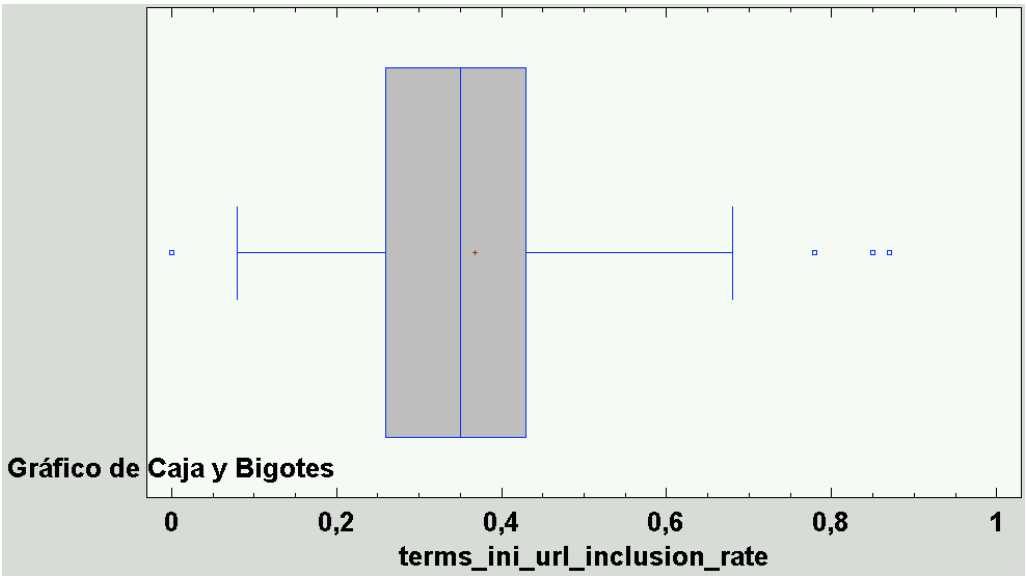


Figura 269. Series: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate

En la Figura 269 se puede observar que no existen valores anómalos de tipo extremo.

k) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de todas las variables de predicción y éxito. De esta manera se obtienen todos los datos de las variables del modelo en un mismo análisis. Se pueden observar los siguientes datos de estas:

Tabla 87

Series: Resumen estadístico de las variables de predicción

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
terms_ini_num_tweets	6.139.830	3,45941	0,0195947
terms_ini_retweet_count_total	1.704.700.000.000	19,6336	57,4145
terms_ini_retweet_count_mean	54.792,4	16,9292	39,4869

terms_ini_favorite_count_total	77.965.500	6,12904	4,54134
terms_ini_favorite_count_mean	2,97933	4,77846	3,37139
terms_ini_followers_talking_rate	0,001381154	19,0747	59,7385
terms_ini_user_num_followers_mean	393.442.000	10,8474	18,9205
terms_ini_user_num_tweets_mean	251.969.000	6,55225	13,4133
terms_ini_user_age_mean	269.663	-3,54567	11,1391
terms_ini_url_inclusion_rate	0,025226	3,19443	1,9391

Se puede observar en la Tabla 87 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple, es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 88

Series: Resumen estadístico de las variables de predicción con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(terms_ini_num_tweets)	2,28406	-5,14058	2,5413

log(terms_ini_retweet_count_total)	9,8223	-4,4522	1,77319
log(terms_ini_retweet_count_mean)	3,34076	-2,47293	0,345047
log(terms_ini_favorite_count_total)	2,87613	-6,22645	5,43921
log(terms_ini_favorite_count_mean)	0,305572	-1,61932	1,3486
log(terms_ini_followers_talking_rate)	0,493699	3,74555	1,13519
log(terms_ini_user_num_followers_mean)	0,721276	-1,54076	1,32348
log(terms_ini_user_num_tweets_mean)	0,107072	-1,56845	7,03026
log(terms_ini_user_age_mean)	0,055211	-9,23736	22,1326
log(terms_ini_url_inclusion_rate)	0,190676	-2,20543	2,22631

Todas las variables, salvo log(terms_ini_favorite_count_mean) y log(terms_ini_user_num_followers_mean), mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. En este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

También se puede apreciar que los valores de varianza son bastante parecidos salvo en el caso de log(terms_ini_retweet_count_total), por lo que se cumple la condición de homocedasticidad menos en esa variable. log(terms_ini_retweet_count_total) se elimina del modelo actual para que dicho requisito se cumpla.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la

prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 89

Series: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(terms_ini_favorite_count_total)	0,0111224
log(terms_ini_user_num_tweets_mean)	0,243206
terms_ini_user_age_mean	0,131927

En la Tabla 89 se puede ver que las variables que supera el valor-p necesario (mayor o igual que 0,05) para confirmar que sigue una distribución normal es log(terms_ini_favorite_count_mean), por lo que dicha variable también es tenida en cuenta en el modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 90

Series: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	retweet_count_mean

log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
terms_ini_user_age_mean	terms_end_num_tweets
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo. También incluye la versión original de: el número total de tuits de la tendencia y el promedio de edad en días de los usuarios que participan.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

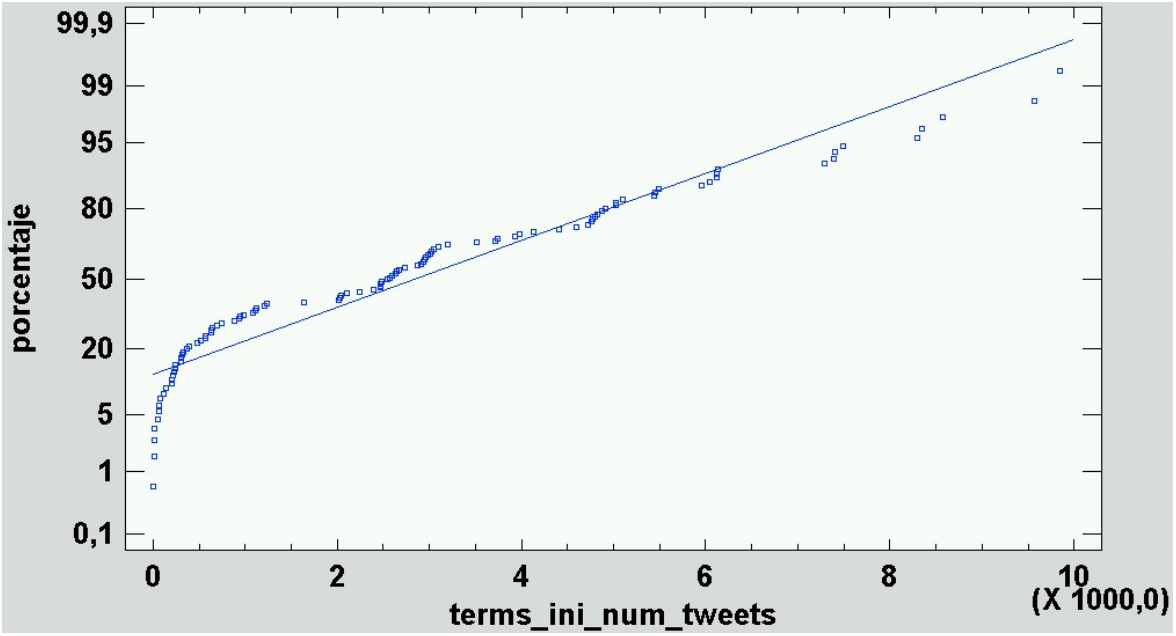


Figura 270. Series: Gráfico de probabilidad normal de la variable terms_ini_num_tweets

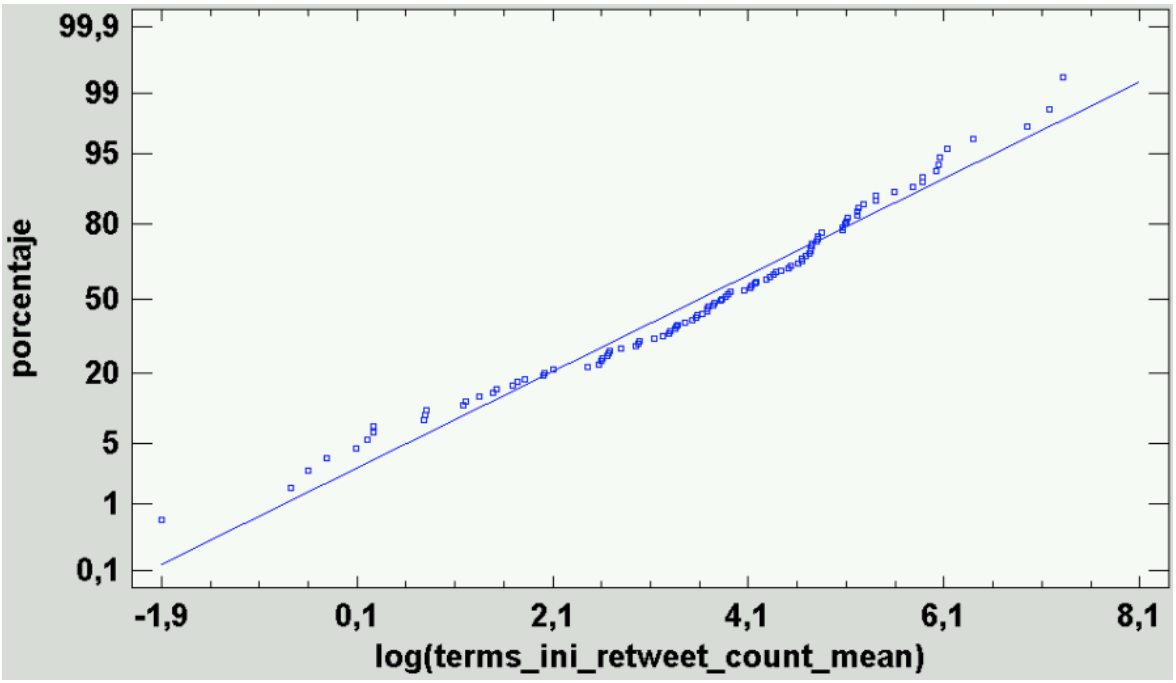


Figura 271. Series: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean)

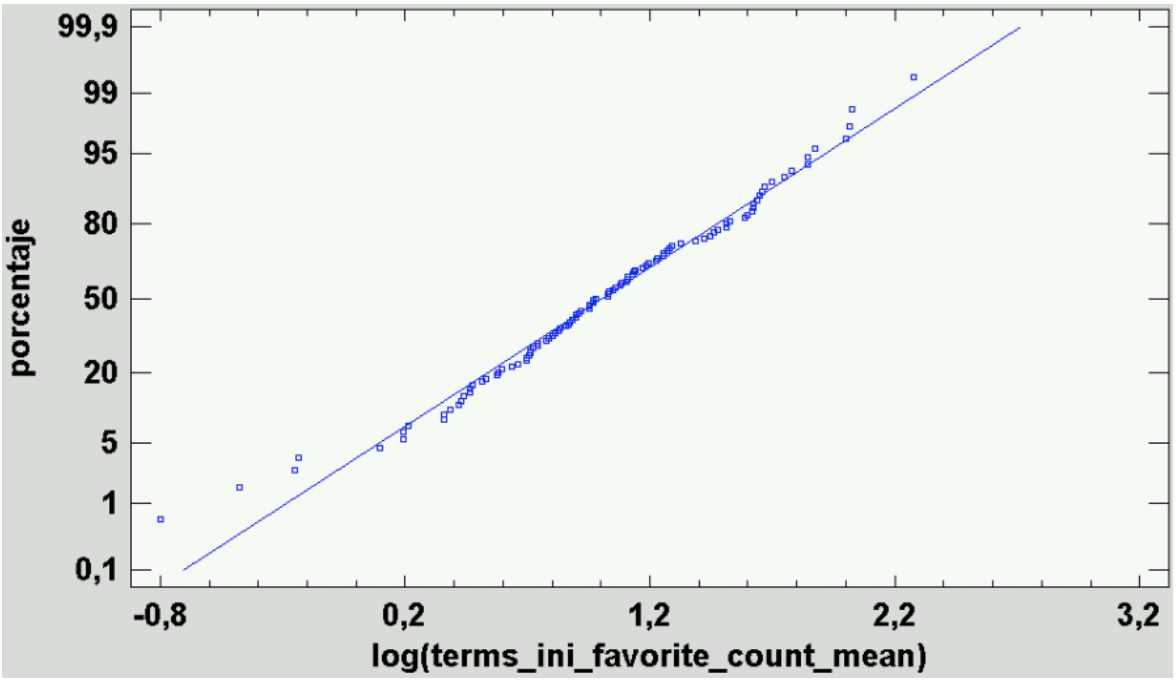


Figura 272. Series: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean)

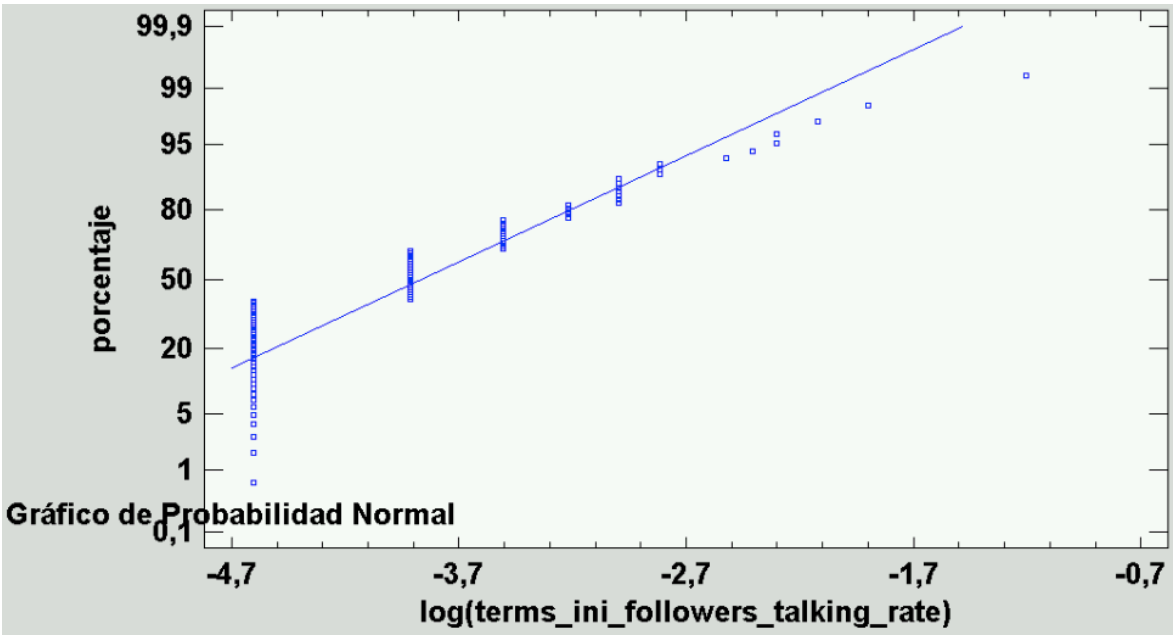


Figura 273. Series: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate)

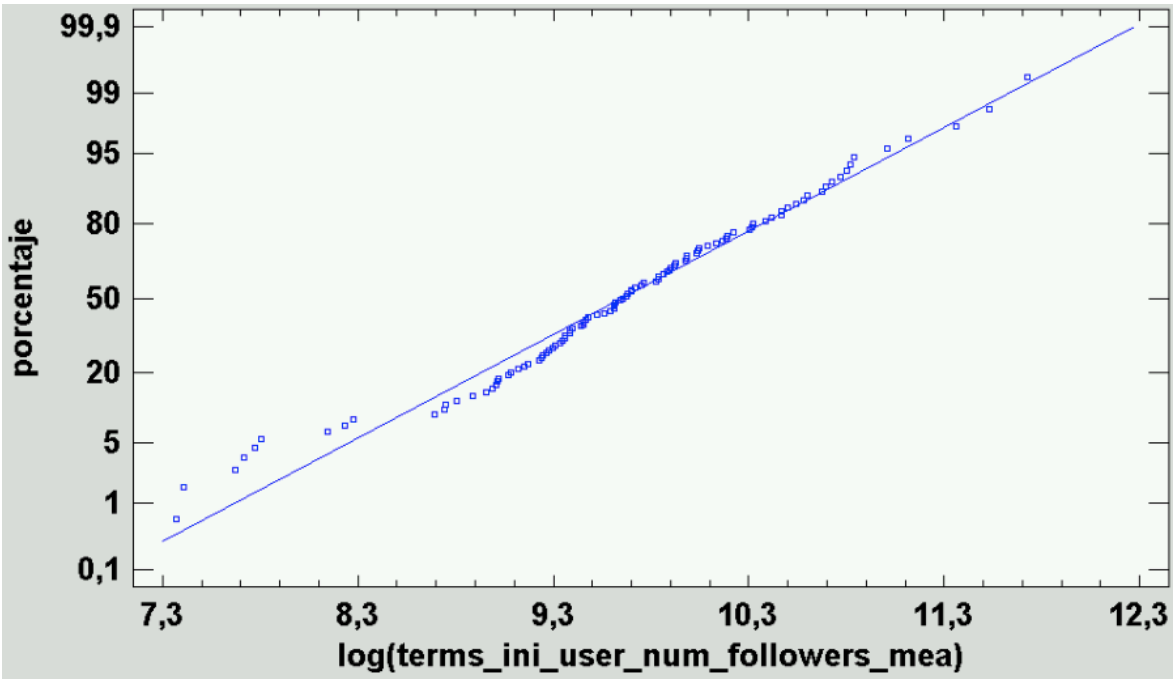


Figura 274. Series: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean)

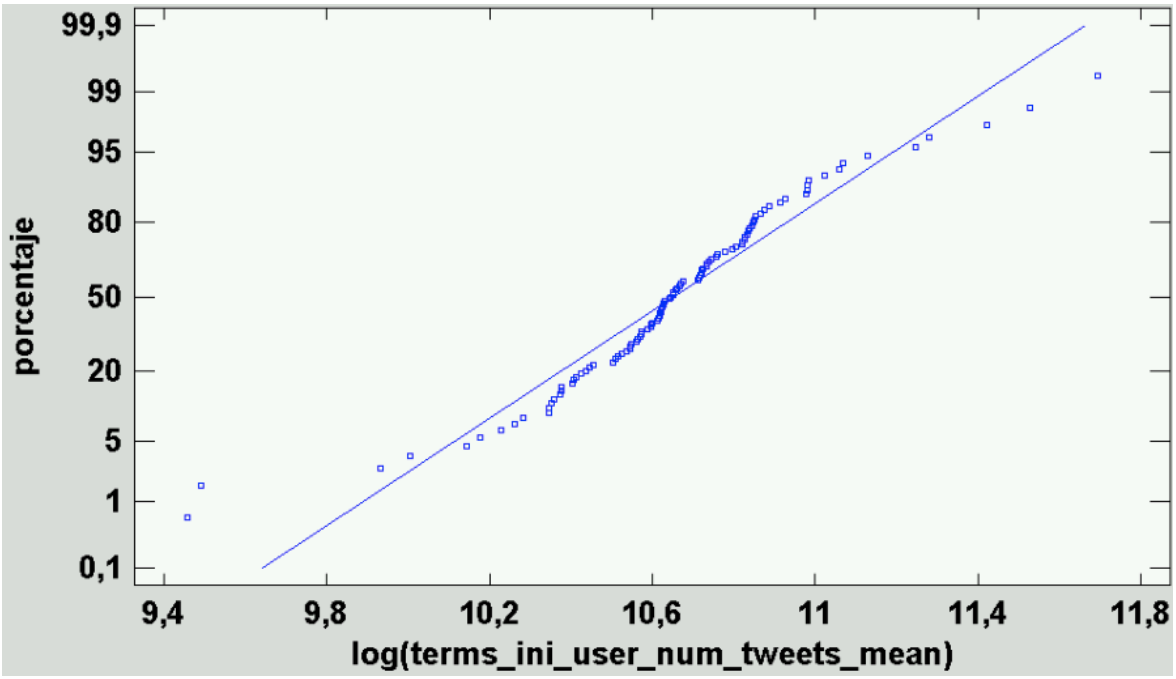


Figura 275. Series: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean)

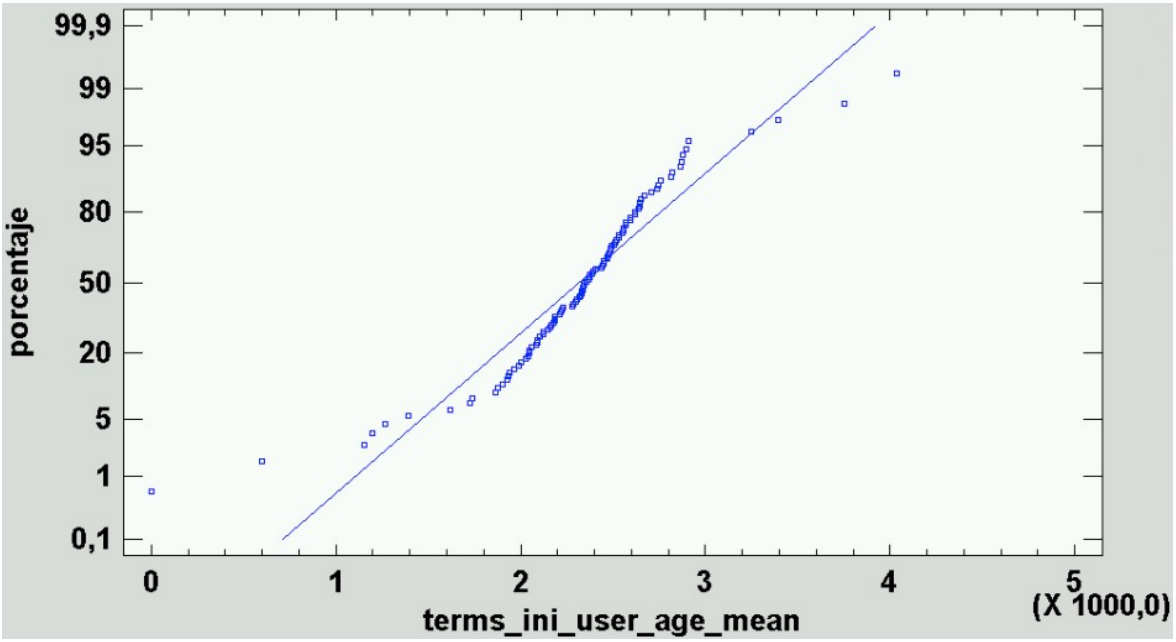


Figura 276. Series: Gráfico de probabilidad normal de la variable terms_ini_user_age_mean

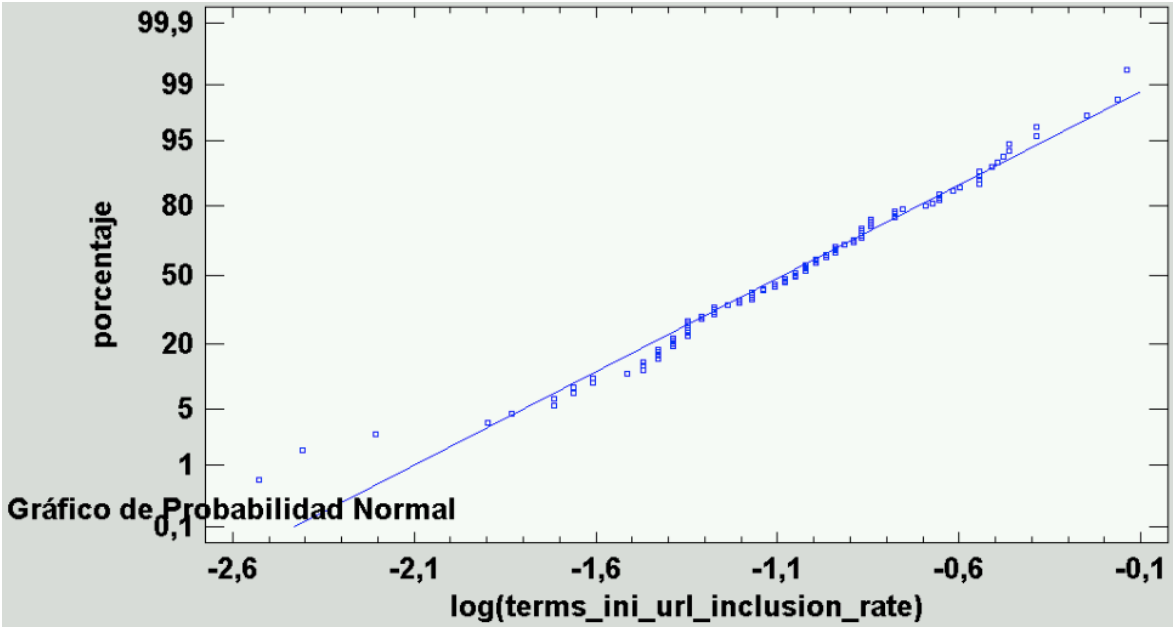


Figura 277. Series: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

l) Filtro de alta correlación (colinealidad)

Antes de proceder, es conveniente analizar la correlación que existe entre las variables con las que contamos para el modelo. Puesto que estas han sido normalizadas en el apartado anterior, dicho análisis de correlación Pearson se realizará con su transformación logarítmica.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

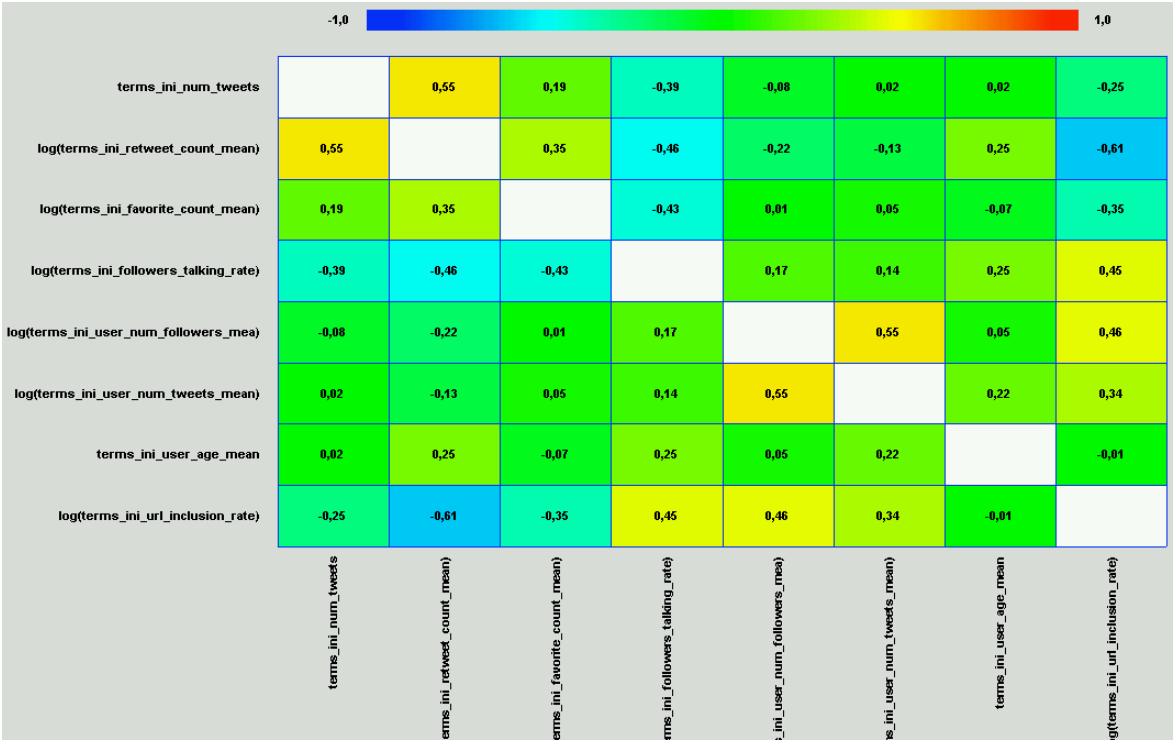


Figura 278. Series: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente

La Figura 278 indica que no hay ninguna correlación fuerte (igual o mayor a 0,7) entre las variables de predicción, por lo que resulta interesante analizar todas por separado.

La tabla de variables quedaría como sigue:

Tabla 91

Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	retweet_count_mean
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)

terms_ini_user_age_mean	terms_end_num_tweets
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

m) Análisis de componentes principales (ACP)

A continuación, se aplica el análisis de componentes principales (ACP, o PCA en inglés), una técnica que sirve par describir un conjunto de datos según nuevas variables no correlacionadas llamadas componentes (Dunteman, 1989).

El objetivo es representar los datos de la mejor manera posible a través de mínimos cuadrados, construyendo una transformación lineal según un nuevo sistema de coordenadas para los datos originales. Es decir, se plantea representar la variabilidad de los datos con el menor número de componentes o fórmulas posible, las cuales son combinaciones lineales de las variables originales (Dunteman, 1989).

Si las variables originales están muy correlacionadas entre sí, la mayor parte de la variabilidad se podrá expresar en pocas componentes. Si están totalmente incorrelacionadas, el número de componentes será igual al de las variables y este análisis carecerá de interés.

Las componentes se ordenan según la varianza original siendo la primer componente el que tenga la varianza de mayor tamaño. Cuanto mayor sea su varianza, mayor será la información que aporta esa componente (Amat Rodrigo, 2017).

Al construir la matriz de coeficientes de correlación, es posible una base de vectores propios, cuya transformación lineal es necesaria para mejora la simplicidad e interpretación que permita tratar de reducir la dimensionalidad de los datos (Dunteman, 1989). Esta reducción se efectuaría seleccionando las componentes principales que más aportan a la varianza e ignorando el resto. Esta selección se produce ordenando las componentes de mayor a menor aportación a la explicación de la variabilidad, y seleccionando tantas como sean necesarias hasta alcanzar un valor propio mayor o igual a 1.

De esta manera, el método ACP condensa la información de múltiples características en unas pocas, ya que se pretende explicar aproximadamente la información con menos valores que los originales.

Realizando el análisis de componentes principales se ha obtenido un total de tres componentes, explicando así el 70,657% de los datos con un valor propio de 1,21531. Las tres componentes tienen la Tabla 92 de pesos, siendo cada peso un valor de entre -1 y 1.

Tabla 92

Series: Tabla de pesos de las componentes

	Componente	*Componente*	*Componente*
	1	2	3
terms_ini_num_tweets	0,339704	0,316314	0,042807
log(terms_ini_retweet_count_mean)	0,472709	0,246993	0,278947
log(terms_ini_favorite_count_mean)	0,308358	0,291543	-0,334224
log(terms_ini_followers_talking_rate)	-0,428277	-0,114527	0,378431
log(terms_ini_user_num_followers_mea)	-0,300848	0,51862	-0,241265
log(terms_ini_user_num_tweets_mean)	-0,238109	0,608895	-0,0544061
terms_ini_user_age_mean	-0,0369735	0,299947	0,76687
log(terms_ini_url_inclusion_rate)	-0,483807	0,105535	-0,127212

De esta manera, por ejemplo, la primer componente principal tiene la fórmula siguiente, en donde los valores de las variables se han estandarizado restándoles su promedio y dividiéndolos entre su desviación estándar:

0,339704 * terms_ini_num_tweets + 0,472709 * log(terms_ini_retweet_count_mean) + 0,308358 * log(terms_ini_favorite_count_mean) – 0,428277 * log(terms_ini_followers_talking_rate) – 0,300848 * log(terms_ini_user_num_followers_mean) – 0,238109 * log(terms_ini_user_num_tweets_mean) – 0,0369735 * terms_ini_user_age_mean – 0,483807 * log(terms_ini_url_inclusion_rate)

Se observa que las variables que aportan positivamente a las tres componentes principales son: el número total de tuits y el promedio de retuits. El resto sí que aportan un valor negativo en alguna de ellas.

La relación entre las variables y las tres componentes principales se puede ver en la siguiente gráfica, ya que las variables se muestran en tres dimensiones formadas por estas componentes:

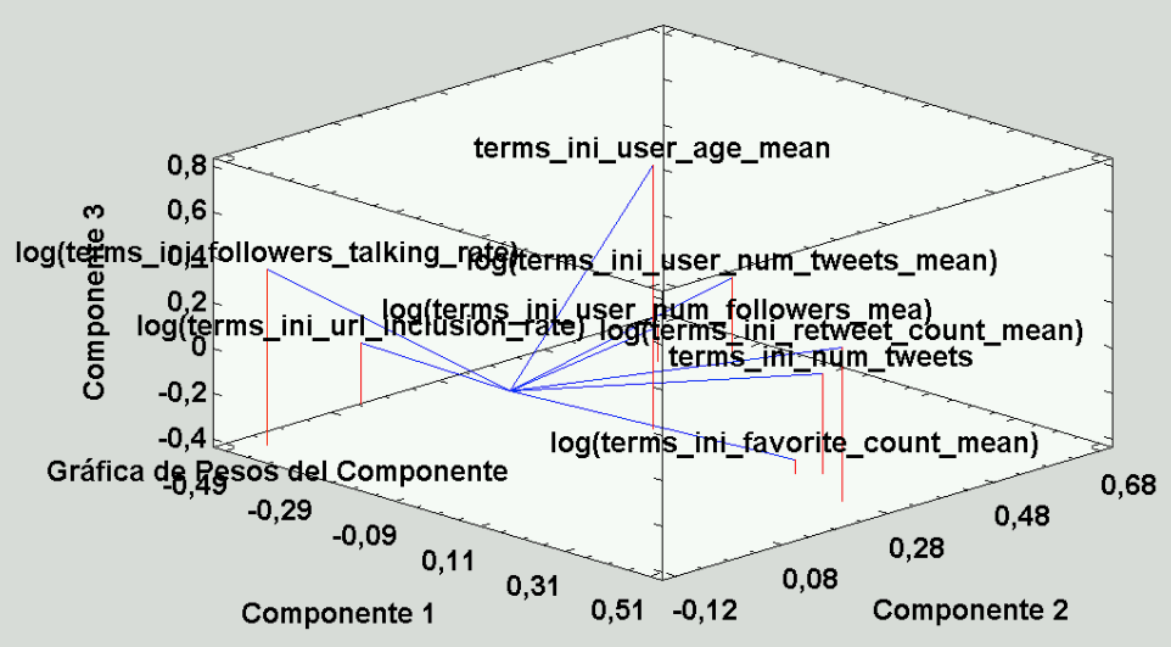


Figura 279. Series: Gráfica de pesos de cada componente principal

En la Tabla 92 se puede comprobar que todas las variables tienen una presencia significativa en alguna de las componentes principales, por lo que no se puede eliminar ninguna de las variables originales mediante esta técnica.

A continuación, se van a analizar todos los artículos de la categoría Series, de manera que se pueda comprobar si las características de los datos y las ecuaciones de predicción varían según la sección a la que pertenezca el artículo.

6.1.2.3. Regresión lineal múltiple

Para estudiar la posible relación entre las variables independientes de predicción de que disponemos y cada variable dependiente de éxito o, dicho de otro modo, para tratar de predecir el cálculo de estas, vamos a realizar un modelo de regresión múltiple.

Para realizar las regresiones múltiples, se cuenta con la Tabla *93* de variables resultante de todos los análisis anteriores:

Tabla 93

Series: Lista final de variables de predicción y de éxito para la regresión lineal múltiple

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)

log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	retweet_count_mean
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
terms_ini_user_age_mean	terms_end_num_tweets
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

Las variables de predicción responden a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores que hablan de la tendencia, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits. Además, incluye también la versión original del número de tuits de la tendencia y la edad en días de los usuarios que participan. Todo lo anterior aplicado a la tendencia el día de la publicación del artículo.

La lista de variables de éxito está formada por la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después. Además, incluye también la versión original del promedio de retuits en la cuenta del medio y el número de tuits de la tendencia 14 días después.

a) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión múltiple con la variable dependiente log(uniquepageviews_total). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 94

Series: Valor-P de las variables de la regresión múltiple de log(uniquepageviews_total)

Variable	Estimación	Valor-P

Constante	5,92251	0,1537
terms_ini_num_tweets	0,000111941	0,0299
log(terms_ini_retweet_count_mean)	-0,00308874	0,9736
log(terms_ini_favorite_count_mean)	0,141138	0,5162
log(terms_ini_followers_talking_rate)	-0,0233037	0,8949
log(terms_ini_user_num_followers_mean)	-0,456791	0,0081
log(terms_ini_user_num_tweets_mean)	0,245897	0,5683
terms_ini_user_age_mean	-0,000389502	0,1751
log(terms_ini_url_inclusion_rate)	0,530329	0,1376
Modelo		0,0265

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 95

Series: Valor-P de las variables de la regresión múltiple simplificada de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	2,82621	0
terms_ini_num_tweets	0,00010482	0,0096
Modelo		0,0096

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(2,82621 - 0,00010482 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

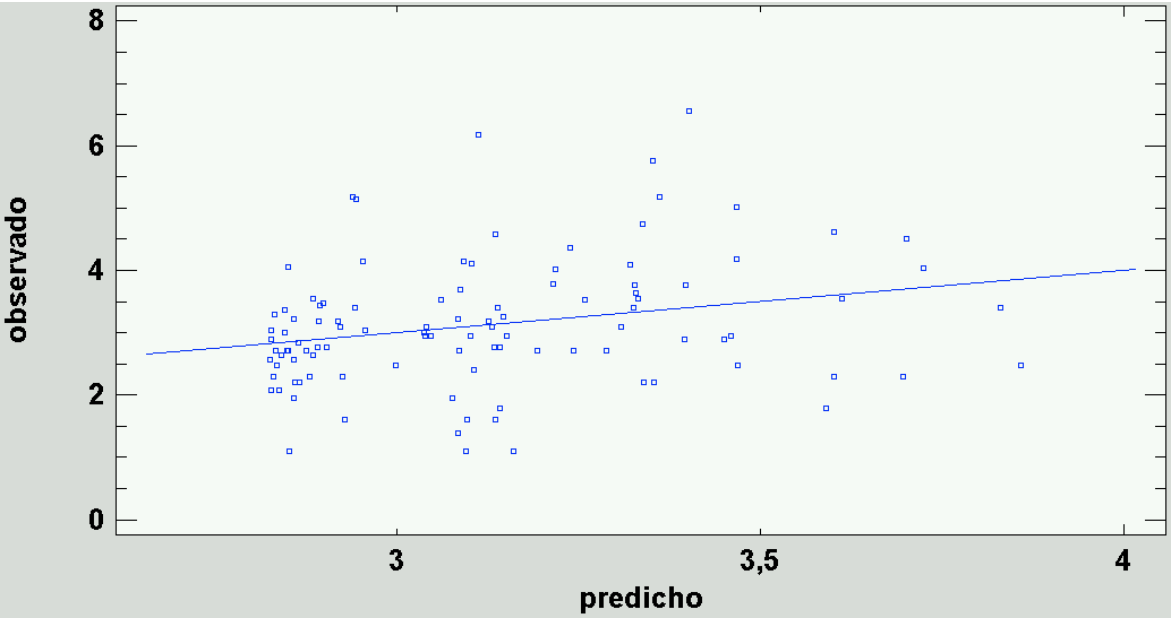


Figura 280. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 6,38706% de la variabilidad de log(uniquepageviews_total), mientras que el R-Cuadrado ajustado indica un 5,46929%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

b) AdSense eCPM (promedio)

Para tratar de predecir el valor de estimación de ingresos de los anuncios por cada 1000 páginas vistas desde los anuncios de Google AdSense es necesario realizar la regresión múltiple con la variable dependiente log(adsense_ecpm_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 96

Series: Valor-P de las variables de la regresión múltiple de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P
Constante	-9,50939	0,072
terms_ini_num_tweets	-0,0000427824	0,4165
log(terms_ini_retweet_count_mean)	-0,0991653	0,3029
log(terms_ini_favorite_count_mean)	-0,0502467	0,8291
log(terms_ini_followers_talking_rate)	-0,237201	0,2106
log(terms_ini_user_num_followers_mean)	-0,324144	0,0837
log(terms_ini_user_num_tweets_mean)	0,748564	0,1475
terms_ini_user_age_mean	0,000353451	0,2479
log(terms_ini_url_inclusion_rate)	-0,0224222	0,9519
Modelo		0,3599

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 97

Series: Valor-P de las variables de la regresión múltiple simplificada de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P
Constante	-1,66402	0,192
log(terms_ini_user_num_followers_mean)	-0,184761	0,1623

Modelo	0,1623

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable log(adsense_ecpm_mean) no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

g) Duración de la visita (promedio)

Para tratar de predecir el valor de la duración de la visita (promedio) es necesario realizar la regresión múltiple con la variable dependiente log(avgtimeonpage_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 98

Series: Valor-P de las variables de la regresión múltiple de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	9,7173	0,0114
terms_ini_num_tweets	-0,0000835527	0,0746
log(terms_ini_retweet_count_mean)	0,119623	0,1625
log(terms_ini_favorite_count_mean)	0,02199	0,9117
log(terms_ini_followers_talking_rate)	-0,0575488	0,7208
log(terms_ini_user_num_followers_mean)	0,0526862	0,7327
log(terms_ini_user_num_tweets_mean)	-0,532796	0,1776
terms_ini_user_age_mean	-0,000148004	0,5708
log(terms_ini_url_inclusion_rate)	-0,00985089	0,9757
Modelo		0,3404

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 99

Series: Valor-P de las variables de la regresión múltiple simplificada de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	5,29092	0
terms_ini_num_tweets	-0,0000982606	0,0288
log(terms_ini_retweet_count_mean)	0,172141	0,007
terms_ini_user_age_mean	-0,000418889	0,0434
Modelo		0,022

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

avgtimeonpage_mean = exp(5,29092 - 0,0000982606 * terms_ini_num_tweets + 0,172141 * log(terms_ini_retweet_count_mean) - 0,000418889 * terms_ini_user_age_mean)

Para realizar el cálculo de avgtimeonpage_mean será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

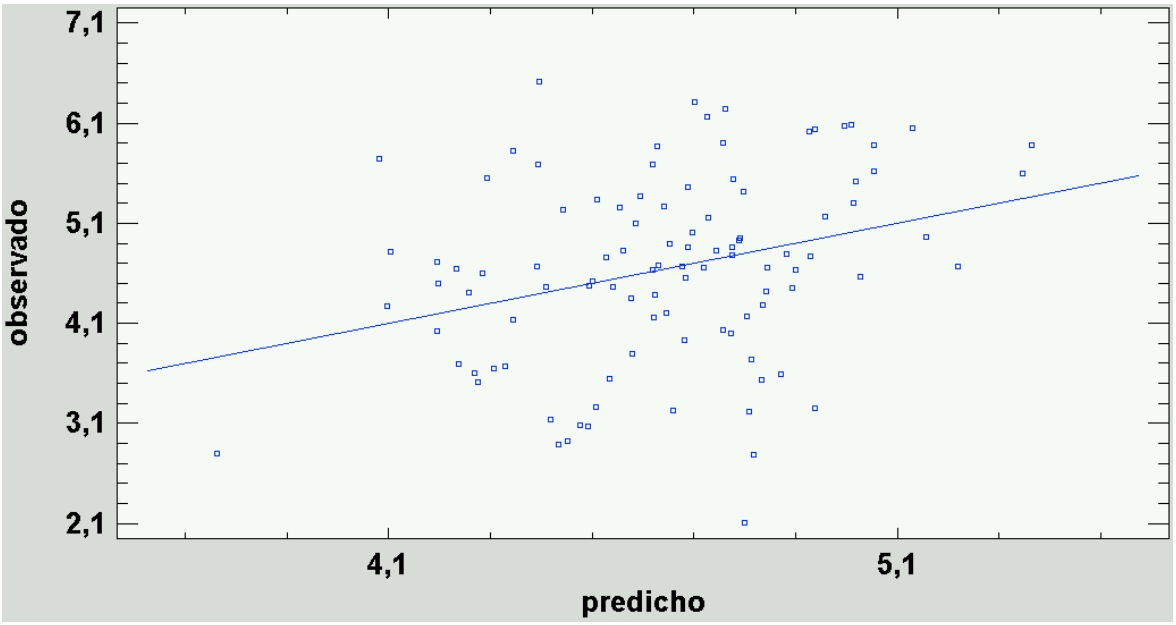


Figura 281. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 9,31683% de la variabilidad de log(avgtimeonpage_mean), mientras que el R-Cuadrado ajustado indica un 6,54081%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

h) Páginas vistas por sesión (promedio)

Para tratar de predecir el valor de las páginas vistas por sesión (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(pageviewspersession_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 100

Series: Valor-P de las variables de la regresión múltiple de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	-1,05491	0,5798
terms_ini_num_tweets	-0,00000567793	0,8088
log(terms_ini_retweet_count_mean)	-0,00975694	0,8207
log(terms_ini_favorite_count_mean)	-0,0175753	0,8607

log(terms_ini_followers_talking_rate)	0,151863	0,0645
log(terms_ini_user_num_followers_mean)	0,0779859	0,3182
log(terms_ini_user_num_tweets_mean)	0,107701	0,5879
terms_ini_user_age_mean	-0,000102825	0,4361
log(terms_ini_url_inclusion_rate)	-0,241834	0,1421
Modelo		0,4819

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 101

Series: Valor-P de las variables de la regresión múltiple simplificada de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	0,664353	0,006
log(terms_ini_followers_talking_rate	0,119508	0,0492
Modelo		0,0492

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

pageviewspersession_mean = exp(0,664353 + 0,119508 * log(terms_ini_followers_talking_rate))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

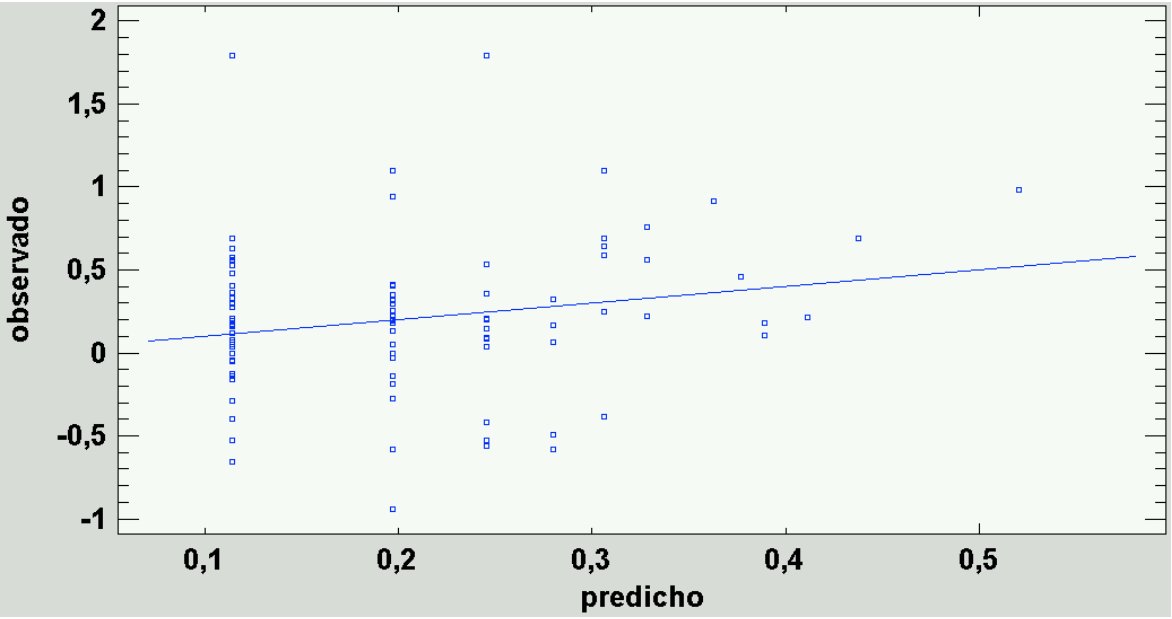


Figura 282. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 4,09697% de la variabilidad de log(pageviewspersession_mean), mientras que el R-Cuadrado ajustado indica un 3,06576%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Nº de retuits en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de retuits en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente retweet_count_mean. Se utilizan como variables independientes las disponibles para la predicción.

Tabla 102

Series: Valor-P de las variables de la regresión múltiple de retweet_count_mean

Variable	Estimación	Valor-P
Constante	-0,0525295	0,9781
terms_ini_num_tweets	0,0000533706	0,0255

log(terms_ini_retweet_count_mean)	-0,0664104	0,1268
log(terms_ini_favorite_count_mean)	-0,0257904	0,7976
log(terms_ini_followers_talking_rate)	-0,0700942	0,3917
log(terms_ini_user_num_followers_mean)	-0,013479	0,8632
log(terms_ini_user_num_tweets_mean)	0,0898871	0,6523
terms_ini_user_age_mean	-0,0000410543	0,7564
log(terms_ini_url_inclusion_rate)	-0,0651417	0,6919
Modelo		0,3882

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 103

Series: Valor-P de las variables de la regresión múltiple simplificada de retweet_count_mean

Variable	Estimación	Valor-P
Constante	0,829906	0
terms_ini_num_tweets	0,0000263004	0,1785
Modelo		0,1785

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable retweet_count_mean no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

d) Nº de favoritos en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de favoritos en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(favorite_count_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 104

Series: Valor-P de las variables de la regresión múltiple de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	-0,186538	0,928
terms_ini_num_tweets	0,0000655985	0,0048
log(terms_ini_retweet_count_mean)	0,017719	0,6921
log(terms_ini_favorite_count_mean)	-0,00510468	0,9594
log(terms_ini_followers_talking_rate)	-0,0178004	0,8289
log(terms_ini_user_num_followers_mean)	0,0323877	0,6734
log(terms_ini_user_num_tweets_mean)	0,0141467	0,9457
terms_ini_user_age_mean	-0,0000861722	0,5182
log(terms_ini_url_inclusion_rate)	0,00361389	0,9825
Modelo		0,0451

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 105

Series: Valor-P de las variables de la regresión múltiple simplificada de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	0,242027	0,001
terms_ini_num_tweets	0,000064501	0,0009
Modelo		0,0009

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

favorite_count_mean = exp(0,242027 + 0,000064501 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

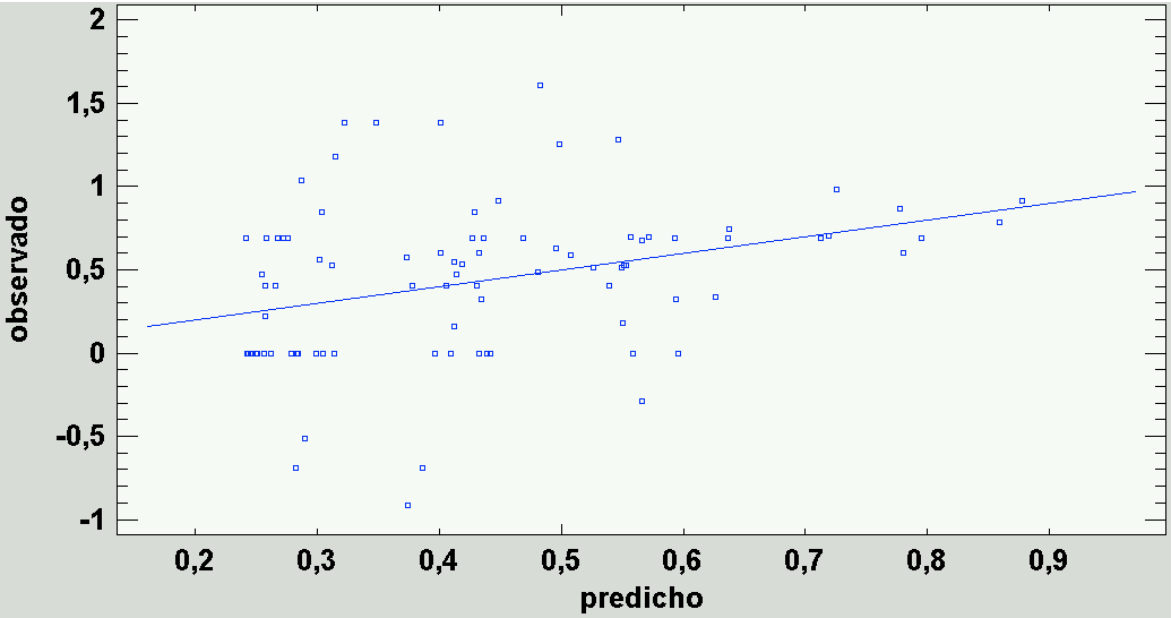


Figura 283. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 11,6459% de la variabilidad de log(favorite_count_mean), mientras que el R-Cuadrado ajustado indica un 10,6642%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor de el número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión múltiple con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Tabla 106

Series: Valor-P de las variables de la regresión múltiple de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	3.538,34	0,4645
terms_ini_num_tweets	0,826592	0
log(terms_ini_retweet_count_mean)	73,023	0,5041
log(terms_ini_favorite_count_mean)	-804,409	0,0021
log(terms_ini_followers_talking_rate)	-23,31	0,91
log(terms_ini_user_num_followers_mean)	-115,176	0,5604
log(terms_ini_user_num_tweets_mean)	-59,865	0,9054
terms_ini_user_age_mean	-0,376416	0,262
log(terms_ini_url_inclusion_rate)	359,272	0,3879
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 107

Series: Valor-P de las variables de la regresión múltiple simplificada de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	659,645	0,0172
terms_ini_num_tweets	0,807158	0
log(terms_ini_favorite_count_mean)	-597,626	0,0088
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = 659,645 + 0,807158 * terms_ini_num_tweets - 597,626 * log(terms_ini_favorite_count_mean)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

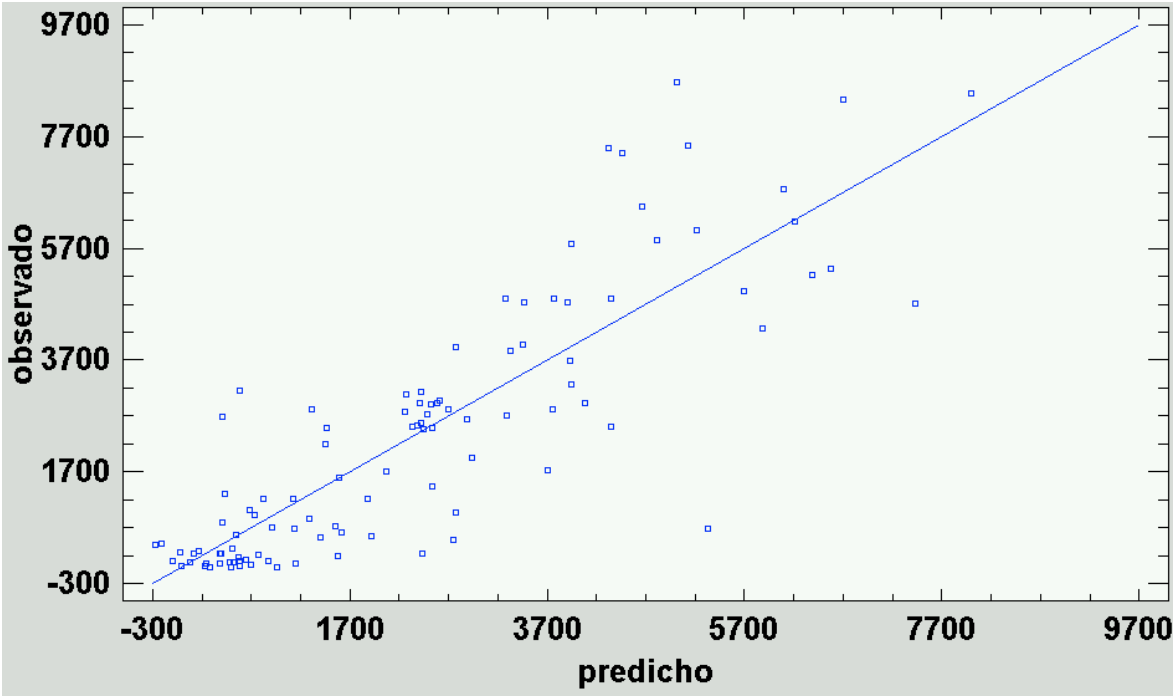


Figura 284. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 72,2708% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 71,7163%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen más la línea que en las regresiones lineales múltiples anteriores.

f) Nº de retuits de la tendencia 14 días después (promedio)

Para tratar de predecir el valor de el número de retuits de la tendencia 14 días después (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 108

Series: Valor-P de las variables de la regresión múltiple de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	1,80148	0,7673
terms_ini_num_tweets	0,000201535	0,009

log(terms_ini_retweet_count_mean)	0,478955	0,0013
log(terms_ini_favorite_count_mean)	-0,411181	0,2093
log(terms_ini_followers_talking_rate)	-0,660614	0,0127
log(terms_ini_user_num_followers_mean)	-0,108553	0,6743
log(terms_ini_user_num_tweets_mean)	-0,135164	0,8319
terms_ini_user_age_mean	-0,0000203138	0,9625
log(terms_ini_url_inclusion_rate)	0,694845	0,1939
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 109

Series: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	-0,88257	0,2803
terms_ini_num_tweets	0,000220248	0,0032
log(terms_ini_retweet_count_mean)	0,36843	0,0012
log(terms_ini_followers_talking_rate)	-0,488484	0,0339
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_mean = exp(-0,88257 + 0,000220248 * terms_ini_num_tweets + 0,36843 * log(terms_ini_retweet_count_mean) - 0,488484 * log(terms_ini_followers_talking_rate))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

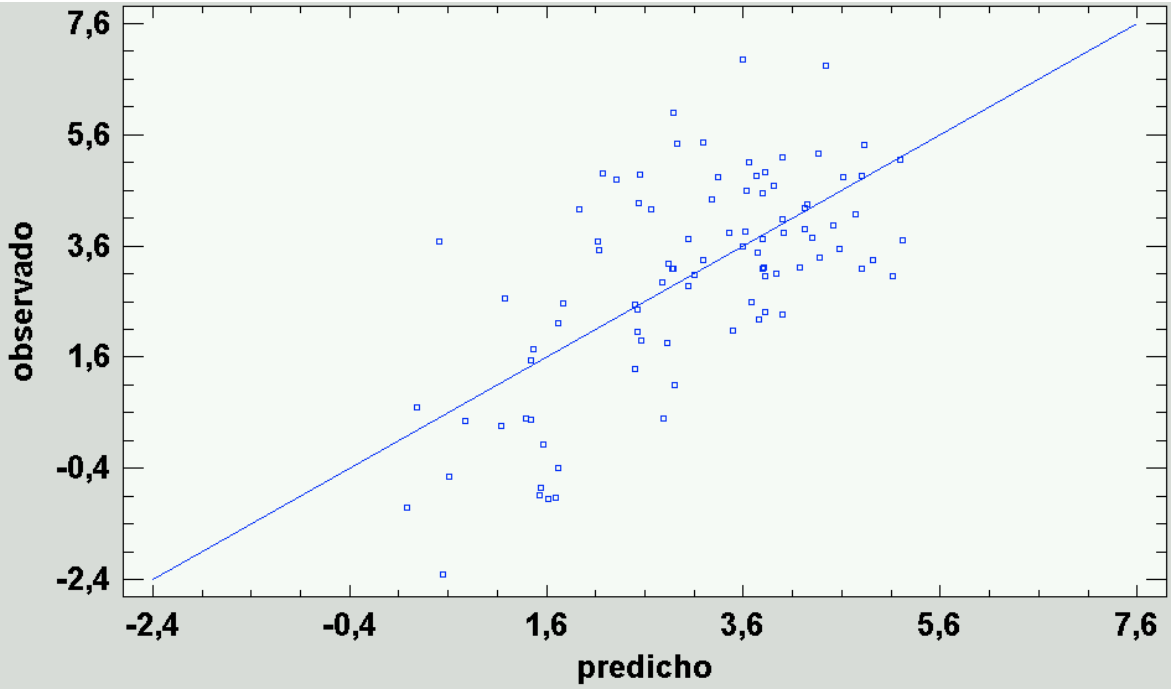


Figura 285. Series: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 43,4481% de la variabilidad de log(terms_end_retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 41,5202%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen más la línea que en la mayoría de las regresiones lineales múltiples anteriores.

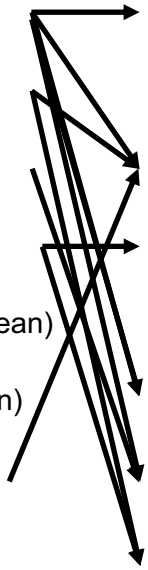
g) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones lineales múltiples de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 110

Series: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones lineales múltiples

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	retweet_count_mean
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
terms_ini_user_age_mean	terms_end_num_tweets
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)



Se puede observar en la Tabla *110* que solo terms_ini_num_tweets, log(terms_ini_retweet_count_mean), log(terms_ini_favorite_count_mean), log(terms_ini_followers_talking_rate) y terms_ini_user_age_mean participan en alguna de las ecuaciones de predicción, por lo que solamente estas son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits de la tendencia explica parte de los datos de páginas vistas únicas.
- El número de tuits, el promedio de retuits y la edad en días de los usuarios explica parte de los datos de la duración de la visita.

- La ratio de seguidores del medio que hablan de la tendencia explica parte de los datos de la media de páginas vistas por sesión.
- El número de tuits de la tendencia explica parte de los datos del número medio de favoritos en la cuenta del medio.
- El número de tuits y el promedio de favoritos explican parte de los datos del número de tuits 14 días después.
- El número de tuits, el promedio de retuits y la ratio de seguidores del medio que hablan de la tendencia explican parte de los datos del promedio de retuits 14 días después.
- El número de tuits de la tendencia, por tanto, participa en la predicción de todas las variables de éxito salvo las páginas vistas por sesión.
- El promedio de eCPM de los anuncios de Google AdSense y el promedio de retuits en la cuenta del medio no se han podido predecir con las variables de predicción existentes en el modelo.

6.1.2.4. Regresión binomial negativa o de Poisson

A continuación, se tratará de predecir todas las variables de éxito que sean de conteo (enteros y sin números negativos) a partir de todas las variables de predicción que sean independientes entre sí según la regresión binomial negativa o la regresión de Poisson.

a) Filtro de alta correlación (colinealidad)

Las variables que aporten información para tratar de realizar la regresión deben ser independientes, motivo por el cual es necesario hacer un filtro de alta correlación de manera que se asegure que todas aportan información diferente.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

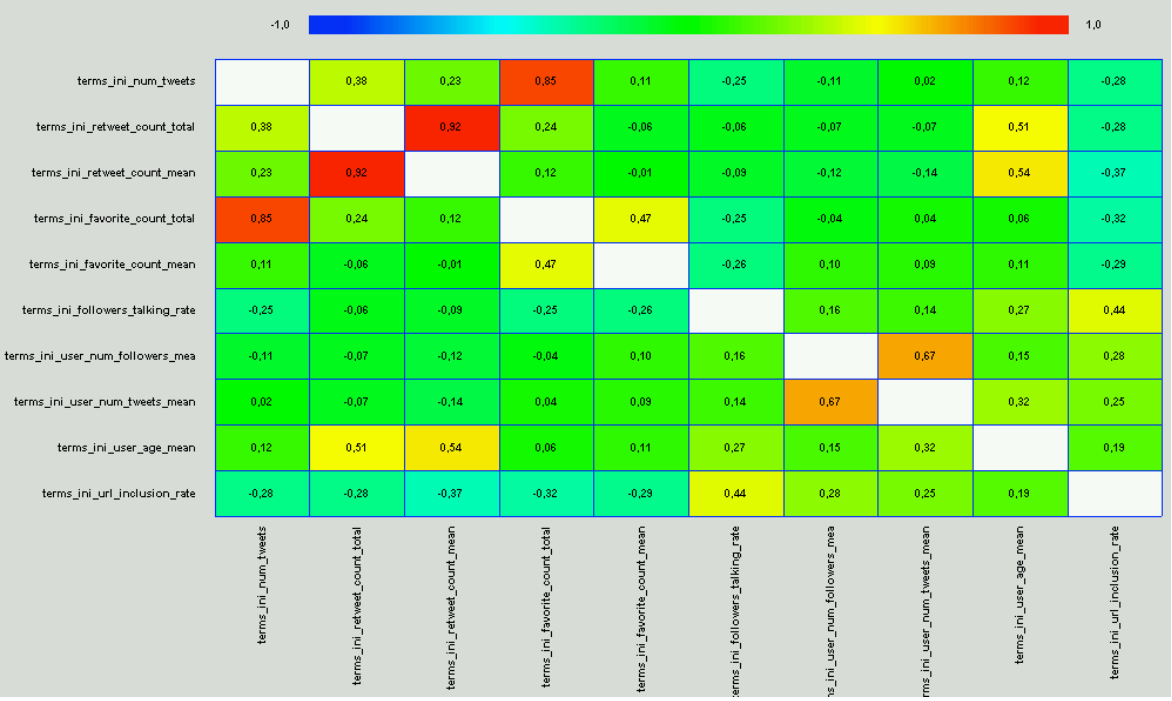


Figura 286. Series: Matriz de correlaciones Pearson entre las variables de predicción

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_ini_num_tweets y terms_ini_favorite_count_total tienen un coeficiente de correlación de 0,852 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- tems_ini_retweet_count_total y terms_ini_retweet_count_mean tienen un coeficiente de correlación de 0,9235 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige terms_ini_num_tweets y terms_ini_retweet_count_mean por tener un sesgo y una curtosis estandarizados menores, como se puede comprobar en el anexo 6.1.2.2.

La tabla de variables quedaría como sigue:

Tabla 111

Series: Lista de variables de predicción y de éxito para la regresión binomial negativa o de Poisson tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito

terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_mean	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a: el número total de tuits, el número total de retuits, el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de edad en días de la cuenta de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

b) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión con la variable dependiente uniquepageviews_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable uniquepageviews_total, el Chi-cuadrado calculado es 846.668 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 112

Series: Valor-P de las variables de la regresión binomial negativa de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,6679	0
terms_ini_num_tweets	0,000165651	0
terms_ini_retweet_count_mean	-0,00166904	0
terms_ini_favorite_count_mean	0,0401883	1
terms_ini_followers_talking_rate	-0,283798	1
terms_ini_user_num_followers_mean	-0,0000363378	0
terms_ini_user_num_tweets_mean	0,0000036733	0,0001
terms_ini_user_age_mean	-0,000104381	0
terms_ini_url_inclusion_rate	1,06018	0
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 113

Series: Valor-P de las variables de la regresión binomial negativa simplificada de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,77086	0
terms_ini_num_tweets	0,000152789	0
terms_ini_retweet_count_mean	-0,00177825	0

terms_ini_followers_talking_rate	-0,792307	0
terms_ini_user_num_followers_mean	-0,0000337097	0
terms_ini_url_inclusion_rate	0,904511	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(3,77086 + 0,000152789 * terms_ini_num_tweets - 0,00177825 * terms_ini_retweet_count_mean - 0,792307 * terms_ini_followers_talking_rate - 0,0000337097 * terms_ini_user_num_followers_mean + 0,904511 * terms_ini_url_inclusion_rate)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

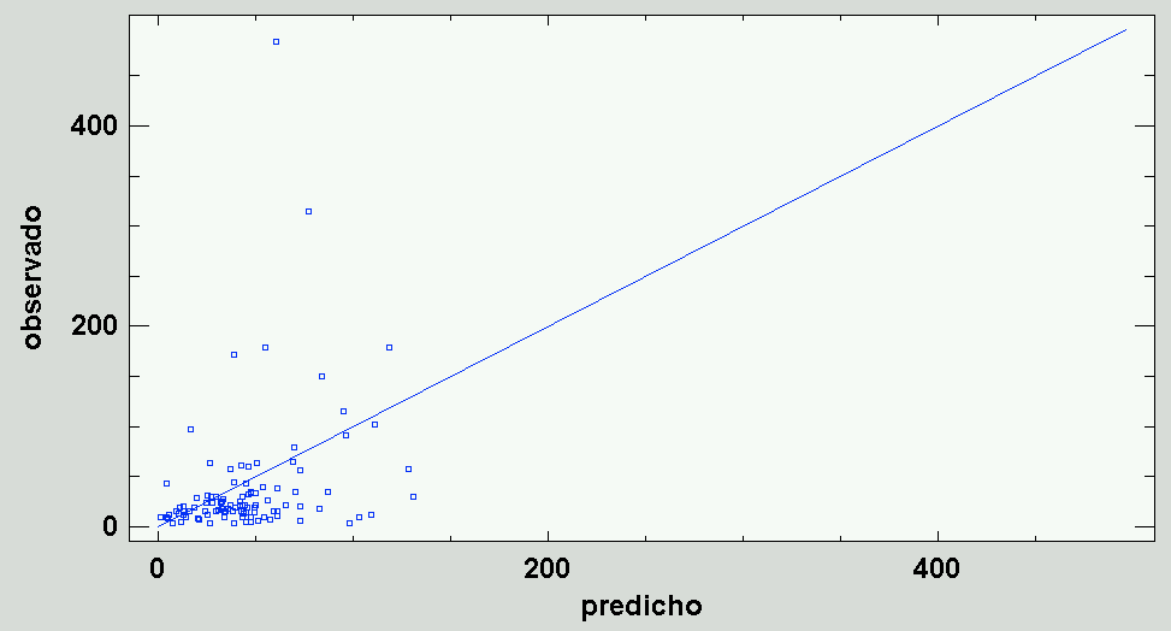


Figura 287. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 9,2758% de la variabilidad de uniquepageviews_total, mientras que el R-Cuadrado ajustado indica un 8,97332%. Este

número bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_num_tweets, el Chi-cuadrado calculado es 544.154.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 114

Series: Valor-P de las variables de la regresión binomial negativa de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	18,5432	0
terms_ini_num_tweets	0,000309556	0,0003
terms_ini_retweet_count_mean	-0,00114114	1
terms_ini_favorite_count_mean	-0,314142	0,0752
terms_ini_followers_talking_rate	5,00515	1
terms_ini_user_num_followers_mean	-0,00000465616	0,8581
terms_ini_user_num_tweets_mean	0,000000435354	1
terms_ini_user_age_mean	0,000258464	1
terms_ini_url_inclusion_rate	-0,962058	0,5622
Modelo		0,0341

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 115

Series: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	17,9916	0
terms_ini_num_tweets	0,000282693	0,0002
Modelo		0,0002

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(17,9916 + 0,000282693 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

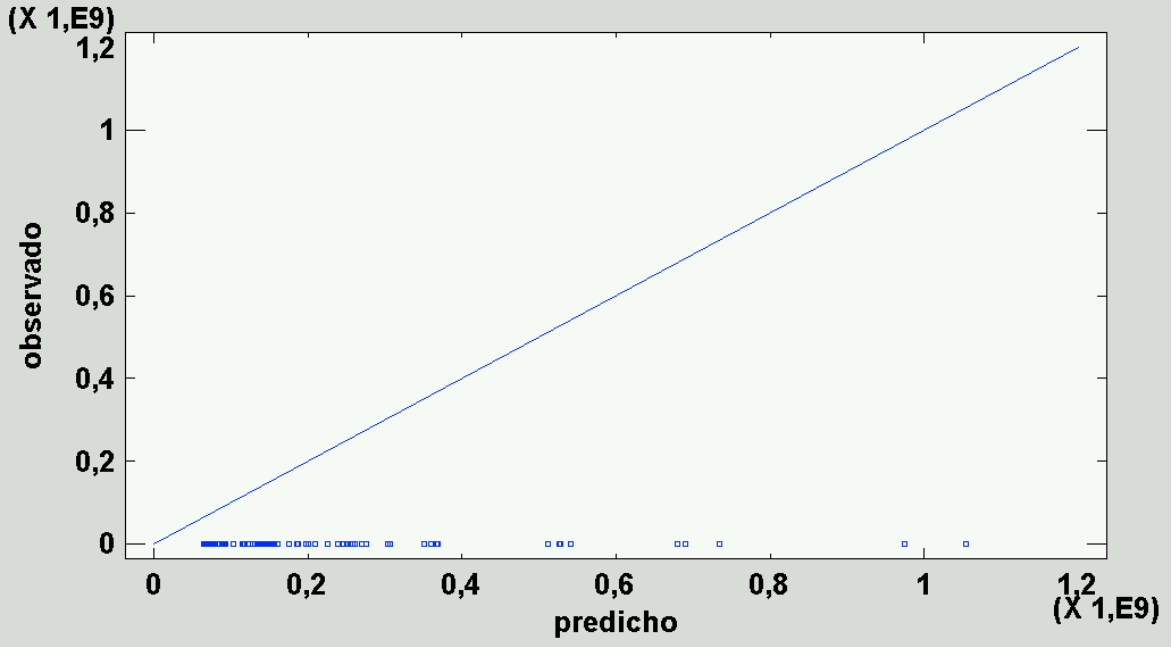


Figura 288. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 6,99566% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 4,91903%. Este número se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se acercan a la línea que muestra una predicción acertada.

d) Nº de retuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de retuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_retweet_count_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_retweet_count_total, el Chi-cuadrado calculado es 67.601.400.000.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 116

Series: Valor-P de las variables de la regresión binomial negativa de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	31,79991	0
terms_ini_num_tweets	0,000254414	0,0106
terms_ini_retweet_count_mean	0,00294609	0,138
terms_ini_favorite_count_mean	0,332293	0,1515
terms_ini_followers_talking_rate	-41,6551	0,0176
terms_ini_user_num_followers_mean	0,0000669959	0,0946
terms_ini_user_num_tweets_mean	-0,0000956929	0,0168
terms_ini_user_age_mean	-0,000493736	0,6814
terms_ini_url_inclusion_rate	6,38437	0,0454
Modelo		0,0748

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 117

Series: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	32,9422	0
terms_ini_retweet_count_mean	0,0014843	0,0083

terms_ini_followers_talking_rate	-37,008	0,0412
Modelo		0,0439

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_total = exp(32,9422 + 0,0014843 * terms_ini_retweet_count_mean - 37,008 * terms_ini_followers_talking_rate)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

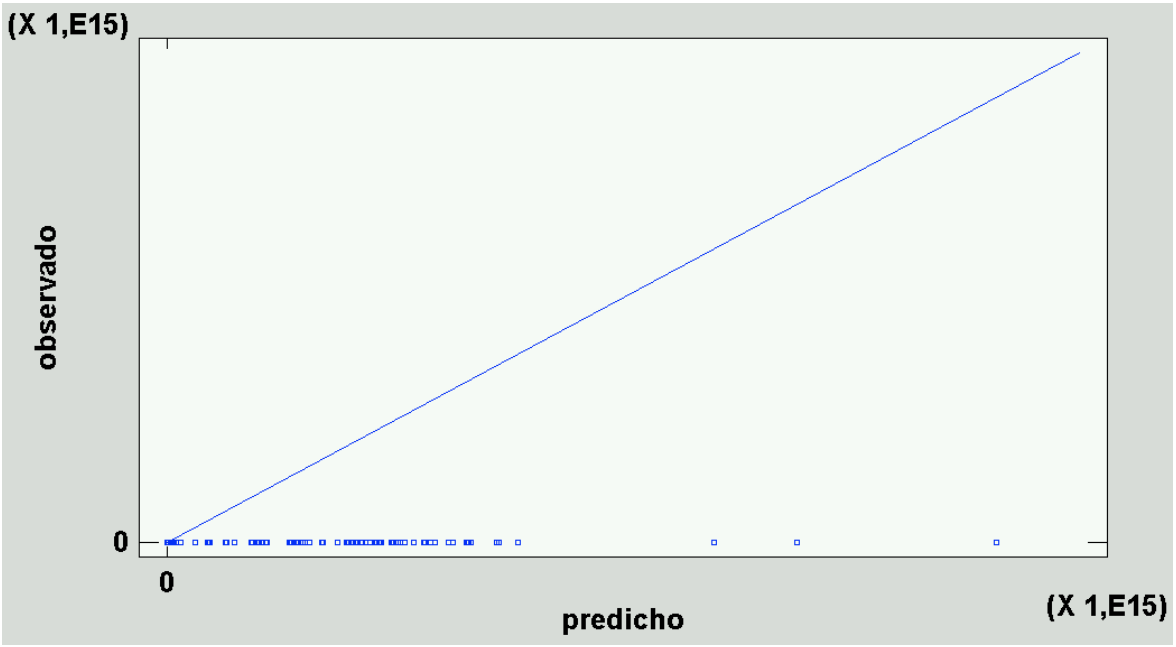


Figura 289. Series: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 3,30127% de la variabilidad de terms_end_retweet_count_total, mientras que el R-Cuadrado ajustado indica un 0,134049%. Este número bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones binomiales negativas de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 118

Series: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones binomiales negativas

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_mean	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

Se puede observar en la Tabla *118* que no todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que no todas son necesarias para las variables de éxito elegidas y que se pueden estudiar. Es el caso del promedio de favoritos, el número de tuits de los usuarios que participan y el promedio de edad en días de los usuarios que participan.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits, el promedio de retuits, la ratio de seguidores de la cuenta que participan, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits de la tendencia explican parte de los datos de páginas vistas únicas.
- El número de tuits explica parte de los datos del número de tuits 14 días después.

- El promedio de retuits y la ratio de seguidores de la cuenta del medio que participan explican en parte el número de retuits 14 días después.

6.1.3. Análisis de los artículos de la categoría Videojuegos

6.1.3.1. Variables de éxito

El objetivo de esta fase es la predicción de los valores de éxito mediante el resto de los indicadores. Por ello, es importante comenzar por analizar estos escenarios por separado para ver tanto sus características como tratar de describir su comportamiento anómalo, si lo tuvieran.

a) Páginas vistas únicas (total)

Este escenario de éxito se ve identificado por la columna uniquepageviews_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 35 valores con un rango de entre 5 y 146.

Presenta un sesgo estandarizado de 5,42251 y una curtosis estandarizada de 6,71391. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

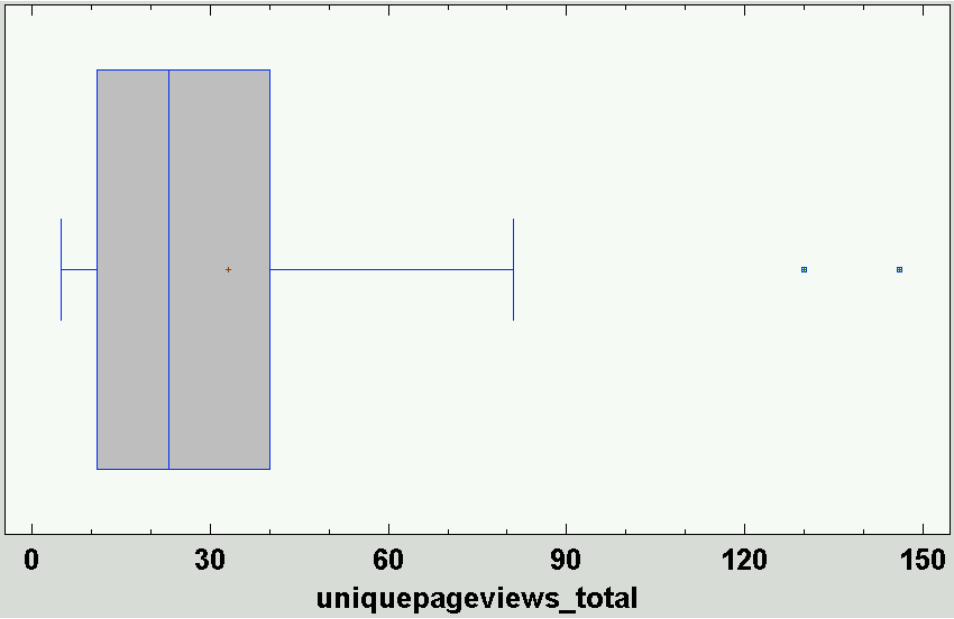


Figura 290. Videojuegos: Gráfico de Caja y Bigotes para el valor uniquepageviews_total

En la Figura 1 se puede comprobar que existen valores anómalos de tipo extremo de 130 páginas vistas o más.

b) AdSense eCPM (promedio)

Este escenario de éxito se identifica con la columna adsense_ecpm_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 0,17.

Presenta un sesgo estandarizado de 4,41179 y una curtosis estandarizada de 4,33278. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

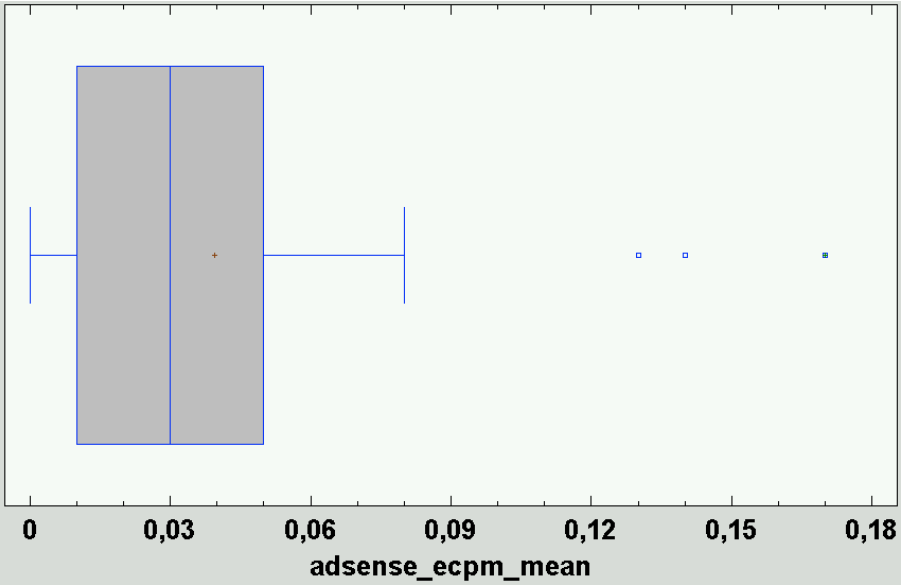


Figura 291. Videojuegos: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean

En la Figura 291 se puede observar que existen valores anómalos de tipo extremo de 0,17.

c) Duración de la visita (promedio)

Este escenario de éxito se identifica con la columna avgtimeonpage_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 554,86.

Presenta un sesgo estandarizado de 5,26691 y una curtosis estandarizada de 7,63204. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

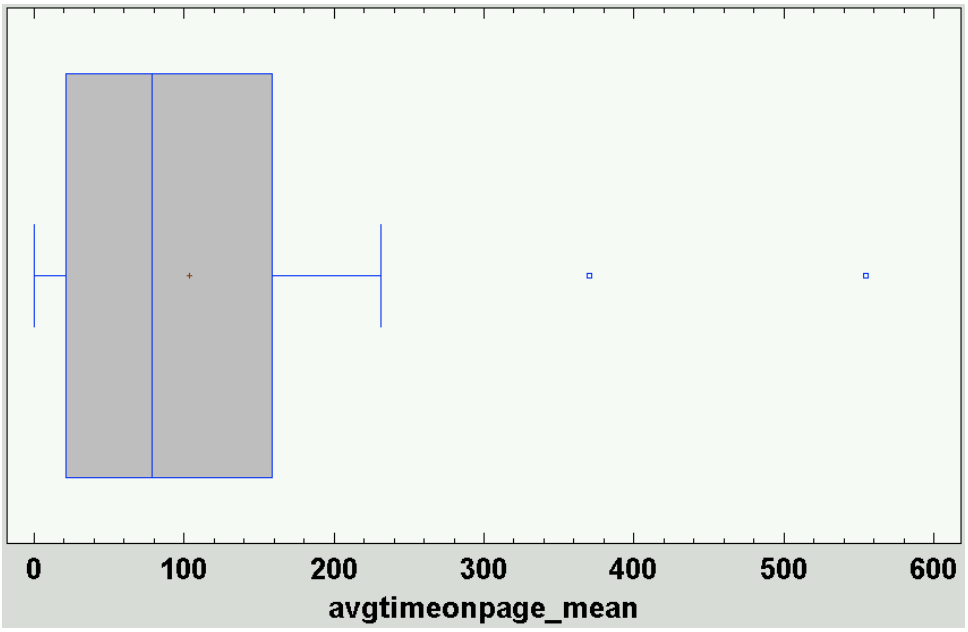


Figura 292. Videojuegos: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean

En la Figura 292 se puede observar que no existen valores anómalos de tipo extremo.

d) Páginas vistas por sesión (promedio)

Este escenario de éxito se identifica con la columna pageviewspersession_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0,59 y 3.

Presenta un sesgo estandarizado de 4,85948 y una curtosis estandarizada de 7,86926 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

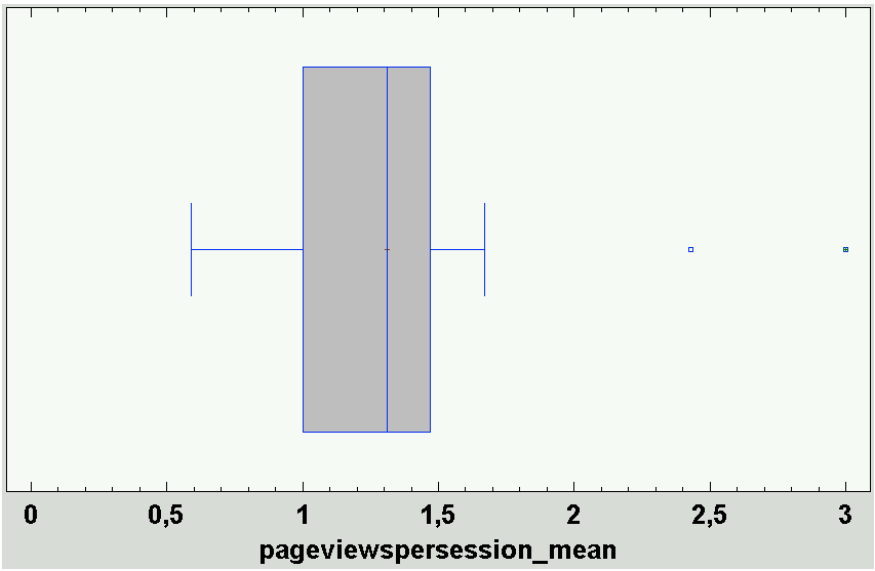


Figura 293. Videojuegos: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean

En la Figura 293 se puede observar que existe un valor anómalo de tipo extremo de 3.

e) Nº de retuits en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 1,74.

Presenta un sesgo estandarizado de -2,42363 y una curtosis estandarizada de 3,07178. Puesto que ambos valores estadísticos están dentro del rango de -2 a +2, indican que siguen una normalidad, es decir, que ambos se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

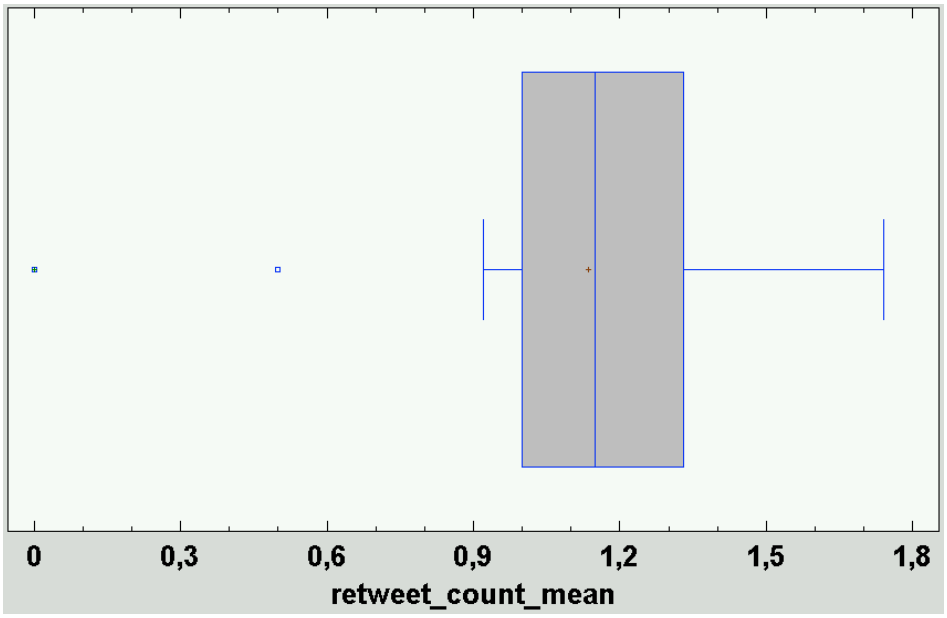


Figura 294. Videojuegos: Gráfico de Caja y Bigotes para el valor retweet_count_mean

En la Figura 294 se observa un valor anómalo de tipo extremo de 0.

f) Nº de favoritos en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 2,4.

Presenta un sesgo estandarizado de -0,572 y una curtosis estandarizada de 0,251316. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

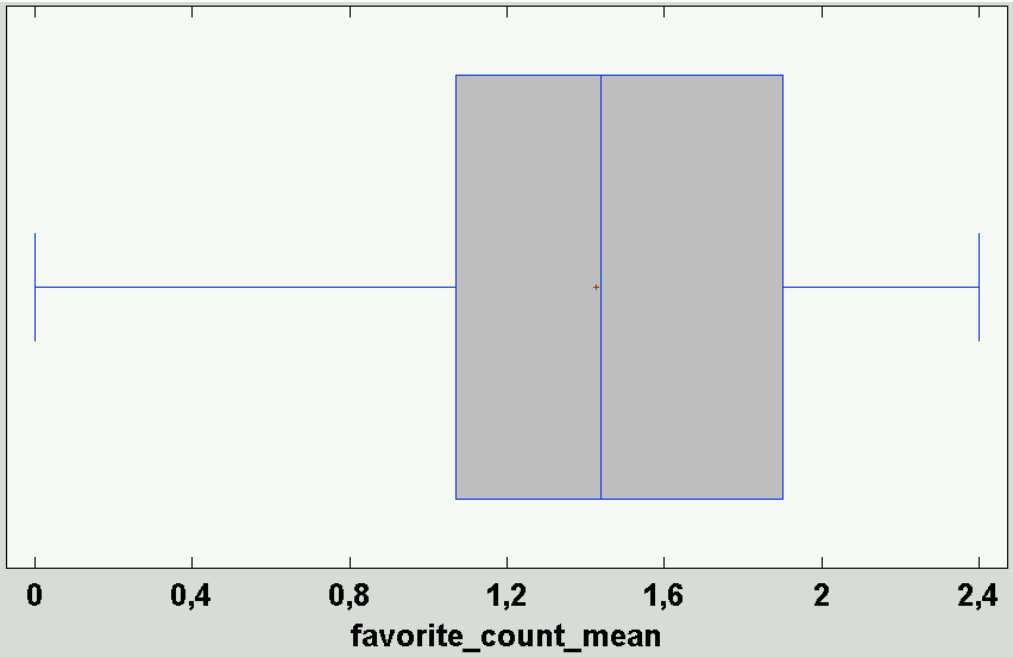


Figura 295. Videojuegos: Gráfico de Caja y Bigotes para el valor favorite_count_mean

En la Figura 295 no se observa ningún valor anómalo de tipo extremo.

g) Nº de tuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 35 valores con un rango de entre 0 y 15.038.

Presenta un sesgo estandarizado de 3,79158 y una curtosis estandarizada de 3,23143 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

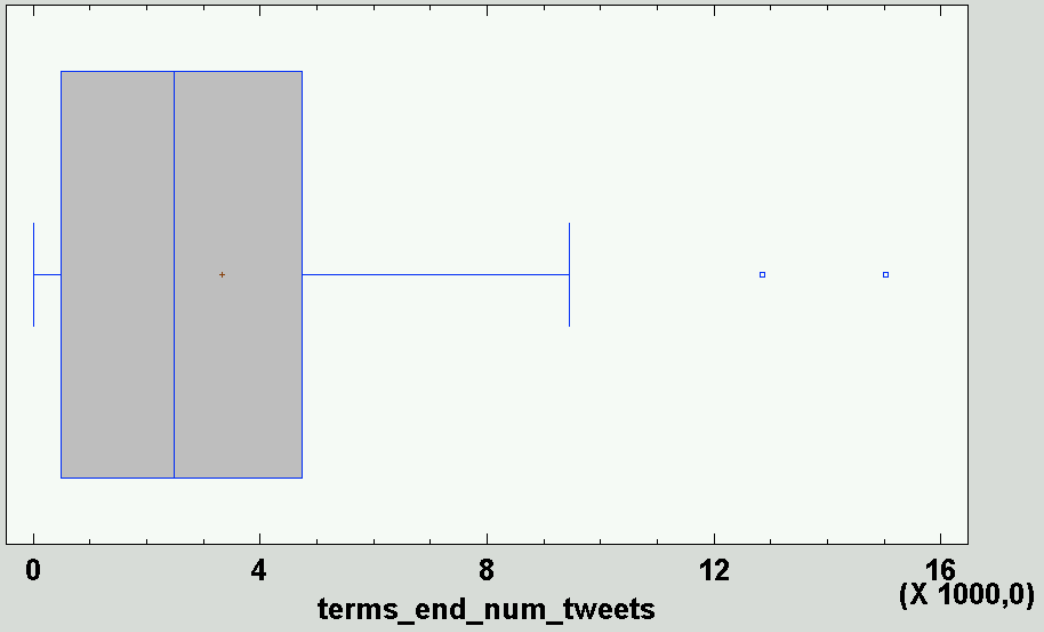


Figura 296. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets

En la Figura 296 no se puede observar ningún valor anómalo de tipo extremo.

h) Nº de retuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 20.336.400.

Presenta un sesgo estandarizado de 9,07754 y una curtosis estandarizada de 19,9158. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

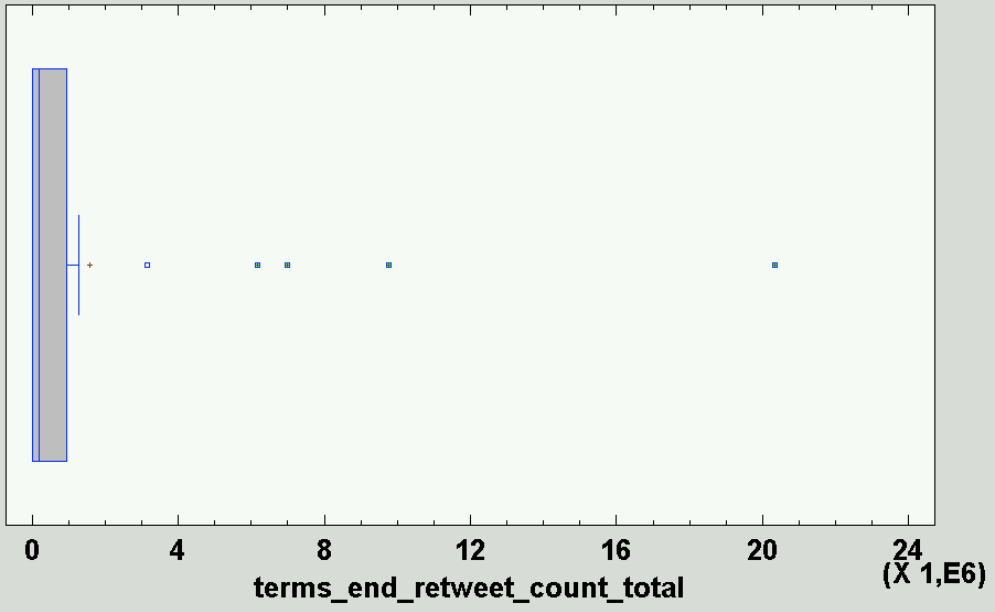


Figura 297. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total

En la Figura 297 se puede observar que existen valores anómalos de tipo extremo de 6.157.160 o más.

i) Nº de retuits de la tendencia 14 días después (promedio)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 3.038,46.

Presenta un sesgo estandarizado de 11,7063 y una curtosis estandarizada de 31,0566. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

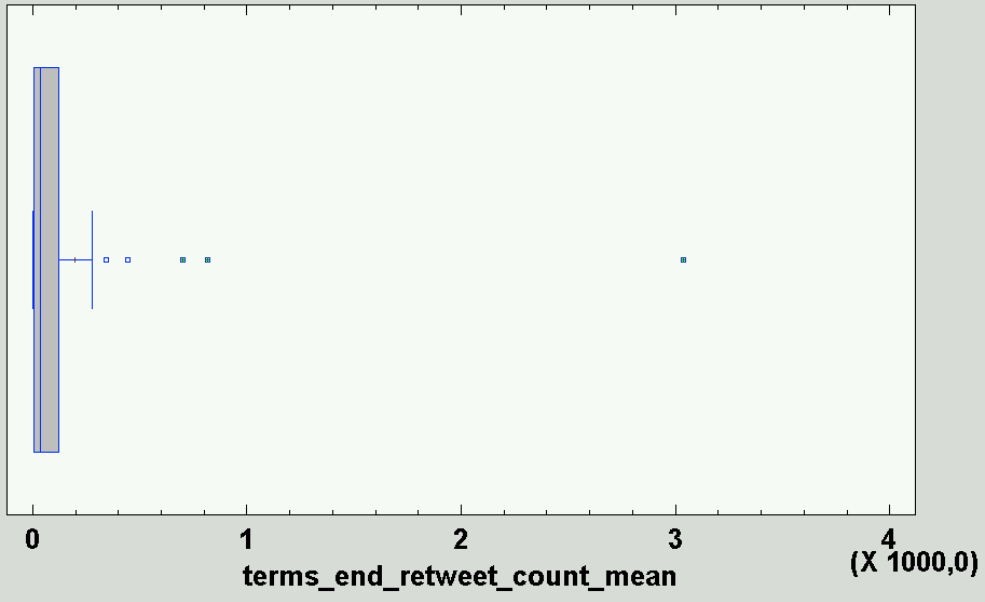


Figura 298. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean

En la Figura 298 se puede observar que existen valores anómalos de tipo extremo de 699,67 o más.

j) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de las variables de éxito. Se pueden observar los siguientes datos de estas:

Tabla 119

Videojuegos: Resumen estadístico de las variables de éxito

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
uniquepageviews_total	1.019,82	5,42251	6,71391
adsense_ecpm_mean	0,00156437	4,41179	4,33278
avgtimeonpage_mean	13.249,7	5,26691	7,63204
pageviewspersession_mean	0,191443	4,85948	7,86926

retweet_count_mean	0,119141	-2,42363	3,07178
favorite_count_mean	0,283953	-0,572	0,251316
terms_end_num_tweets	13.358.700	3,79158	3,23143
terms_end_retweet_count_total	15.472.300.000.000	9,07754	19,0158
terms_end_retweet_count_mean	281.255	11,7063	31,0566

Se puede observar en la Tabla 119 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2 salvo favorite_count_mean, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 120

Videojuegos: Resumen estadístico de las variables de éxito con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(uniquepageviews_total)	0,717545	0,45845	-0,528857
log(adsense_ecpm_mean)	0,692847	0,134004	-0,807969
log(avgtimeonpage_mean)	1,5674	-1,63849	0,436772

log(pageviewspersession_mean)	0,0892304	0,893811	2,51523
log(retweet_count_mean)	0,0772491	-2,98193	3,33006
favorite_count_mean	0,283953	-0,572	0,251316
log(terms_end_num_tweets)	4,39223	-3,46615	2,26087
log(terms_end_retweet_count_total)	13,76	-1,97794	0,131197
log(terms_end_retweet_count_mean)	4,23118	-0,570198	0,0206953

log(pageviewspersession_mean), log(retweet_count_mean) y log(terms_end_num_tweets) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. Sin embargo, mantienen valores mucho menores, próximos al rango de -2 a +2 y con una dispersión muy parecida, por lo que en este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

Nos quedamos, por tanto, con las variables que tengan menores sesgo y curtosis estandarizados de su forma original o transformación logarítmica.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5, la tabla queda así:

Tabla 121

Videojuegos: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	retweet_count_mean

terms_ini_followers_talking_rate	favorite_count_mean
terms_ini_user_num_followers_mean	log(terms_end_num_tweets)
terms_ini_user_num_tweets_mean	log(terms_end_retweet_count_total)
terms_ini_user_age_mean	log(terms_end_retweet_count_mean)
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, por tanto, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el número de tuits de la tendencia 14 días después, el número total de retuits de la tendencia 14 días después y el promedio de retuits de la tendencia 14 días después. También incluye la versión original del promedio de retuits en la cuenta del medio y el promedio de favoritos en la cuenta del medio.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

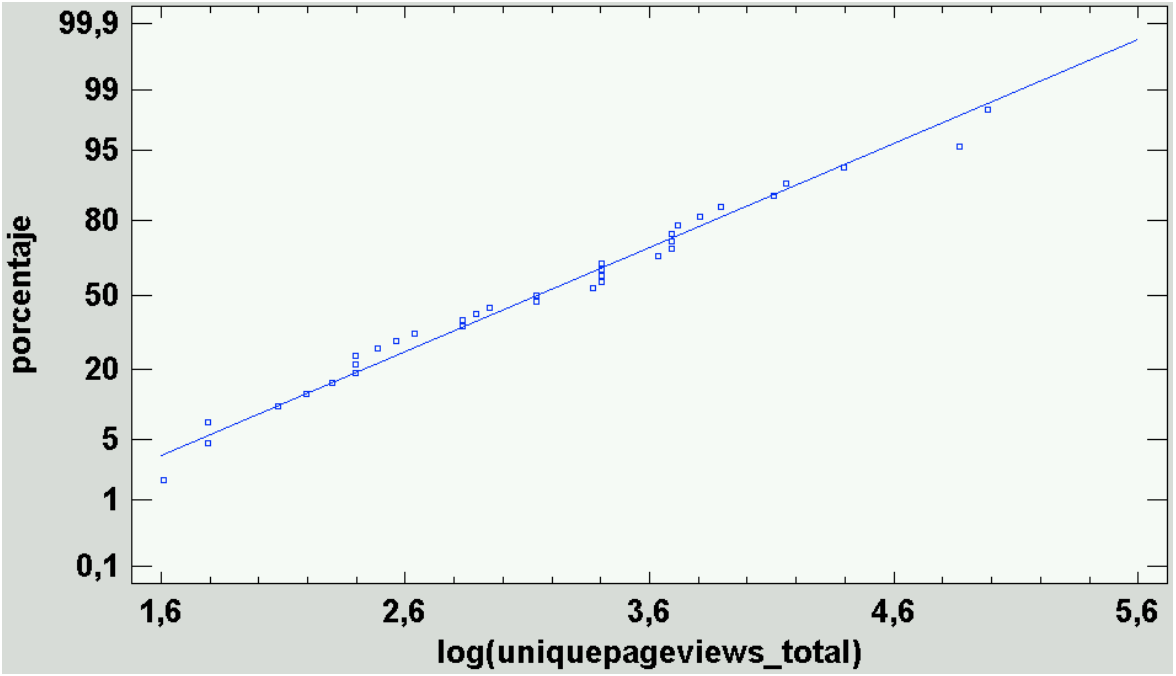


Figura 299. Videojuegos: Gráfico de probabilidad normal de la variable log(uniquepageviews_total)

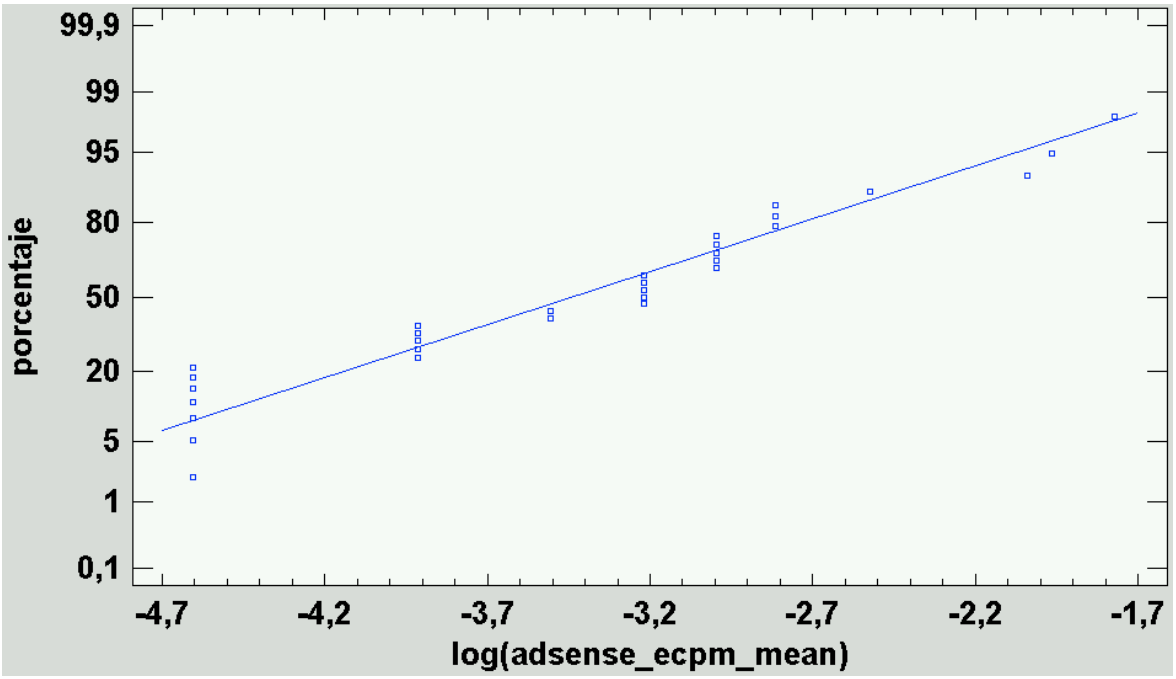


Figura 300. Videojuegos: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean)

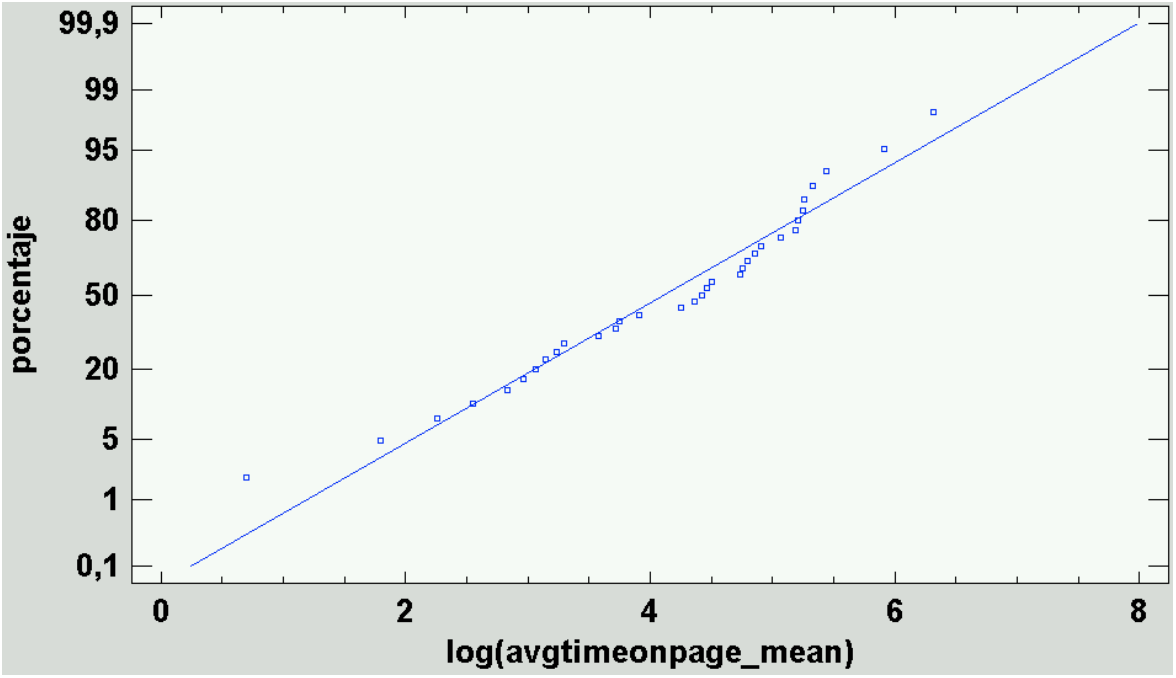


Figura 301. Videojuegos: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean)

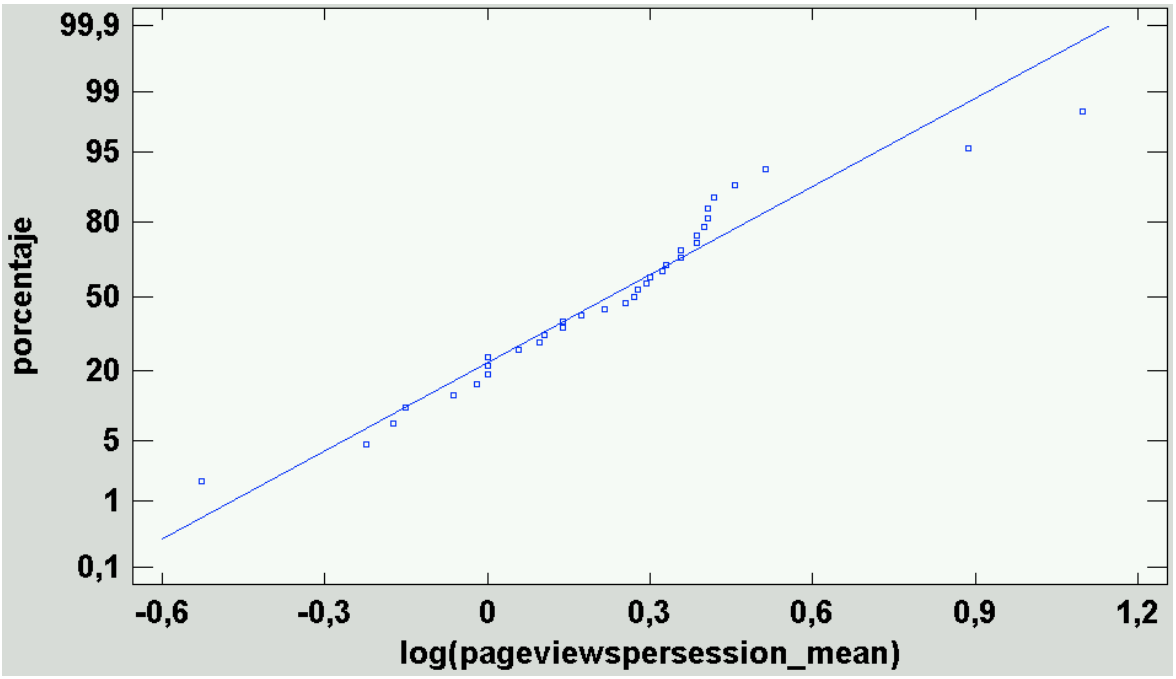


Figura 302. Videojuegos: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean)

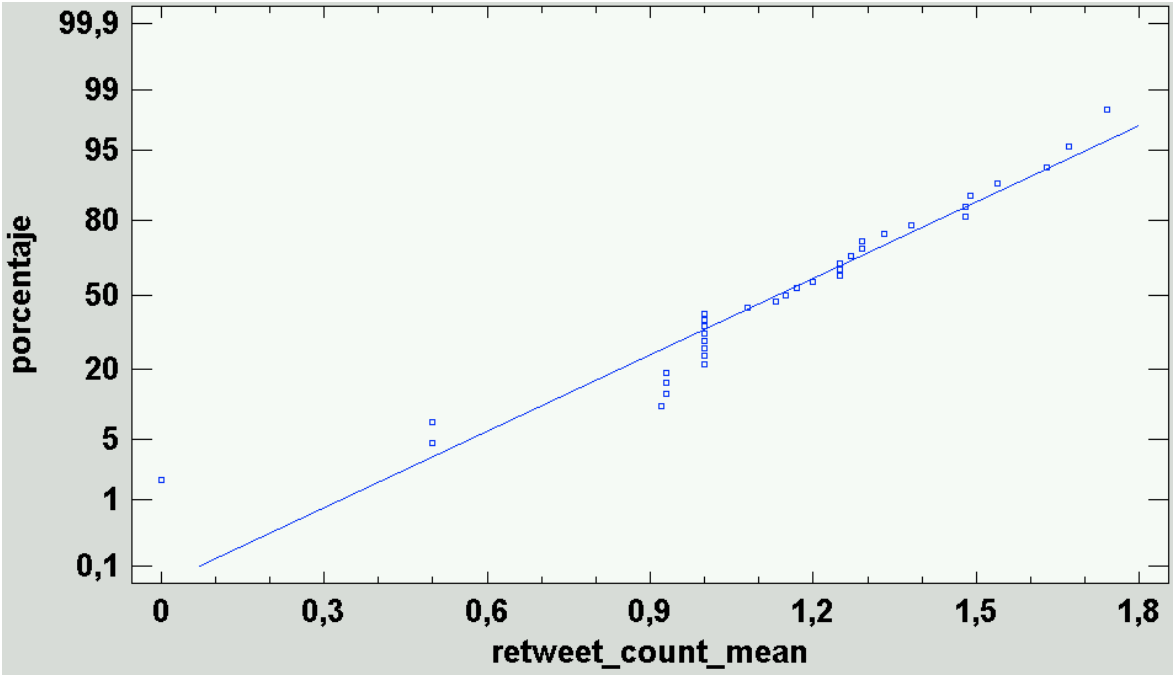


Figura 303. Videojuegos: Gráfico de probabilidad normal de la variable retweet_count_mean

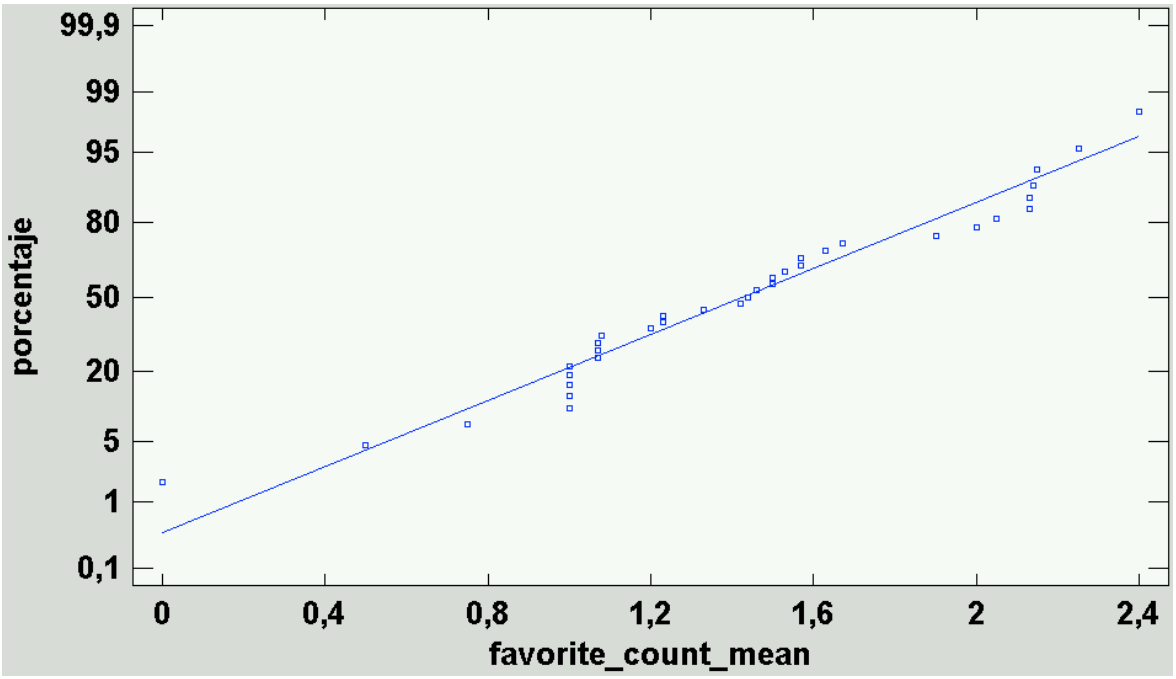


Figura 304. Videojuegos: Gráfico de probabilidad normal de la variable favorite_count_mean

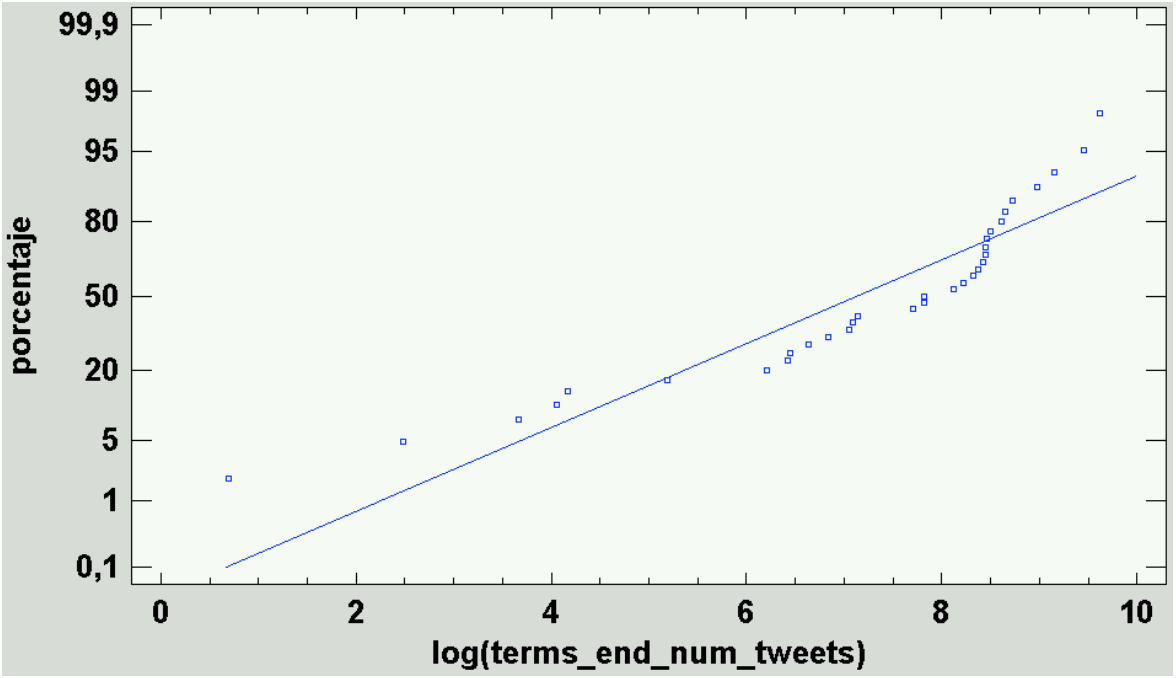


Figura 305. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_num_tweets)

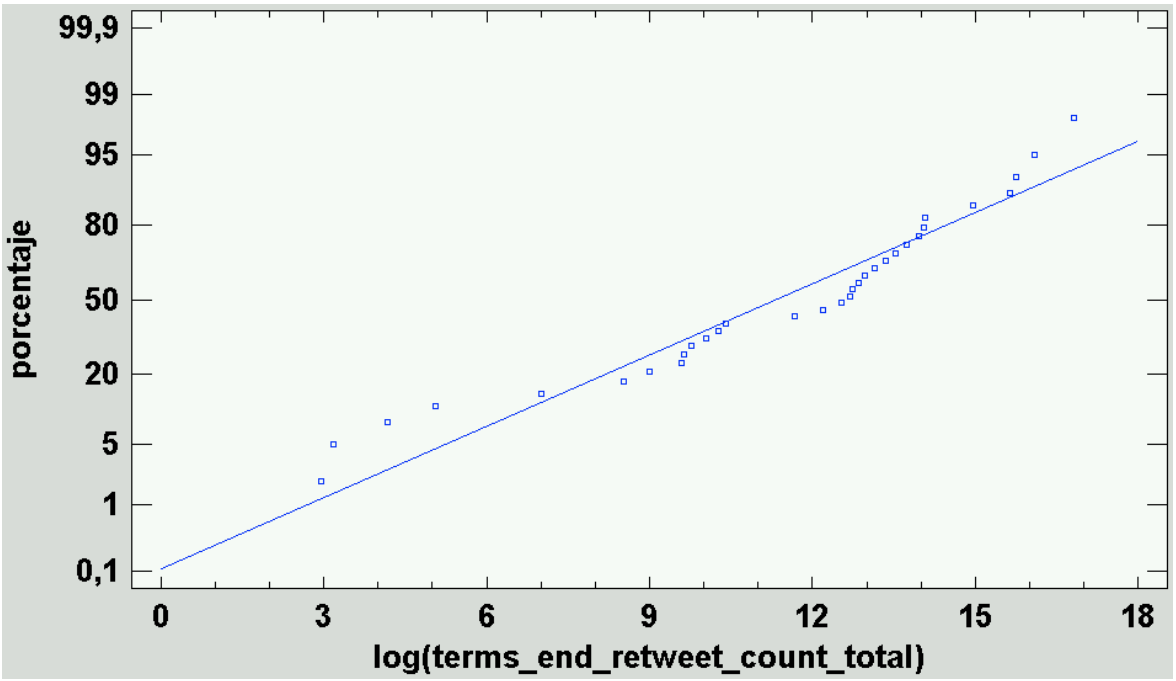


Figura 306. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total)

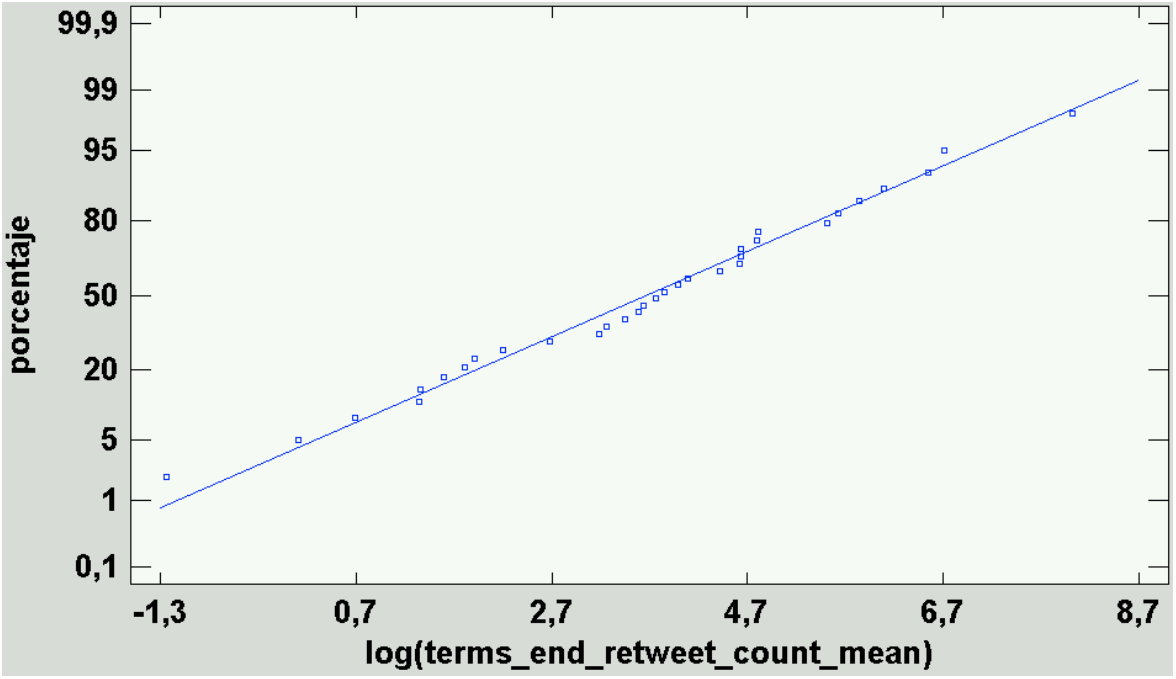


Figura 307. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

k) Filtro de alta correlación

Puesto que se han propuesto más de una variable de éxito, es conveniente evaluar si están asociadas mediante una fuerte correlación. Si fuera ese el caso, el análisis de este estudio

se simplificaría ya que las conclusiones de una variable servirían para las que estén fuertemente correlacionadas con ella, lo que permitiría dedicar menos recursos a la predicción y toma de decisiones.

Para ello, es necesario realizar un análisis multivariado de las variables de éxito con su correspondiente transformación logarítmica, cuya matriz de correlaciones Pearson se puede observar en la Figura 308:

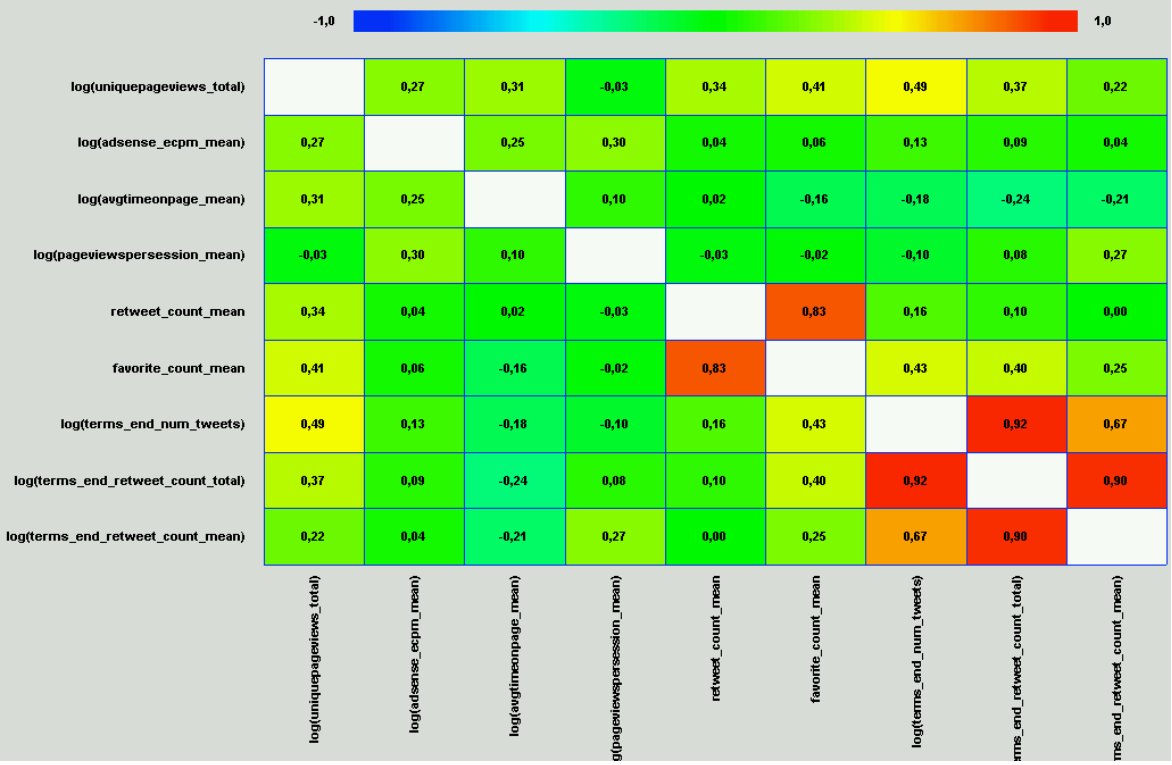


Figura 308. Videojuegos: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- retweet_count_mean y favorite_count_mean tienen un coeficiente de correlación de 0,8277 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_end_num_tweets) y log(terms_end_retweet_count_total) tienen un coeficiente de correlación de 0,9172 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_end_retweet_count_total) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,9001 y un valor-P cercano a 0, por lo que sí existe

una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

- log(terms_end_num_tweets) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,6738 y un valor-P cercano a 0, por lo que no existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto log(terms_end_num_tweets) y log(terms_end_retweet_count_mean), porque log(terms_end_retweet_count_total) está fuertemente correlacionada con estas dos, pero entre ellas no lo están. Se elige también favorite_count_mean por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que retweet_count_mean.

La tabla de variables quedaría como sigue:

Tabla 122

Videojuegos: Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	favorite_count_mean
terms_ini_followers_talking_rate	log(terms_end_num_tweets)
terms_ini_user_num_followers_mean	log(terms_end_retweet_count_mean)
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, finalmente, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el número de tuits de la tendencia 14 días después y el promedio de retuits de la tendencia 14 días después. También incluye la versión original del promedio del promedio de favoritos en la cuenta del medio.

6.1.3.2. Variables de predicción

Se parte de un conjunto de datos con diez características, por lo que es conveniente tratar de reducir la dimensión del conjunto, restableciendo la varianza sin modificar la información relevante de los datos en sí. Esto posibilitará que se reduzca el tiempo y el coste de la computación y facilita la visualización y el análisis de los datos. Además, es una condición necesaria para aplicar la regresión lineal múltiple (Anon., 2017).

a) Número de tuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 35 valores con un rango de entre 0 y 15.305.

Presenta un sesgo estandarizado de 3,5469 y una curtosis estandarizada de 2,2705. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

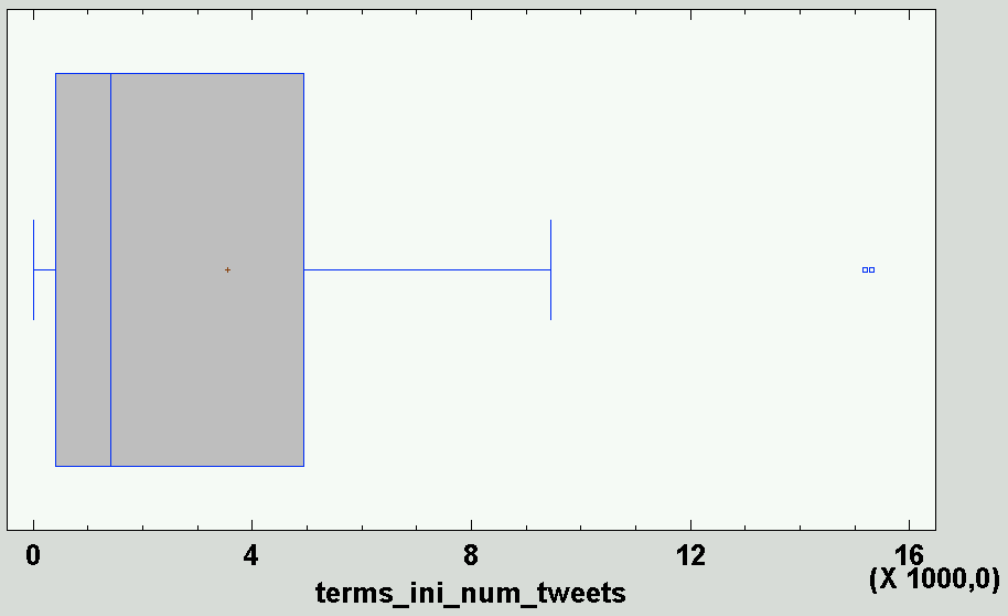


Figura 309. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets

En la Figura 309 se puede observar que no existen valores anómalos de tipo extremo.

b) Número de retuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 35 valores con un rango de entre 0 y 23.578.400.

Presenta un sesgo estandarizado de 8,74337 y una curtosis estandarizada de 15,7851. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

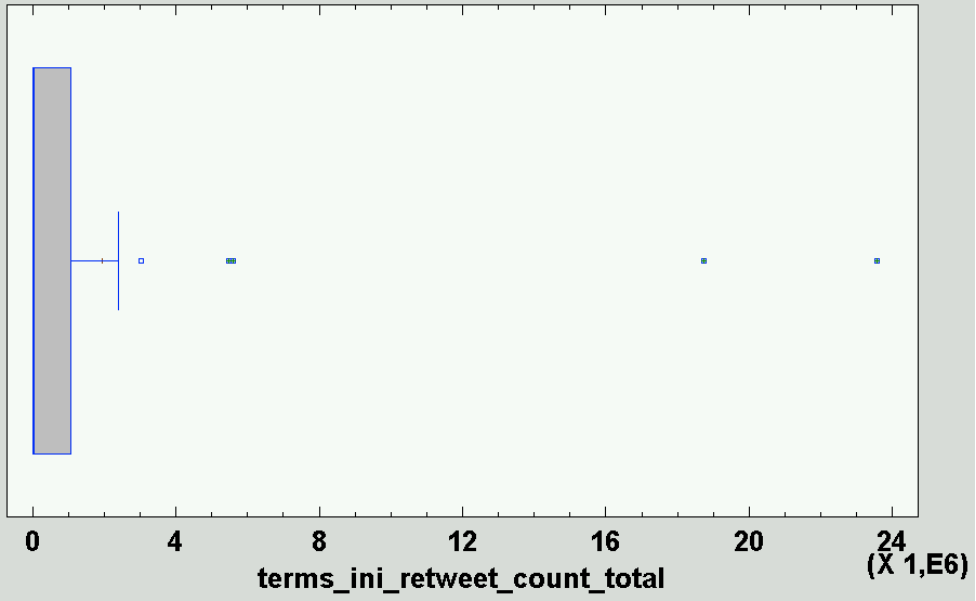


Figura 310. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total

En la Figura 310 se puede observar que existen valores anómalos de tipo extremo de 5.477.440 o más.

c) Número de retuits de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 1.919,35.

Presenta un sesgo estandarizado de 7,78759 y una curtosis estandarizada de 13,9376. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

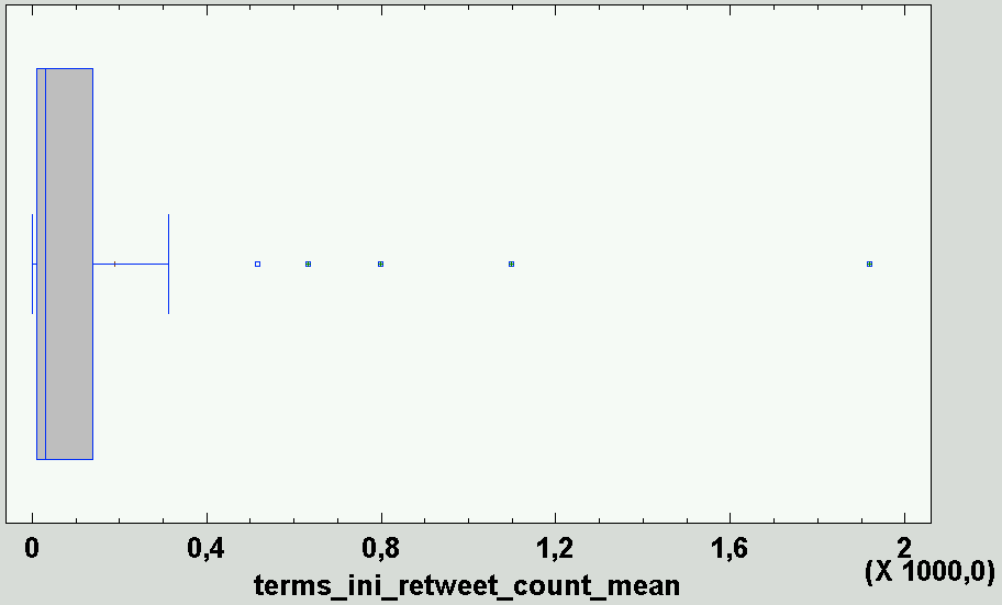


Figura 311. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean

En la Figura 311 se puede observar que existen valores anómalos de tipo extremo de 633,09 o más.

d) Número de favoritos de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 35 valores con un rango de entre 0 y 43.410.

Presenta un sesgo estandarizado de 3,01942 y una curtosis estandarizada de 1,23191. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

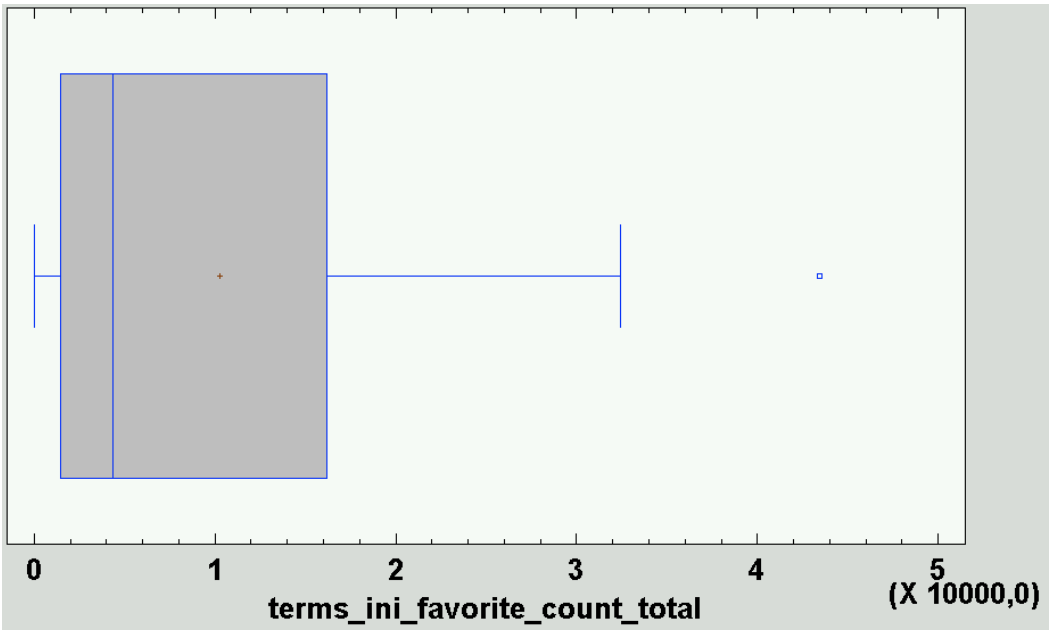


Figura 312. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total

En la Figura 312 se puede observar que no existen valores anómalos de tipo extremo.

e) Número de favoritos de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 8,96.

Presenta un sesgo estandarizado de 3,33923 y una curtosis estandarizada de 4,66576. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

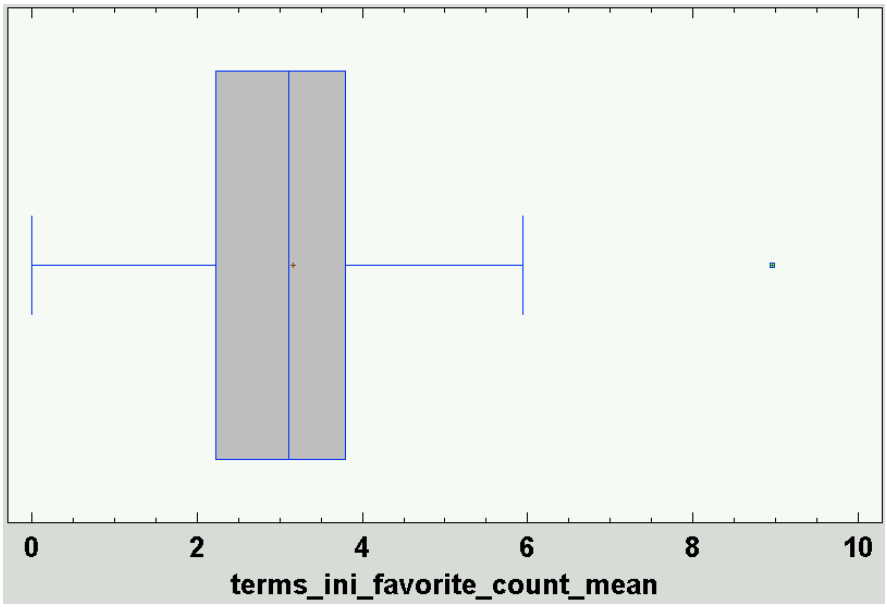


Figura 313. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean

En la Figura 313 se puede observar que existe un valor anómalo de tipo extremo de 8,96.

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_followers_talking_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 0,07.

Presenta un sesgo estandarizado de 6,77774 y una curtosis estandarizada de 13,6342. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

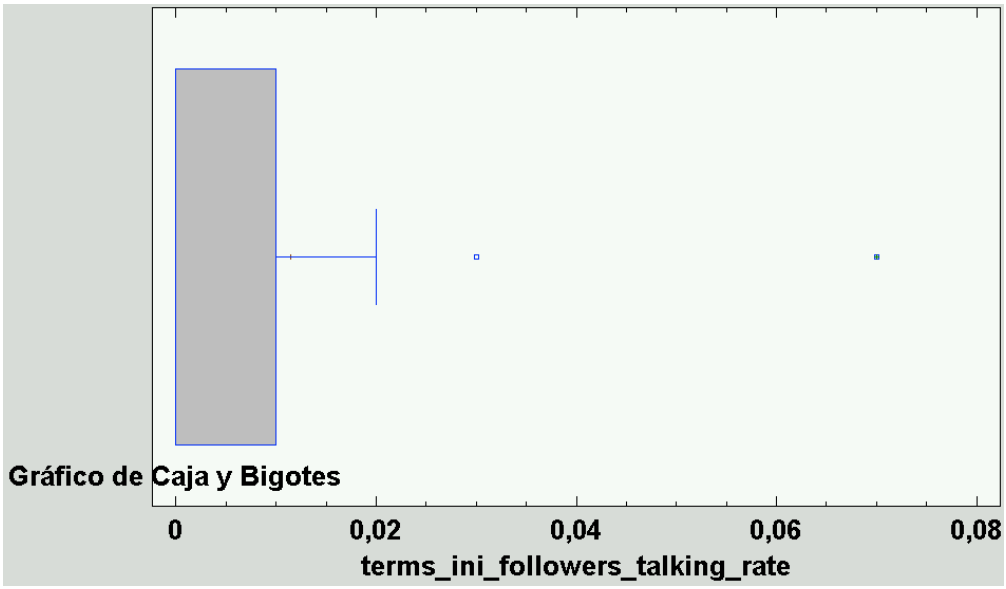


Figura 314. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate

En la Figura 314 se puede observar que existe un valor anómalo de tipo extremo de 0,07.

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_followers_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 6,33523.

Presenta un sesgo estandarizado de 6,33523 y una curtosis estandarizada de 10,864. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

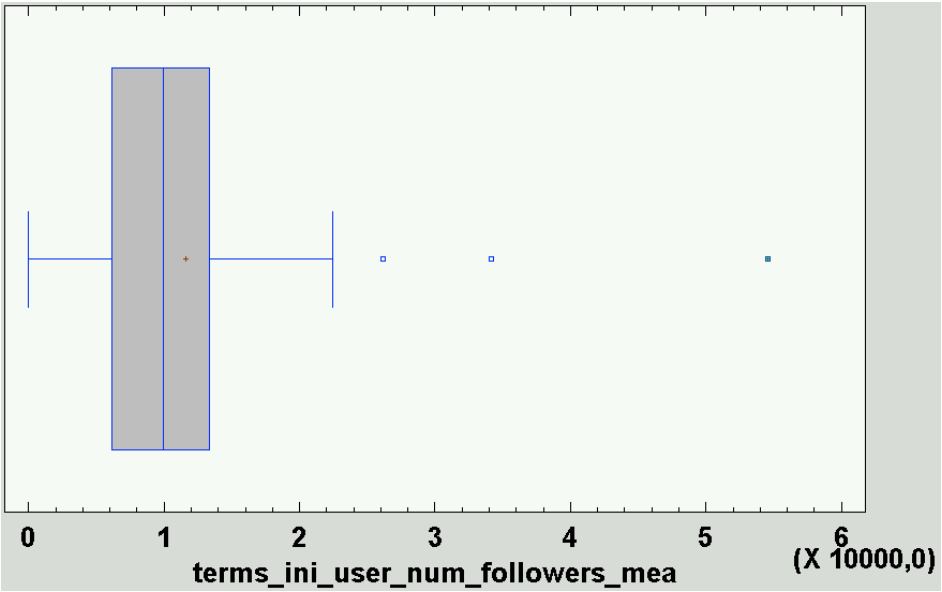


Figura 315. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean

En la Figura 315 se puede observar que existe un valor anómalo de tipo extremo de 54.538,7.

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_tweets_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 57.803,8.

Presenta un sesgo estandarizado de -1,0768 y una curtosis estandarizada de 1,31301. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

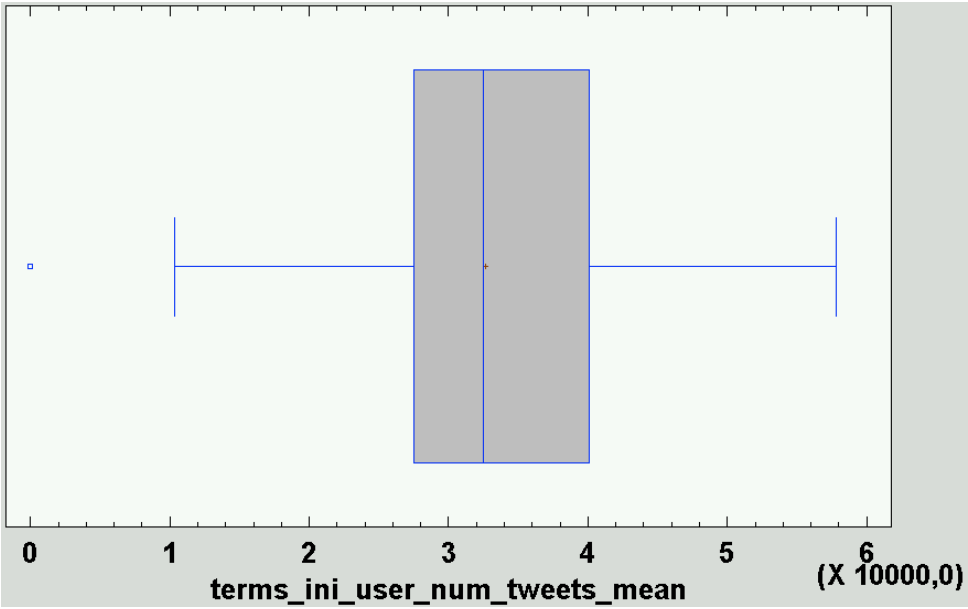


Figura 316. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean

En la Figura 316 se puede observar que no existen valores anómalos de tipo extremo.

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_age_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 3.784,9.

Presenta un sesgo estandarizado de -2,38489 y una curtosis estandarizada de 7,2801. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

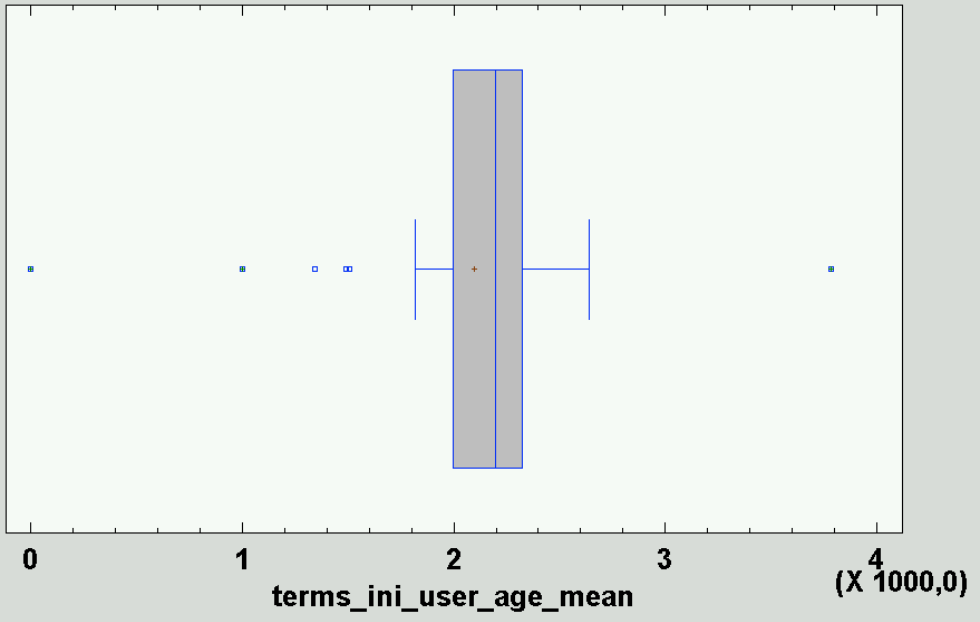


Figura 317. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean

En la Figura 317 se puede observar que existen valores anómalos de tipo extremo de 999,5 o menos, y 3.784,9.

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial

Esta variable de predicción se identifica con la columna terms_ini_url_inclusion_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 35 valores con un rango de entre 0 y 0,88.

Presenta un sesgo estandarizado de 0,634859 y una curtosis estandarizada de 0,493157. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

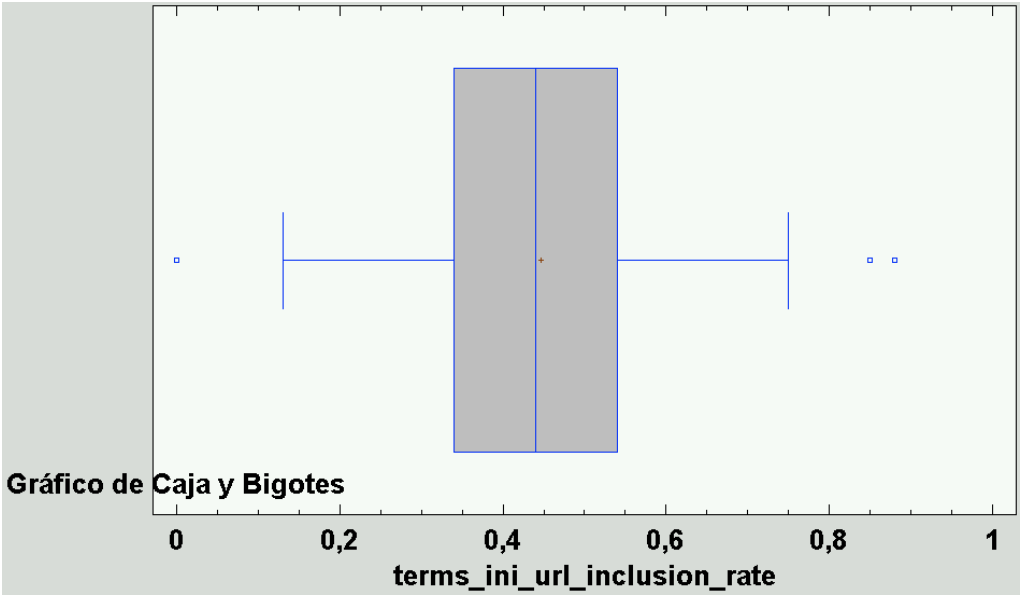


Figura 318. Videojuegos: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate

En la Figura 318 se puede observar que no existen valores anómalos de tipo extremo.

k) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de todas las variables de predicción y éxito. De esta manera se obtienen todos los datos de las variables del modelo en un mismo análisis. Se pueden observar los siguientes datos de estas:

Tabla 123

Videojuegos: Resumen estadístico de las variables de predicción

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
terms_ini_num_tweets	16.987.500	3,5469	2,2705
terms_ini_retweet_count_total	25.356.700.000.000	8,74337	15,7851

terms_ini_retweet_count_mean	152.857	7,78759	13,9376
terms_ini_favorite_count_total	125.911.000	3,01942	1,23191
terms_ini_favorite_count_mean	2,62785	3,33923	4,66576
terms_ini_followers_talking_rate	0,000171429	6,77774	13,6342
terms_ini_user_num_followers_mean	103.845.000	6,33523	10,864
terms_ini_user_num_tweets_mean	134.967.000	-1,0768	1,31301
terms_ini_user_age_mean	335227	-2,38489	7,2801
terms_ini_url_inclusion_rate	0,0369829	0,634859	0,493157

Se puede observar en la Tabla 123 que todas las variables salvo terms_ini_user_num_tweets_mean y terms_ini_url_inclusion_rate presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple, es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

También se transforma logarítmicamente terms_ini_user_num_tweets_mean ya que tiene una varianza tan grande que es necesario una mucho menor para que se cumpla la condición de homocedasticidad.

Tabla 124

Videojuegos: Resumen estadístico de las variables de predicción con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(terms_ini_num_tweets)	3,43245	-1,8986	-0,186368
log(terms_ini_retweet_count_total)	12,9988	-0,878894	-0,7802
log(terms_ini_retweet_count_mean)	3,48029	0,371223	-0,468748
log(terms_ini_favorite_count_total)	3,77106	-2,52709	0,525678
log(terms_ini_favorite_count_mean)	0,189587	0,501117	0,2600095
log(terms_ini_followers_talking_rate)	0,261956	3,61327	3,136
log(terms_ini_user_num_followers_mean)	0,54283	-0,32214	0,873393
log(terms_ini_user_num_tweets_mean)	0,12395	-2,82932	2,91542
log(terms_ini_user_age_mean)	0,0500846	-2,47542	4,84944
terms_ini_url_inclusion_rate	0,0369829	0,634859	0,493157

log(terms_ini_favorite_count_total), log(terms_ini_followers_talking_rate), log(terms_ini_user_num_tweets_mean) y log(terms_ini_user_age_mean) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. En este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

También se puede apreciar que los valores de varianza son bastante parecidos salvo en el caso de log(terms_ini_retweet_count_total), por lo que se cumple la condición de

homocedasticidad menos en esa variable. log(terms_ini_retweet_count_total) se elimina del modelo actual para que dicho requisito se cumpla.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5, la tabla queda así:

Tabla 125

Videojuegos: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_num_tweets)	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_total)	log(avgtimeonpage_mean)
log(terms_ini_favorite_count_mean)	log(pageviewspersession_mean)
log(terms_ini_followers_talking_rate)	favorite_count_mean
log(terms_ini_user_num_followers_mean)	log(terms_end_num_tweets)
log(terms_ini_user_num_tweets_mean)	log(terms_end_retweet_count_mean)
log(terms_ini_user_age_mean)	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo. También incluye la versión original de: el número total de tuits de la tendencia y el promedio de edad en días de los usuarios que participan.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

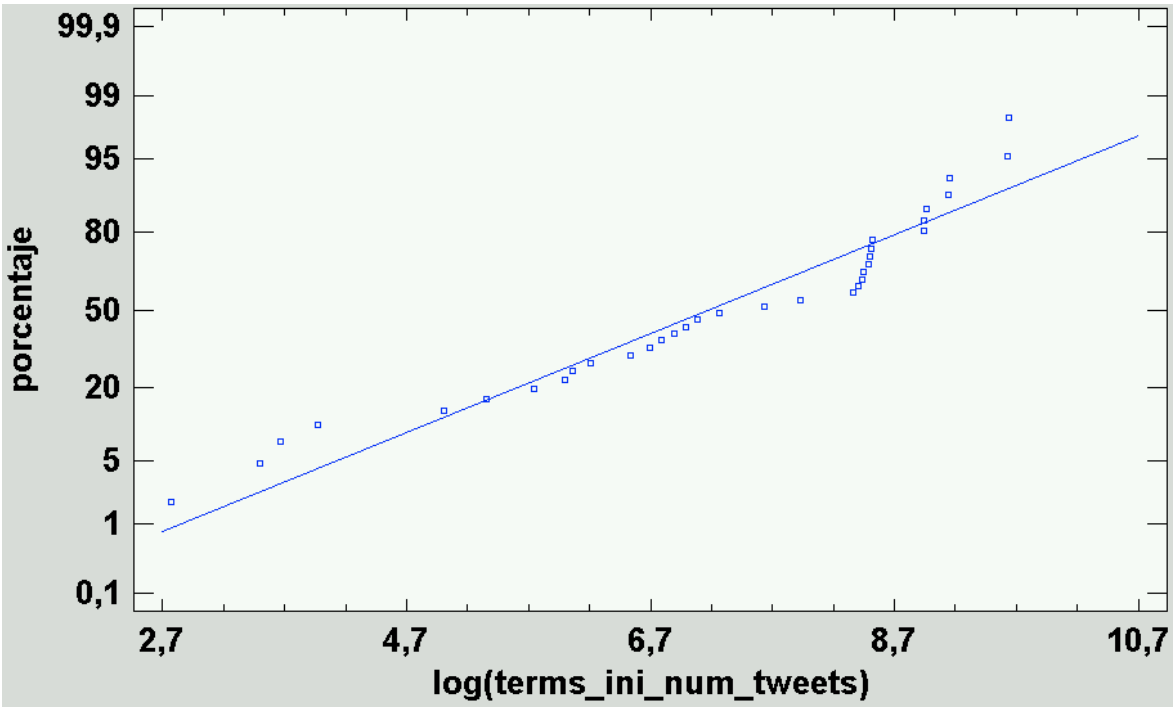


Figura 319. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_num_tweets)

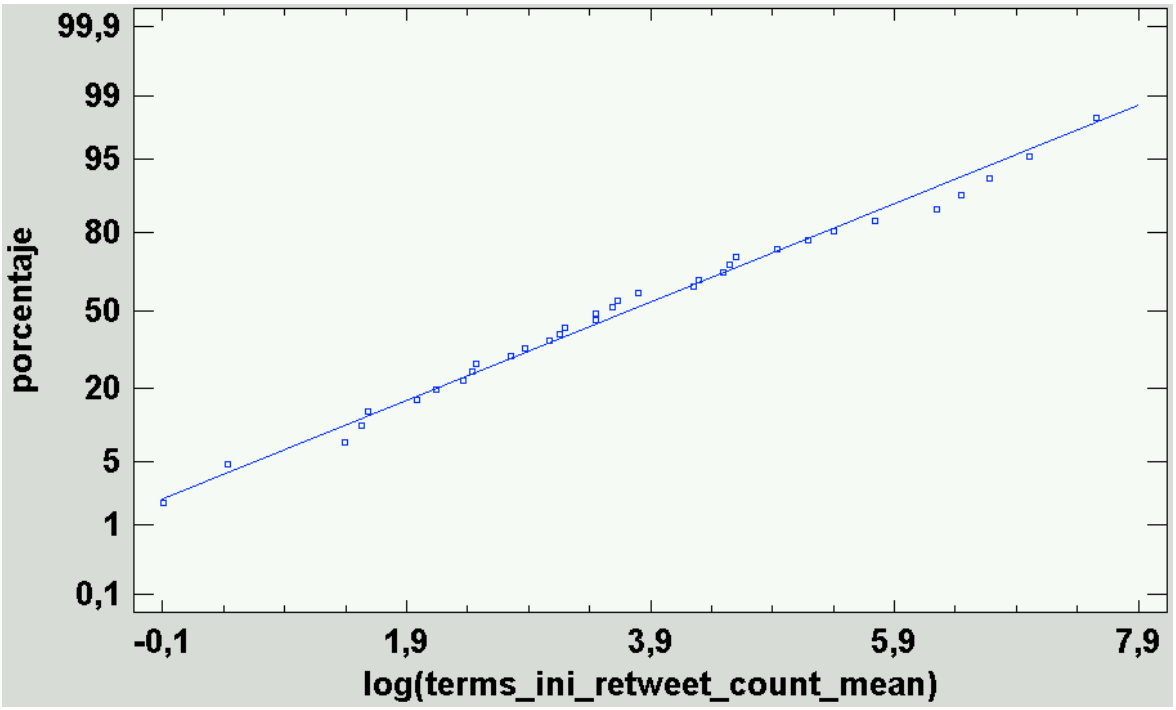


Figura 320. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean)

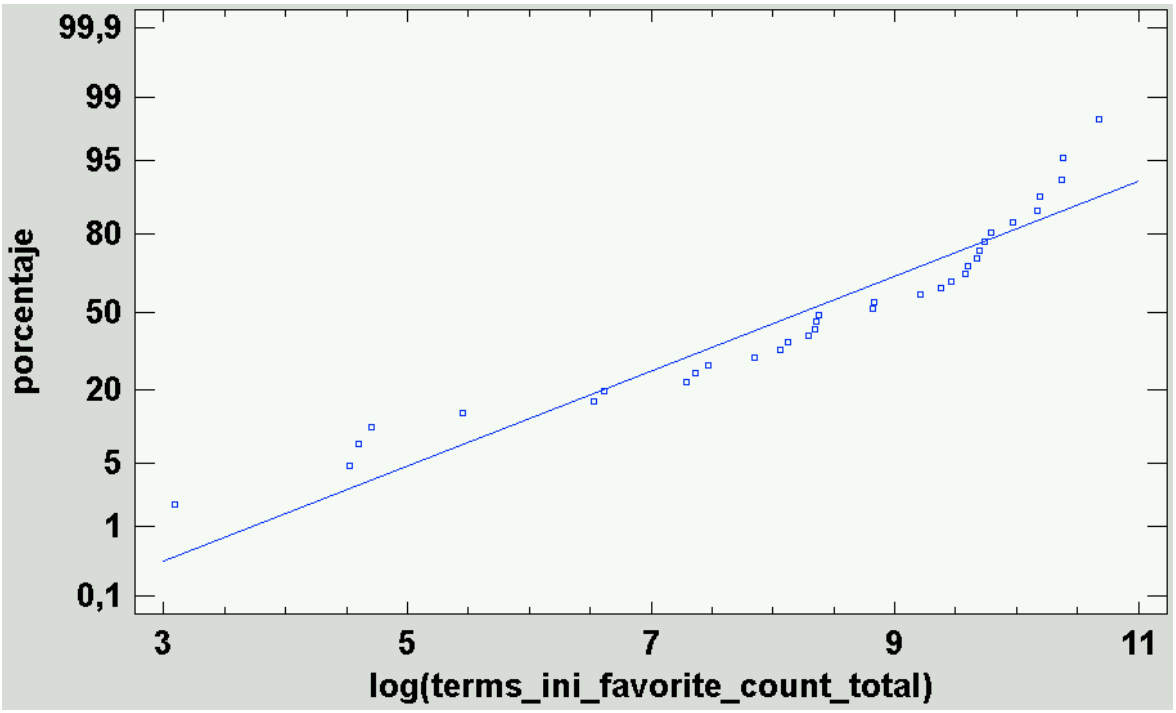


Figura 321. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_total)

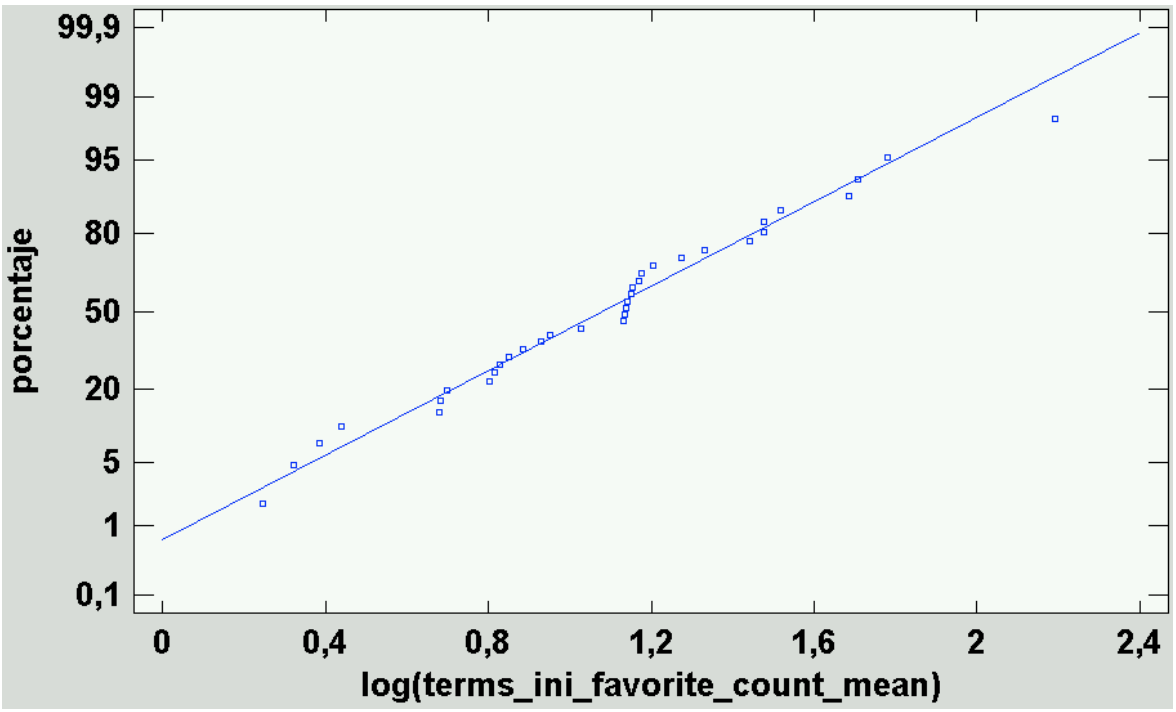


Figura 322. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean)

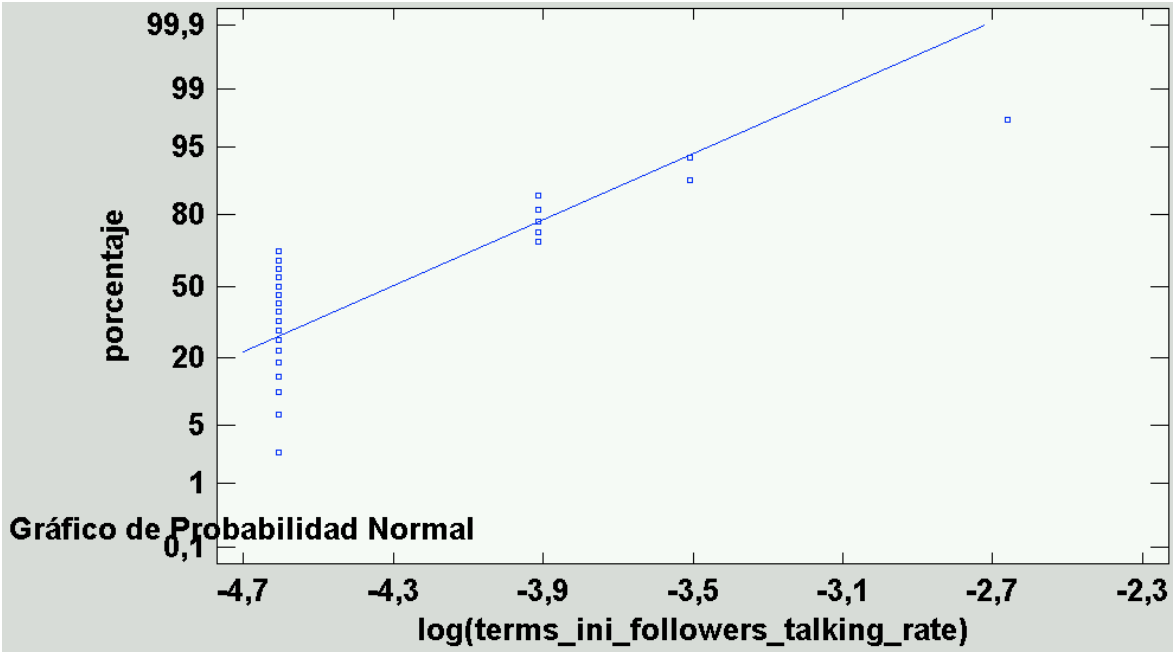


Figura 323. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate)

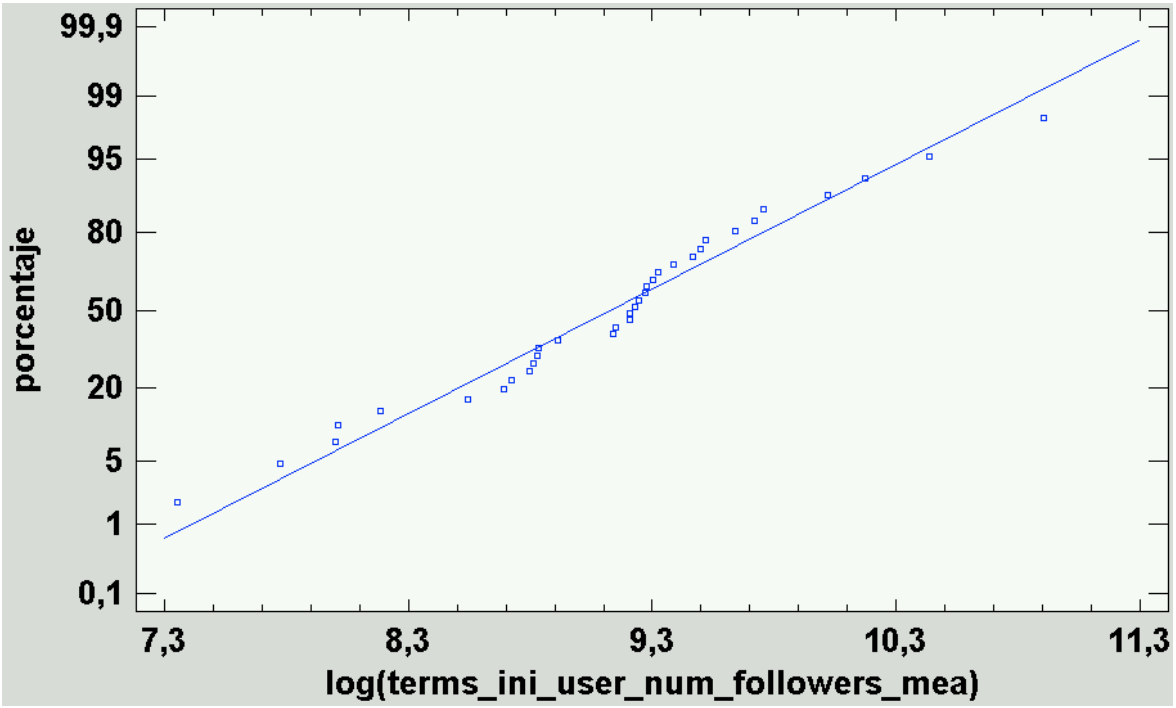


Figura 324. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean)

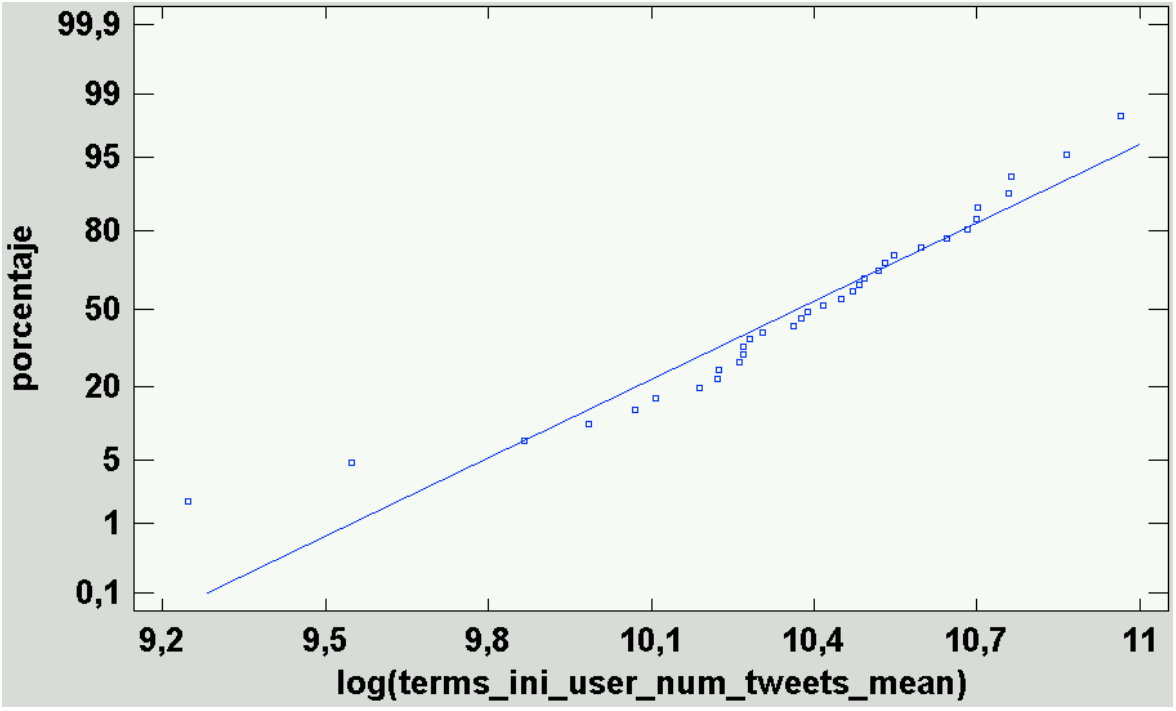


Figura 325. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean)

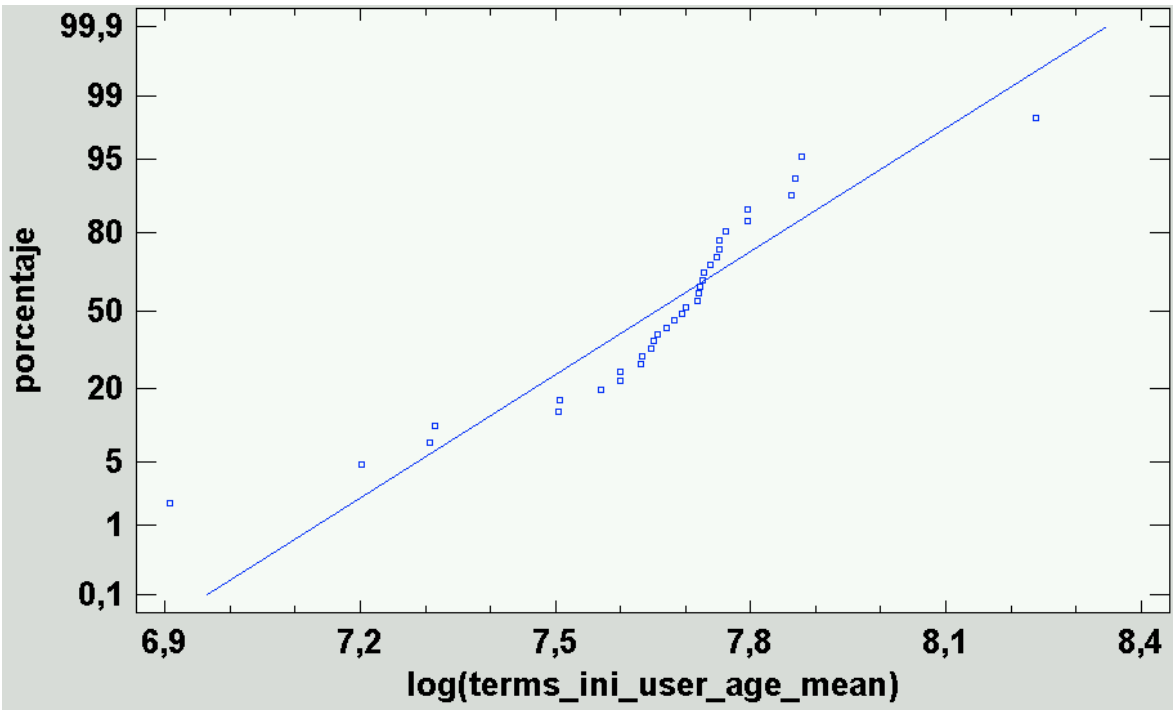


Figura 326. Videojuegos: Gráfico de probabilidad normal de la variable log(terms_ini_user_age_mean)

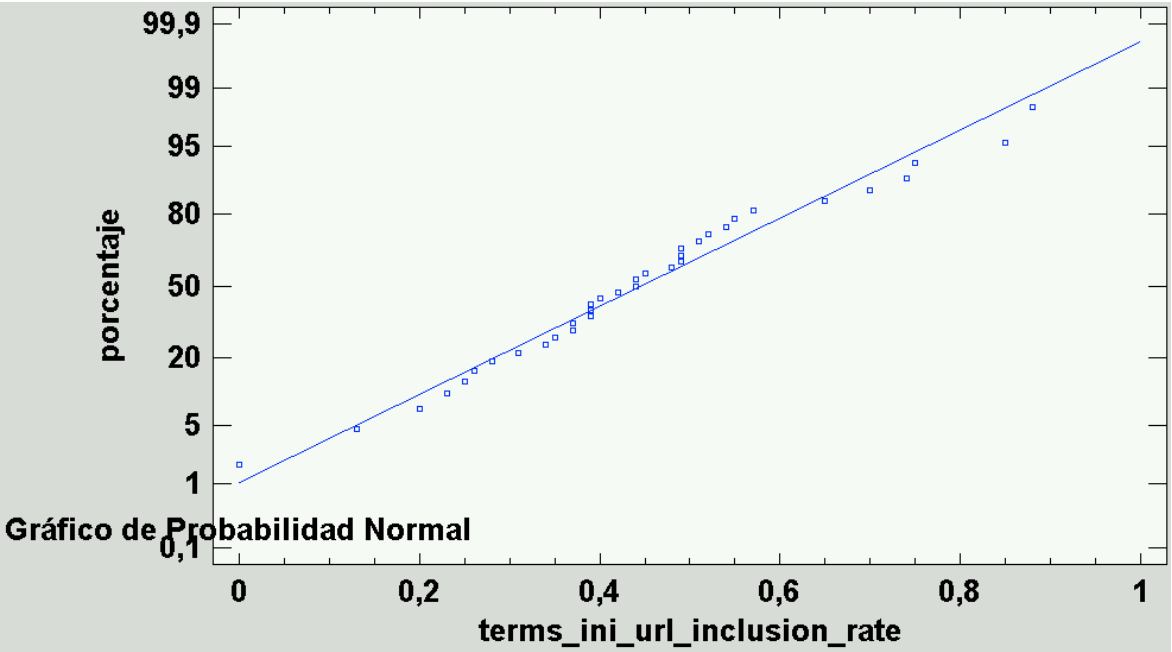


Figura 327. Videojuegos: Gráfico de probabilidad normal de la variable terms_ini_url_inclusion_rate

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

l) Filtro de alta correlación (colinealidad)

Antes de proceder, es conveniente analizar la correlación que existe entre las variables con las que contamos para el modelo. Puesto que estas han sido normalizadas en el apartado anterior, dicho análisis de correlación Pearson se realizará con su transformación logarítmica.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

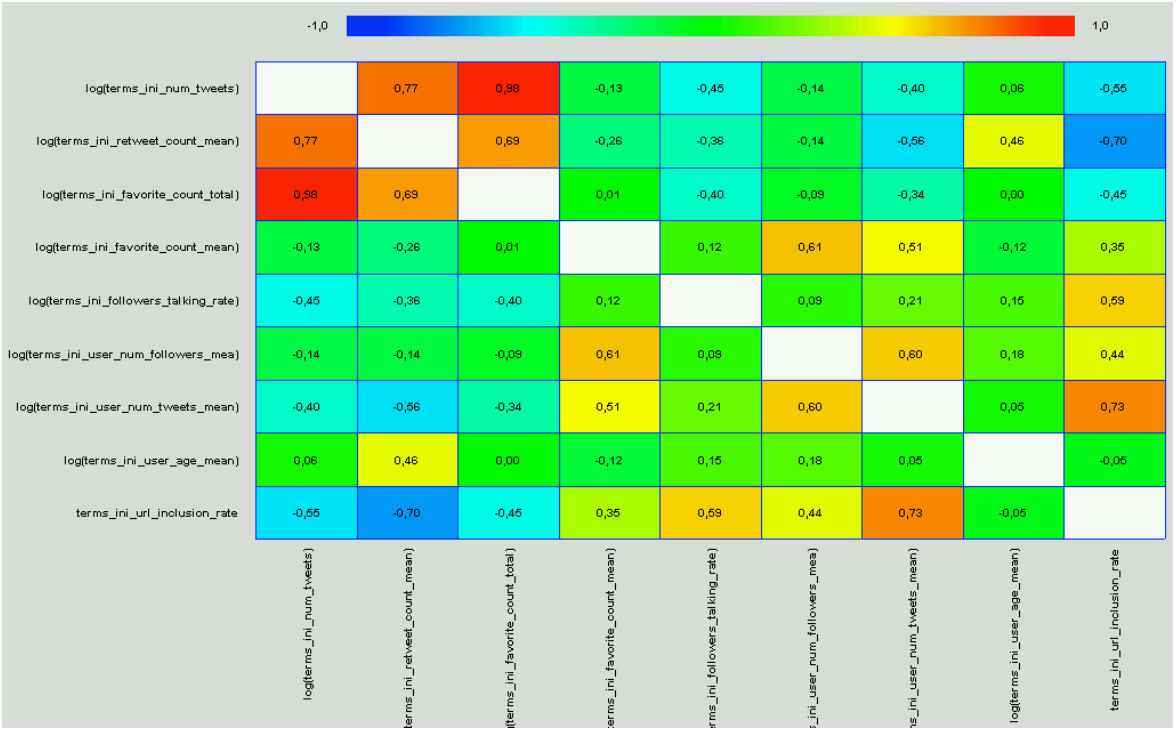


Figura 328. Videojuegos: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- log(terms_ini_num_tweets) y log(terms_ini_retweet_count_mean) tienen un coeficiente de correlación de 0,7669 y un valor-P cercano a 0, por lo que no existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_ini_num_tweets) y log(terms_ini_favorite_count_total) tienen un coeficiente de correlación de 0,9794 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_ini_retweet_count_mean) y log(terms_ini_favorite_count_total) tienen un coeficiente de correlación de 0,6919 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7, redondeando) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_ini_retweet_count_mean) y terms_ini_url_inclusion_rate tienen un coeficiente de correlación de -0,6952 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7, redondeando) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_ini_user_num_tweets_mean) y terms_ini_url_inclusion_rate tienen un coeficiente de correlación de 0,7329 y un valor-P cercano a 0, por lo que no existe

una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto log(terms_ini_retweet_count_mean) porque log(terms_ini_num_tweets y log(terms_ini_favorite_count_total) están fuertemente correlacionadas con esta, y entre ellas también. Además, se elige log(terms_ini_retweet_count_mean) por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que terms_ini_url_inclusion_rate, y log(terms_ini_user_num_tweets_mean) deja de tener esa correlación fuerte con terms_ini_url_inclusion_rate al desaparecer esta del modelo.

La tabla de variables quedaría como sigue:

Tabla 126

Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_followers_talking_rate)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_user_num_tweets_mean)	favorite_count_mean
log(terms_ini_user_age_mean)	log(terms_end_num_tweets)
	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

m) Análisis de componentes principales (ACP)

A continuación, se aplica el análisis de componentes principales (ACP, o PCA en inglés), una técnica que sirve par describir un conjunto de datos según nuevas variables no correlacionadas llamadas componentes (Dunteman, 1989).

El objetivo es representar los datos de la mejor manera posible a través de mínimos cuadrados, construyendo una transformación lineal según un nuevo sistema de coordenadas para los datos originales. Es decir, se plantea representar la variabilidad de los datos con el menor número de componentes o fórmulas posible, las cuales son combinaciones lineales de las variables originales (Dunteman, 1989).

Si las variables originales están muy correlacionadas entre sí, la mayor parte de la variabilidad se podrá expresar en pocas componentes. Si están totalmente incorrelacionadas, el número de componentes será igual al de las variables y este análisis carecerá de interés.

Las componentes se ordenan según la varianza original siendo la primer componente el que tenga la varianza de mayor tamaño. Cuanto mayor sea su varianza, mayor será la información que aporta esa componente (Amat Rodrigo, 2017).

Al construir la matriz de coeficientes de correlación, es posible una base de vectores propios, cuya transformación lineal es necesaria para mejora la simplicidad e interpretación que permita tratar de reducir la dimensionalidad de los datos (Dunteman, 1989). Esta reducción se efectuaría seleccionando las componentes principales que más aportan a la varianza e ignorando el resto. Esta selección se produce ordenando las componentes de mayor a menor aportación a la explicación de la variabilidad, y seleccionando tantas como sean necesarias hasta alcanzar un valor propio mayor o igual a 1.

De esta manera, el método ACP condensa la información de múltiples características en unas pocas, ya que se pretende explicar aproximadamente la información con menos valores que los originales.

Realizando el análisis de componentes principales se ha obtenido un total de tres componentes, explicando así el 82,917% de los datos con un valor propio de 1,08464. Las tres componentes tienen la Tabla 127 de pesos, siendo cada peso un valor de entre -1 y 1.

Tabla 127

Videojuegos: Tabla de pesos de las componentes

	Componente	*Componente*	*Componente*
	1	2	3
log(terms_ini_retweet_count_mean)	-0,421557	0,53654	-0,193897
log(terms_ini_favorite_count_mean)	0,484636	0,123088	-0,31942
log(terms_ini_followers_talking_rate)	0,241147	-0,0405094	0,821516
log(terms_ini_user_num_followers_mea)	0,470304	0,405361	-0,224149
log(terms_ini_user_num_tweets_mean)	0,550511	0,0673117	0,0116511
log(terms_ini_user_age_mean)	-0,0708397	0,725593	0,36757

De esta manera, por ejemplo, la primer componente principal tiene la fórmula siguiente, en donde los valores de las variables se han estandarizado restándoles su promedio y dividiéndolos entre su desviación estándar:

-0,421557 * log(terms_ini_retweet_count_mean) + 0,484636 * log(terms_ini_favorite_count_mean) + 0,241147 * log(terms_ini_followers_talking_rate) + 0,470304 * log(terms_ini_user_num_followers_mean) + 0,550511 * log(terms_ini_user_num_tweets_mean) - 0,0708397 * log(terms_ini_user_age_mean)

Se observa que no hay ninguna variable que aporte positivamente a las tres componentes principales. Todas aportan un valor negativo en alguno de ellos.

La relación entre las variables y las tres componentes principales se puede ver en la siguiente gráfica, ya que las variables se muestran en tres dimensiones formadas por estas componentes:

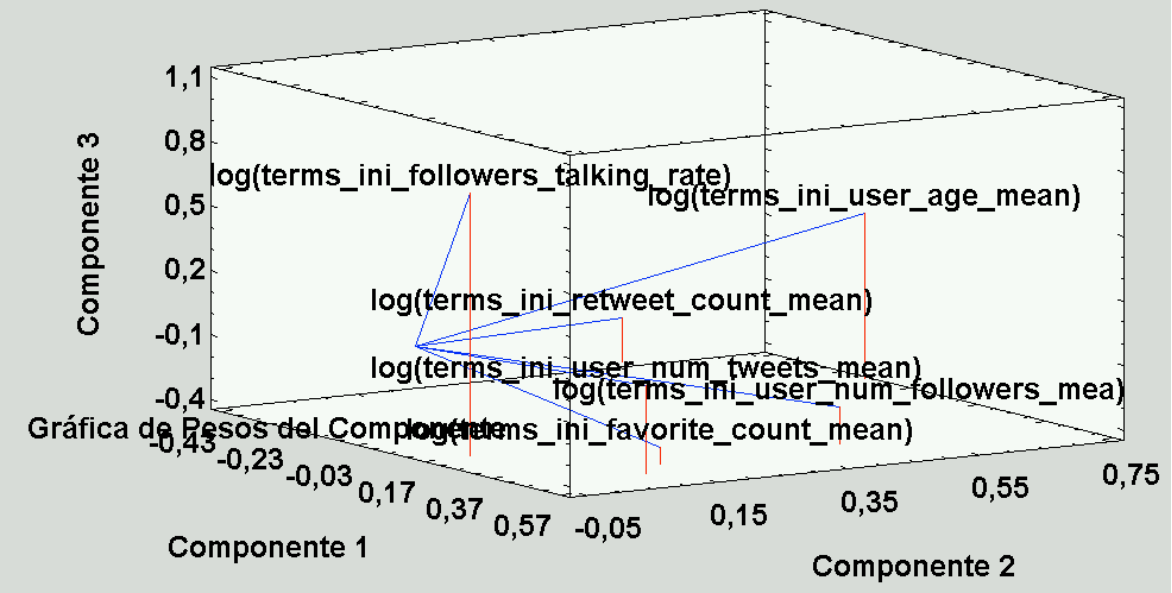


Figura 329. Videojuegos: Gráfica de pesos de cada componente principal

En la Tabla 127 se puede comprobar que todas las variables tienen una presencia significativa en alguna de las componentes principales, por lo que no se puede eliminar ninguna de las variables originales mediante esta técnica. A continuación, se van a analizar todos los artículos de la categoría Videojuegos, de manera que se pueda comprobar si las

características de los datos y las ecuaciones de predicción varían según la sección a la que pertenezca el artículo.

6.1.3.3. Regresión lineal múltiple

Para estudiar la posible relación entre las variables independientes de predicción de que disponemos y cada variable dependiente de éxito o, dicho de otro modo, para tratar de predecir el cálculo de estas, vamos a realizar un modelo de regresión múltiple.

Para realizar las regresiones múltiples, se cuenta con la Tabla *128* de variables resultante de todos los análisis anteriores:

Tabla 128

Videojuegos: Lista final de variables de predicción y de éxito para la regresión lineal múltiple

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_followers_talking_rate)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_user_num_tweets_mean)	favorite_count_mean
log(terms_ini_user_age_mean)	log(terms_end_num_tweets)
	log(terms_end_retweet_count_mean)

Las variables de predicción responden a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores que hablan de la tendencia, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits. Además, incluye también la versión original del número de tuits de la tendencia y la edad en días de los usuarios que participan. Todo lo anterior aplicado a la tendencia el día de la publicación del artículo.

La lista de variables de éxito está formada por la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de

favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después. Además, incluye también la versión original del promedio de retuits en la cuenta del medio y el número de tuits de la tendencia 14 días después.

a) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión múltiple con la variable dependiente log(uniquepageviews_total). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 129

Videojuegos: Valor-P de las variables de la regresión múltiple de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	-4,22632	0,6626
log(terms_ini_retweet_count_mean)	0,391606	0,0055
log(terms_ini_favorite_count_mean)	-1,1255	0,0204
log(terms_ini_followers_talking_rate)	0,411271	0,1977
log(terms_ini_user_num_followers_mean)	-0,216973	0,5286
log(terms_ini_user_num_tweets_mean)	1,63121	0,0715
log(terms_ini_user_age_mean)	-0,784849	0,5466
Modelo		0,0065

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 130

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	2,78799	0
log(terms_ini_retweet_count_mean)	0,278855	0,0001
log(terms_ini_favorite_count_mean)	-0,645254	0,0184
Modelo		0,0001

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(2,78799 + 0,278855 * log(terms_ini_retweet_count_mean) - 0,645254 * log(terms_ini_favorite_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

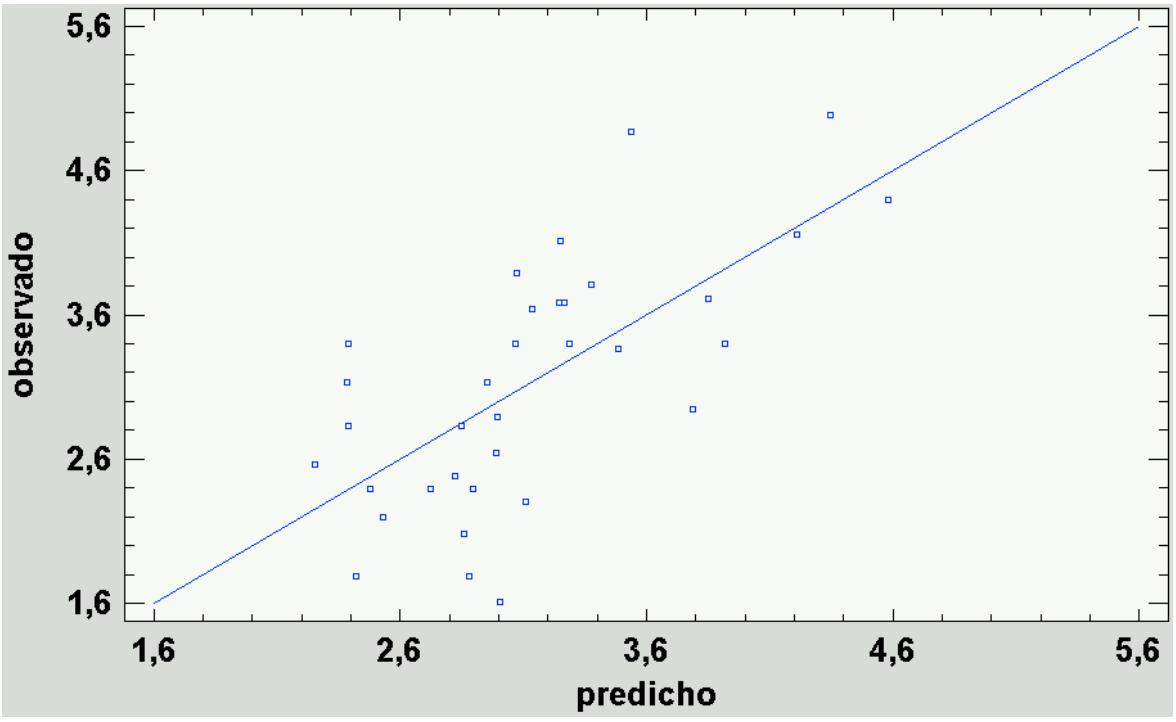


Figura 330. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 46,0113% de la variabilidad de log(uniquepageviews_total), mientras que el R-Cuadrado ajustado indica un 42,5282%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen la línea.

b) AdSense eCPM (promedio)

Para tratar de predecir el valor de estimación de ingresos de los anuncios por cada 1000 páginas vistas desde los anuncios de Google AdSense es necesario realizar la regresión múltiple con la variable dependiente log(adsense_ecpm_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 131

Videojuegos: Valor-P de las variables de la regresión múltiple de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P
Constante	-19,437	0,0953
log(terms_ini_retweet_count_mean)	0,164876	0,2644
log(terms_ini_favorite_count_mean)	-0,455882	0,3935
log(terms_ini_followers_talking_rate)	0,263255	0,4687
log(terms_ini_user_num_followers_mean)	-0,60388	0,1365
log(terms_ini_user_num_tweets_mean)	1,93928	0,0638
log(terms_ini_user_age_mean)	0,288151	0,8463
Modelo		0,3203

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 132

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P
Constante	-2,65561	0
log(terms_ini_favorite_count_mean)	-0,704478	0,0601
Modelo		0,0601

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable log(adsense_ecpm_mean) no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

c) Duración de la visita (promedio)

Para tratar de predecir el valor de la duración de la visita (promedio) es necesario realizar la regresión múltiple con la variable dependiente log(avgtimeonpage_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 133

Videojuegos: Valor-P de las variables de la regresión múltiple de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	-7,286	0,6857
log(terms_ini_retweet_count_mean)	-0,203281	0,3898
log(terms_ini_favorite_count_mean)	-0,780055	0,3942
log(terms_ini_followers_talking_rate)	0,706126	0,2372
log(terms_ini_user_num_followers_mean)	0,442792	0,5119
log(terms_ini_user_num_tweets_mean)	-1,20878	0,4758

log(terms_ini_user_age_mean)	3,17175	0,2114
Modelo		0,289

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 134

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	8,08745	0,0007
log(terms_ini_followers_talking_rate)	0,959294	0,0553
Modelo		0,0553

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable log(avgtimeonpage_mean) no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

d) Páginas vistas por sesión (promedio)

Para tratar de predecir el valor de las páginas vistas por sesión (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(pageviewspersession_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 135

Videojuegos: Valor-P de las variables de la regresión múltiple de log(pageviewspersession_mean)

Variable	Estimación	Valor-P

Constante	-5,10614	0,277
log(terms_ini_retweet_count_mean)	-0,042543	0,4834
log(terms_ini_favorite_count_mean)	0,278979	0,2041
log(terms_ini_followers_talking_rate)	-0,20056	0,1894
log(terms_ini_user_num_followers_mean)	-0,235695	0,1616
log(terms_ini_user_num_tweets_mean)	0,0286456	0,9447
log(terms_ini_user_age_mean)	0,801942	0,2057
Modelo		0,6162

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 136

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	1,02235	0,1236
log(terms_ini_user_num_followers_mean)	-0,0879484	0,2219
Modelo		0,2219

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del

95%. Por tanto, la variable log(pageviewspersession_mean) no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

e) Nº de favoritos en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de favoritos en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente favorite_count_mean. Se utilizan como variables independientes las disponibles para la predicción.

Tabla 137

Videojuegos: Valor-P de las variables de la regresión múltiple de favorite_count_mean

Variable	Estimación	Valor-P
Constante	-2,31932	-0,305362
log(terms_ini_retweet_count_mean)	0,149358	0,1493
log(terms_ini_favorite_count_mean)	-0,5827	0,1161
log(terms_ini_followers_talking_rate)	0,0587544	0,8133
log(terms_ini_user_num_followers_mean)	0,00121173	0,9965
log(terms_ini_user_num_tweets_mean)	0,968943	0,1708
log(terms_ini_user_age_mean)	-0,775653	0,4562
Modelo		0,4893

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 138

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de favorite_count_mean

Variable	Estimación	Valor-P
Constante	1,20046	0
log(terms_ini_favorite_count_mean)	0,0564979	0,2608
Modelo		0,2608

El valor-P en la tabla ANOVA del modelo es mayor o igual que 0,05, por lo que no existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%. Por tanto, la variable favorite_count_mean no puede ser predicha mediante regresión lineal múltiple con las variables de que se dispone en el modelo.

f) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor de el número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_num_tweets). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 139

Videojuegos: Valor-P de las variables de la regresión múltiple de log(terms_end_num_tweets)

Variable	Estimación	Valor-P
Constante	21,8923	0,2094
log(terms_ini_retweet_count_mean)	1,11775	0,0001
log(terms_ini_favorite_count_mean)	-0,87363	0,3164
log(terms_ini_followers_talking_rate)	0,124397	0,8227
log(terms_ini_user_num_followers_mean)	0,608231	0,3454
log(terms_ini_user_num_tweets_mean)	3,61303	0,3454
log(terms_ini_user_age_mean)	-7,83682	0,0345

Modelo	0,0011

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 140

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_num_tweets)

Variable	Estimación	Valor-P
Constante	-20,4083	
log(terms_ini_retweet_count_mean)	0,859457	0
log(terms_ini_user_num_tweets_mean)	2,34719	0,0088
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(-20,4083 + 0,859457 * log(terms_ini_retweet_count_mean) + 2,34719 * log(terms_ini_user_num_tweets_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

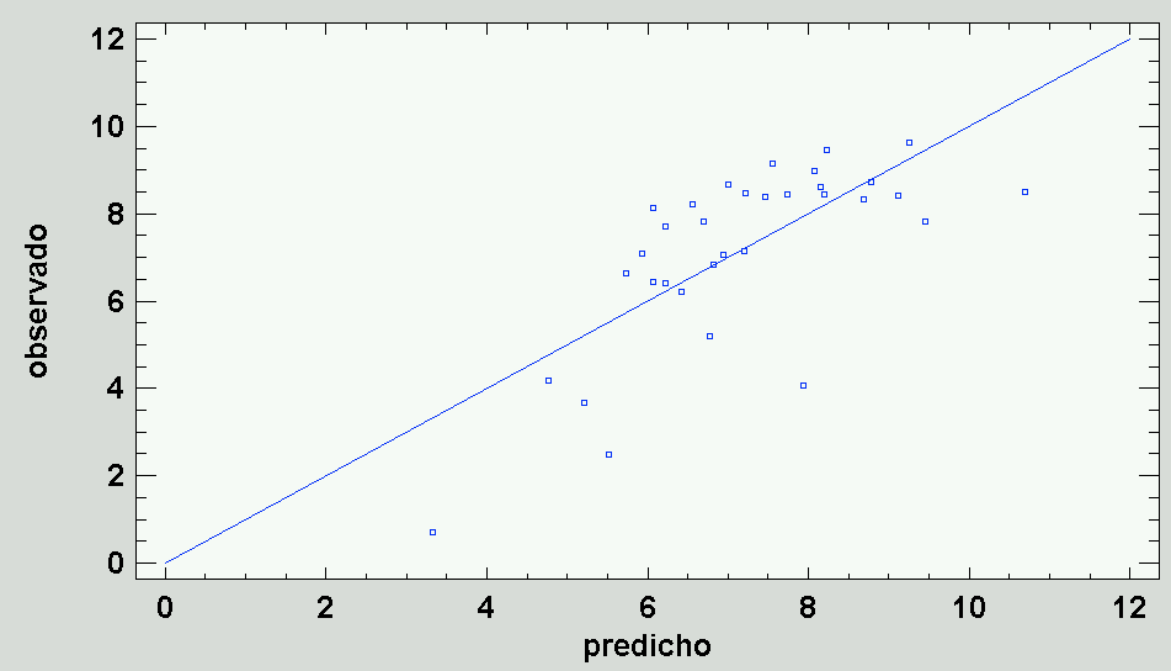


Figura 331. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_num_tweets) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 51,3521% de la variabilidad de log(terms_end_num_tweets), mientras que el R-Cuadrado ajustado indica un 48,109%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos se aproximan a la línea.

g) Nº de retuits de la tendencia 14 días después (promedio)

Para tratar de predecir el valor de el número de retuits de la tendencia 14 días después (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción.

Tabla 141

Videojuegos: Valor-P de las variables de la regresión múltiple de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	0,0497355	0,9981
log(terms_ini_retweet_count_mean)	0,969258	0,0019

log(terms_ini_favorite_count_mean)	0,0320186	0,9754
log(terms_ini_followers_talking_rate)	0,0405686	0,9517
log(terms_ini_user_num_followers_mean)	0,866634	0,2679
log(terms_ini_user_num_tweets_mean)	2,80553	0,1575
log(terms_ini_user_age_mean)	-4,80001	0,1041
Modelo		0,0102

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 142

Videojuegos: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	1,04387	0,144
log(terms_ini_retweet_count_mean)	0,680446	0,0002
Modelo		0,0002

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_mean = exp(1,04387 + 0,680446 * log(terms_ini_retweet_count_mean))

Para realizar el cálculo de terms_end_retweet_count_mean será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

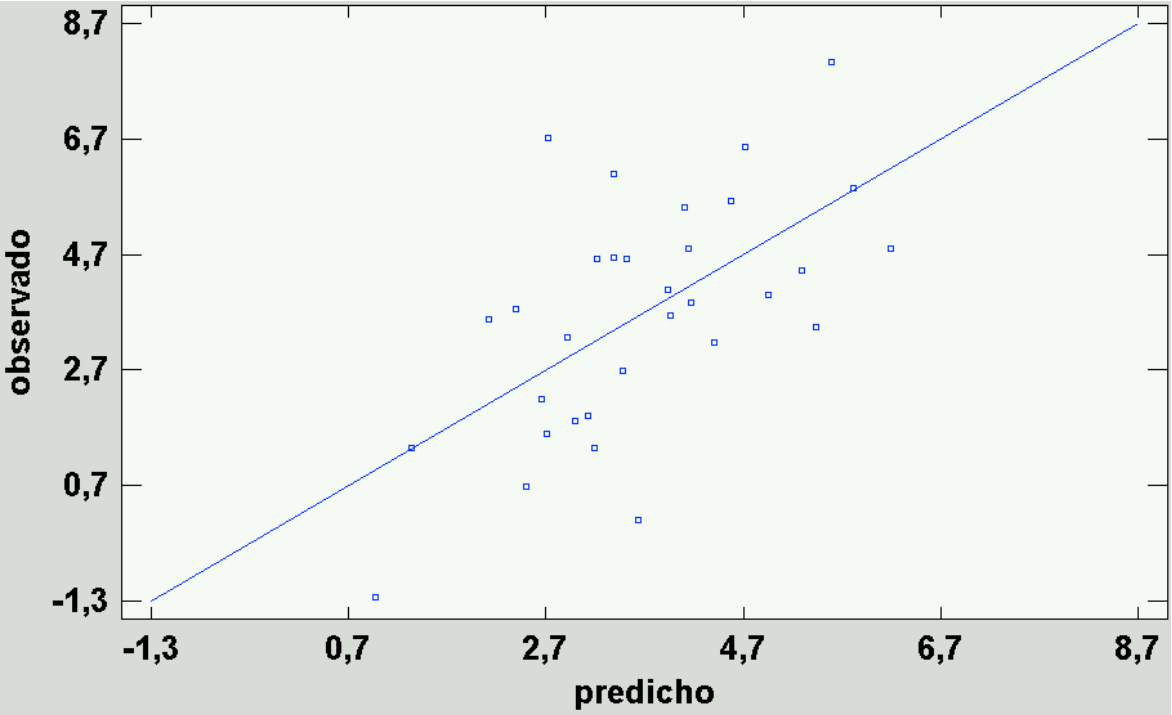


Figura 332. Videojuegos: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 36,7106% de la variabilidad de log(terms_end_retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 34,601%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos se aproximan a la línea.

h) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones lineales múltiples de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 143

Videojuegos: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones lineales múltiples

Variables de predicción	Variables de éxito
log(terms_ini_retweet_count_mean)	log(uniquepageviews_total)
log(terms_ini_favorite_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_followers_talking_rate)	log(avgtimeonpage_mean)
log(terms_ini_user_num_followers_mean)	log(pageviewspersession_mean)
log(terms_ini_user_num_tweets_mean)	favorite_count_mean
log(terms_ini_user_age_mean)	log(terms_end_num_tweets)
	log(terms_end_retweet_count_mean)

Se puede observar en la Tabla *143* que solo log(terms_ini_retweet_count_mean) y log(terms_ini_favorite_count_mean) participan en alguna de las ecuaciones de predicción, por lo que solamente estas son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- El promedio de retuits y el promedio de favoritos explica parte de los datos de páginas vistas únicas.
- El promedio de retuits explica parte de los datos del número de tuits 14 días después.
- El promedio de retuits explica parte de los datos del promedio de retuits 14 días después.
- El promedio de retuits, por tanto, participa en la predicción de todas las variables de éxito que se pueden predecir.
- El promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas por sesión y el promedio de favoritos en la cuenta del medio no se han podido predecir con las variables de predicción existentes en el modelo.

6.1.3.4. Regresión binomial negativa o de Poisson

A continuación, se tratará de predecir todas las variables de éxito que sean de conteo (enteros y sin números negativos) a partir de todas las variables de predicción que sean independientes entre sí según la regresión binomial negativa o la regresión de Poisson.

a) Filtro de alta correlación (colinealidad)

Las variables que aporten información para tratar de realizar la regresión deben ser independientes, motivo por el cual es necesario hacer un filtro de alta correlación de manera que se asegure que todas aportan información diferente.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

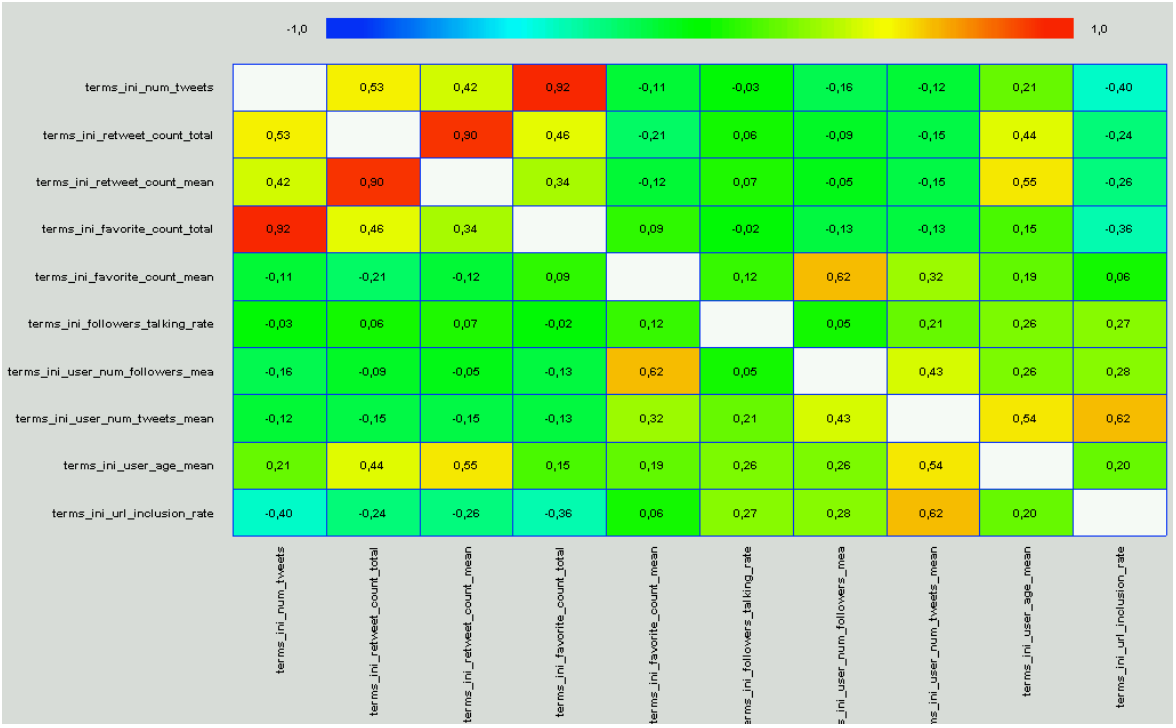


Figura 333. Videojuegos: Matriz de correlaciones Pearson entre las variables de predicción

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_ini_num_tweets y terms_ini_favorite_count_total tienen un coeficiente de correlación de 0,9185 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

- tems_ini_retweet_count_total y terms_ini_retweet_count_mean tienen un coeficiente de correlación de 0,8975 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige terms_ini_favorite_count_total y terms_ini_retweet_count_mean por tener un sesgo y una curtosis estandarizados menores, como se puede comprobar en el anexo 6.1.3.2.

La tabla de variables quedaría como sigue:

Tabla 144

Videojuegos: Lista de variables de predicción y de éxito para la regresión binomial negativa o de Poisson tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_retweet_count_mean	uniquepageviews_total
terms_ini_favorite_count_total	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a: el número total de tuits, el número total de retuits, el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de edad en días de la cuenta de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

b) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión con la variable dependiente uniquepageviews_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable uniquepageviews_total, el Chi-cuadrado calculado es 34.673,9 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 145

Videojuegos: Valor-P de las variables de la regresión binomial negativa de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,44534	0
terms_ini_retweet_count_mean	0,000752962	0,082
terms_ini_favorite_count_total	0,0000153543	0
terms_ini_favorite_count_mean	-0,166932	0
terms_ini_followers_talking_rate	16,2211	0,0001
terms_ini_user_num_followers_mean	-0,000018654	0,1508
terms_ini_user_num_tweets_mean	0,0000416159	0,0013
terms_ini_user_age_mean	-0,0000940685	1
terms_ini_url_inclusion_rate	-2,42176	0,0001
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que

0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 146

Videojuegos: Valor-P de las variables de la regresión binomial negativa simplificada de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,44036	0
terms_ini_retweet_count_mean	0,000626433	0
terms_ini_favorite_count_total	0,0000162856	0
terms_ini_favorite_count_mean	-0,21469	0
terms_ini_followers_talking_rate	16,3856	0,0051
terms_ini_user_num_tweets_mean	0,0000359194	0,0002
terms_ini_url_inclusion_rate	-2,53259	0,0001
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(3,44036 + 0,000626433 * terms_ini_retweet_count_mean + 0,0000162856 * terms_ini_favorite_count_total - 0,21469 * terms_ini_favorite_count_mean + 16,3856 * terms_ini_followers_talking_rate + 0,0000359194 * terms_ini_user_num_tweets_mean - 2,53259 * terms_ini_url_inclusion_rate)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

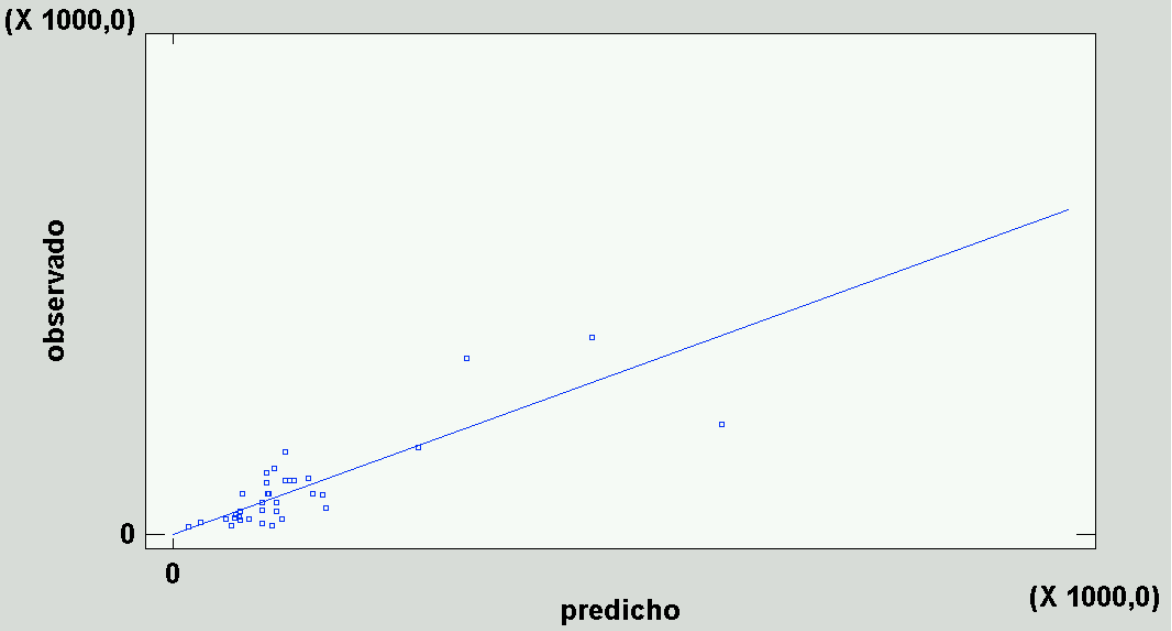


Figura 334. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 60,1601% de la variabilidad de uniquepageviews_total, mientras que el R-Cuadrado ajustado indica un 55,3071%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se acercan de la línea que muestra una predicción acertada.

c) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_num_tweets, el Chi-cuadrado calculado es 454.197.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 147

Videojuegos: Valor-P de las variables de la regresión binomial negativa de terms_end_num_tweets

Variable	Estimación	Valor-P

Constante	17,3859	0
terms_ini_retweet_count_mean	-0,000745232	0,3432
terms_ini_favorite_count_total	0,0000719001	0,0154
terms_ini_favorite_count_mean	-0,24769	1
terms_ini_followers_talking_rate	0,796487	0,9404
terms_ini_user_num_followers_mean	0,0000172302	0,9395
terms_ini_user_num_tweets_mean	0,0000376117	0,433
terms_ini_user_age_mean	0,000918417	0,3061
terms_ini_url_inclusion_rate	-3,42351	0,2392
Modelo		0,3439

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 148

Videojuegos: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_num_tweets

Variable	Estimación	Valor-P
Constante	18,5096	0
terms_ini_favorite_count_total	0,0000783923	0,0062
Modelo		0,0062

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(18,5096 + 0,0000783923 * terms_ini_favorite_count_total)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

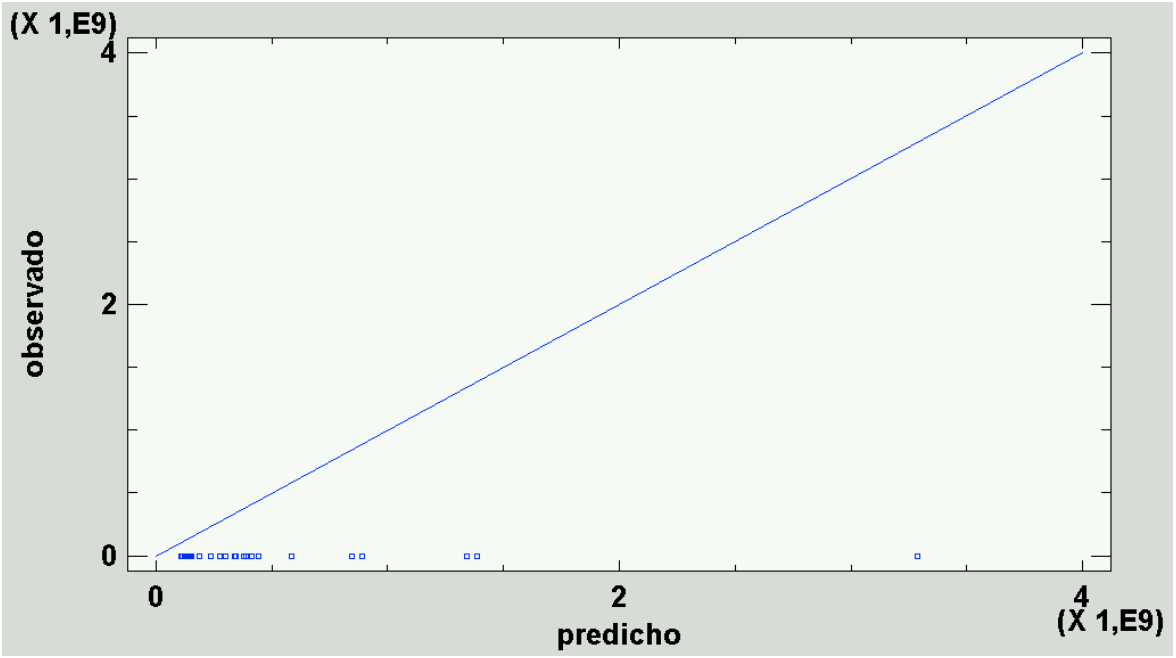


Figura 335. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 10,4097% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 4,84411%. Este número se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan a la línea que muestra una predicción acertada.

d) Nº de retuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de retuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_retweet_count_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_retweet_count_total, el Chi-cuadrado calculado es 526.058.000.000.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Se realiza la regresión binomial negativa sin la variable terms_ini_favorite_count_total porque en este análisis en concreto es una combinación lineal de otras variables y no permitía su procesamiento:

Tabla 149

Videojuegos: Valor-P de las variables de la regresión binomial negativa de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	34,0395	0
terms_ini_retweet_count_mean	-0,00128272	0,7123
terms_ini_favorite_count_mean	-0,332544	0,5336
terms_ini_followers_talking_rate	0,000112511	0,4842
terms_ini_user_num_followers_mean	0,000112511	1
terms_ini_user_num_tweets_mean	-0,000173384	0,3895
terms_ini_user_age_mean	0,00239063	1
terms_ini_url_inclusion_rate	5,09387	0
Modelo		0,9617

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 150

Videojuegos: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	23,4201	0
terms_ini_user_age_mean	0,00532893	0,0358
Modelo		0,0358

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_total = exp(23,4201 + 0,00532893 * terms_ini_user_age_mean)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

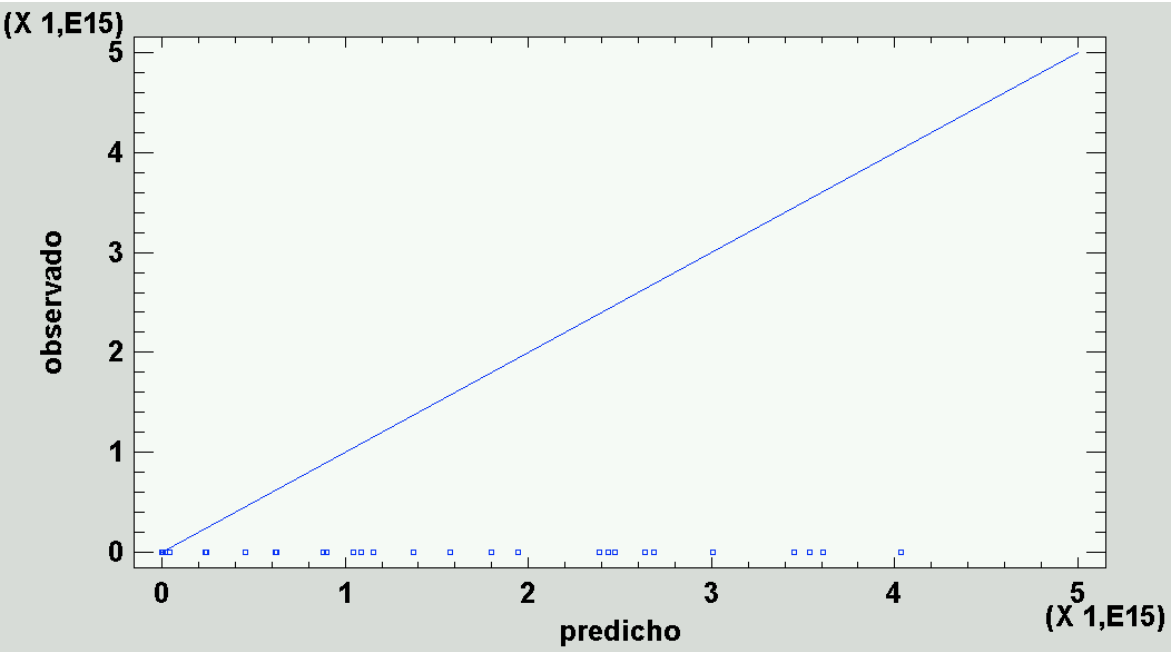


Figura 336. Videojuegos: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 6,38892% de la variabilidad de terms_end_retweet_count_total, mientras que el R-Cuadrado ajustado indica un 0,590067%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

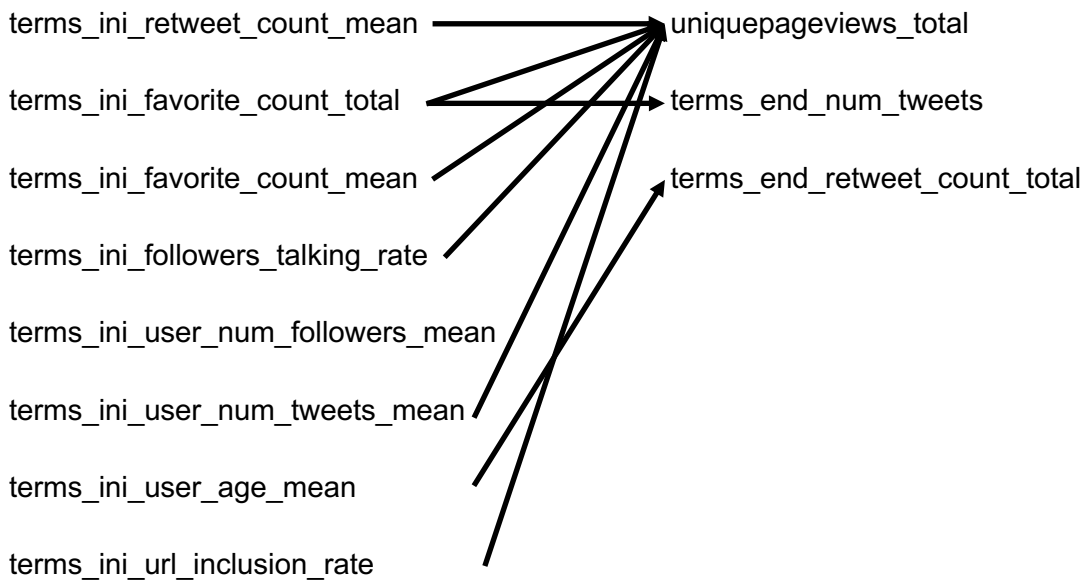
e) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones binomiales negativas de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 151

Videojuegos: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones binomiales negativas

Variables de predicción	Variables de éxito



Se puede observar en la Tabla *151* que todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que todas son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- El promedio de retuits, el número de favoritos, el promedio de favoritos, la ratio de seguidores de la cuenta del medio que participan, el número de tuits de los usuarios que participan, el promedio de la edad en días de los usuarios que participan en la tendencia y la ratio de inclusión de URL en los tuits explican parte de los datos de páginas vistas únicas.
- El número de favoritos explica parte de los datos del número de tuits 14 días después.

  El promedio de la edad en días de los usuarios que participan explica en parte el número de retuits 14 días después.

6.1.4. Análisis de los artículos de la categoría Tráileres

6.1.4.1. Variables de éxito

El objetivo de esta fase es la predicción de los valores de éxito mediante el resto de los indicadores. Por ello, es importante comenzar por analizar estos escenarios por separado para ver tanto sus características como tratar de describir su comportamiento anómalo, si lo tuvieran.

a) Páginas vistas únicas (total)

Este escenario de éxito se ve identificado por la columna uniquepageviews_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 101 valores con un rango de entre 3 y 704.

Presenta un sesgo estandarizado de 37,317 y una curtosis estandarizada de 180,319. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

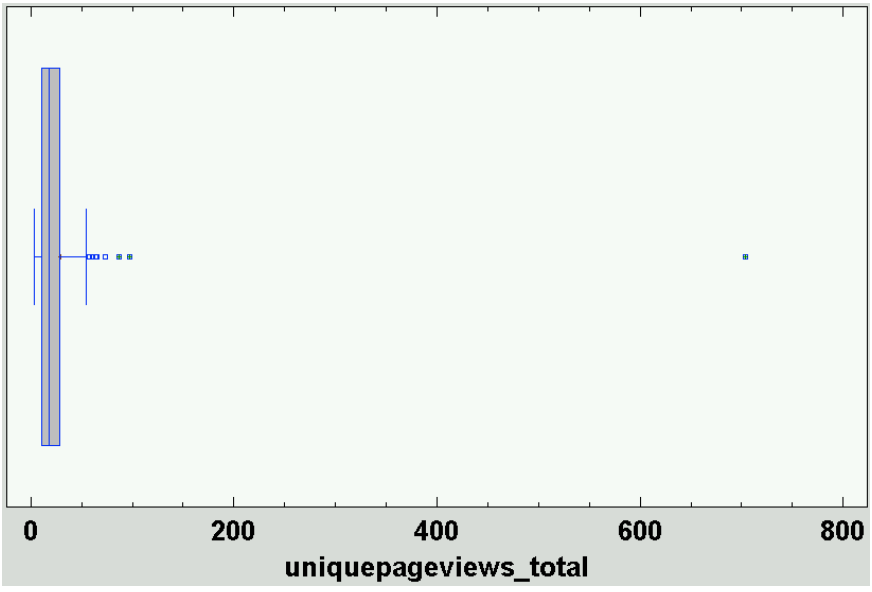


Figura 337. Tráileres: Gráfico de Caja y Bigotes para el valor uniquepageviews_total

En la Figura 1 se puede comprobar que existen valores anómalos de tipo extremo de 87 páginas vistas o más.

b) AdSense eCPM (promedio)

Este escenario de éxito se identifica con la columna adsense_ecpm_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 0,38.

Presenta un sesgo estandarizado de 13,3323 y una curtosis estandarizada de 26,6034. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

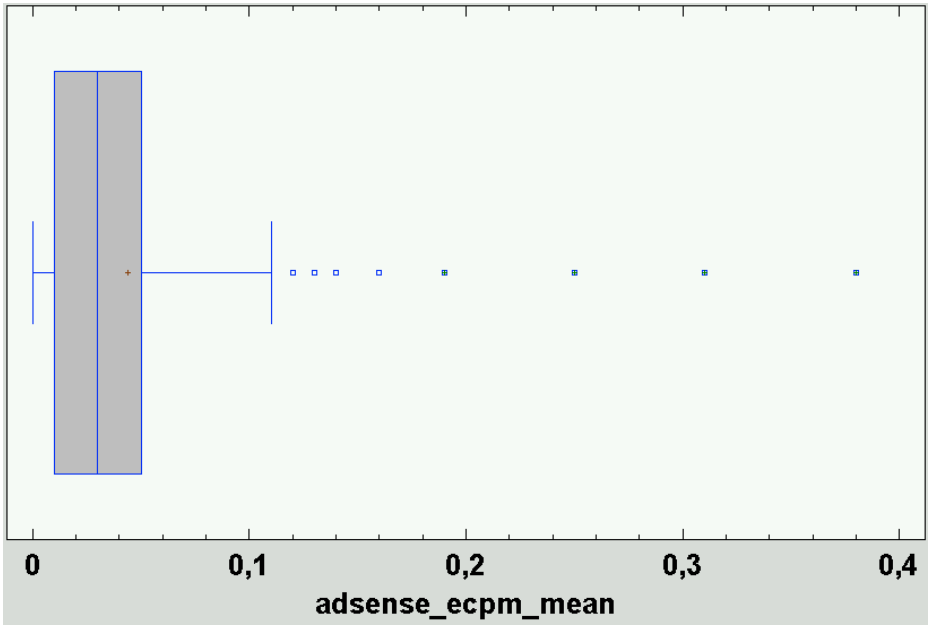


Figura 338. Tráileres: Gráfico de Caja y Bigotes para el valor adsense_ecpm_mean

En la Figura 338 se puede observar que existen valores anómalos de tipo extremo de 0,19 o más.

c) Duración de la visita (promedio)

Este escenario de éxito se identifica con la columna avgtimeonpage_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 11 y 884.

Presenta un sesgo estandarizado de 9,08036 y una curtosis estandarizada de 14,771. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

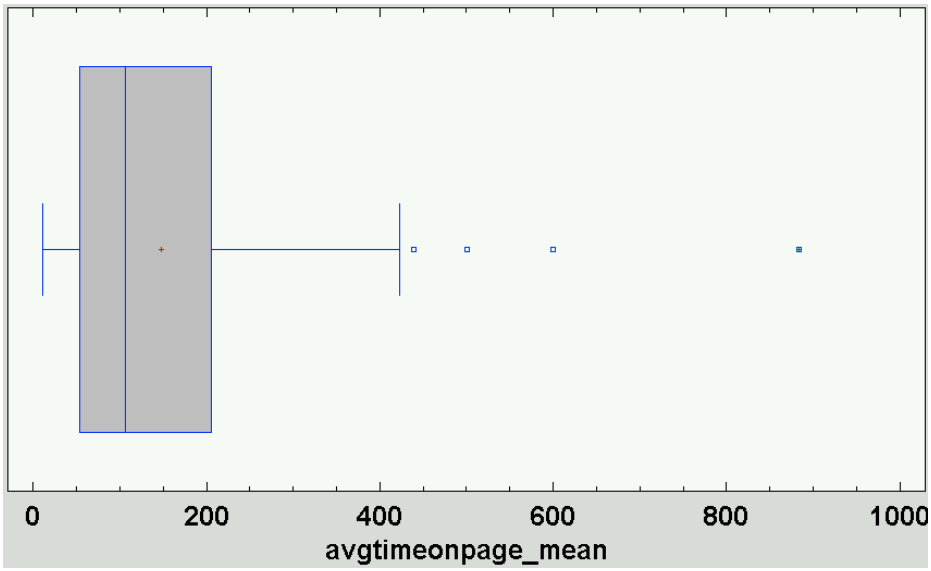


Figura 339. Tráileres: Gráfico de Caja y Bigotes para el valor avgtimeonpage_mean

En la Figura 339 se puede observar que existe un valor anómalo de tipo extremo de 884.

d) Páginas vistas por sesión (promedio)

Este escenario de éxito se identifica con la columna pageviewspersession_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0,5 y 9.

Presenta un sesgo estandarizado de 19,3283 y una curtosis estandarizada de 68,7835 Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

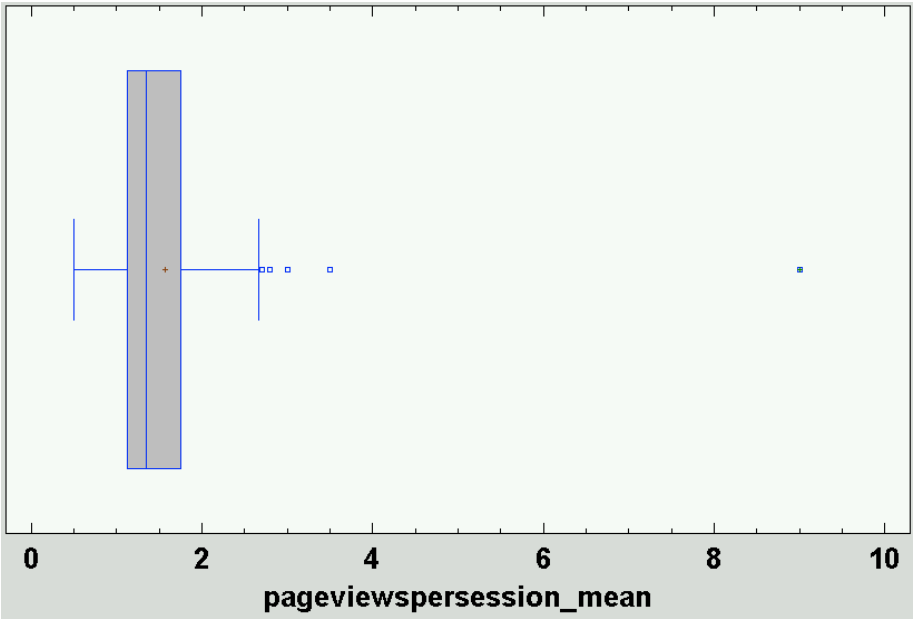


Figura 340. Tráileres: Gráfico de Caja y Bigotes para el valor pageviewspersession_mean

En la Figura 340 se puede observar que existe un valor anómalo de tipo extremo de 9.

e) Nº de retuits en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 4.

Presenta un sesgo estandarizado de 6,32462 y una curtosis estandarizada de 17,2261. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

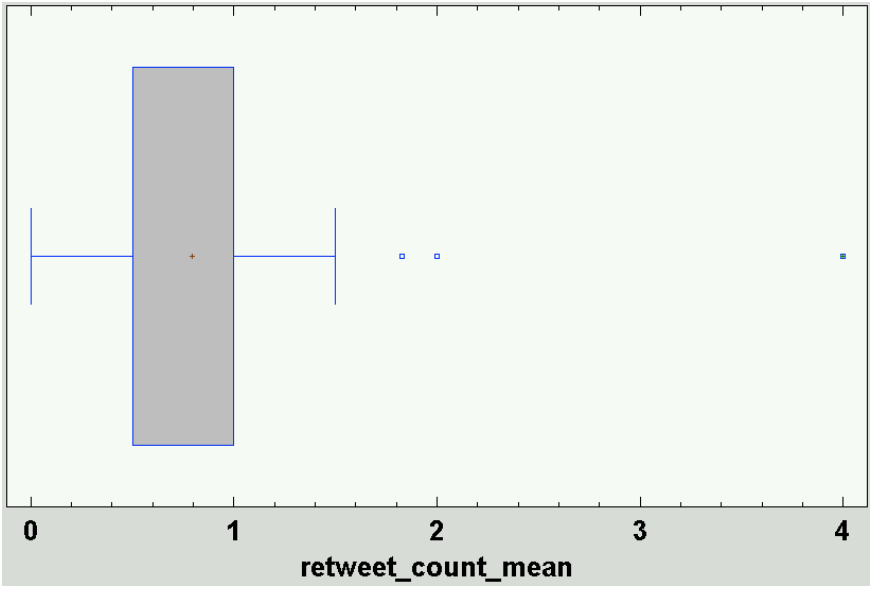


Figura 341. Tráileres: Gráfico de Caja y Bigotes para el valor retweet_count_mean

En la Figura 341 solo se observa un valor anómalo de tipo extremo con 4 retuits de promedio.

f) Nº de favoritos en la cuenta del medio (promedio)

Este escenario de éxito se identifica con la columna favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 11.

Presenta un sesgo estandarizado de 18,1545 y una curtosis estandarizada de 63,3226. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

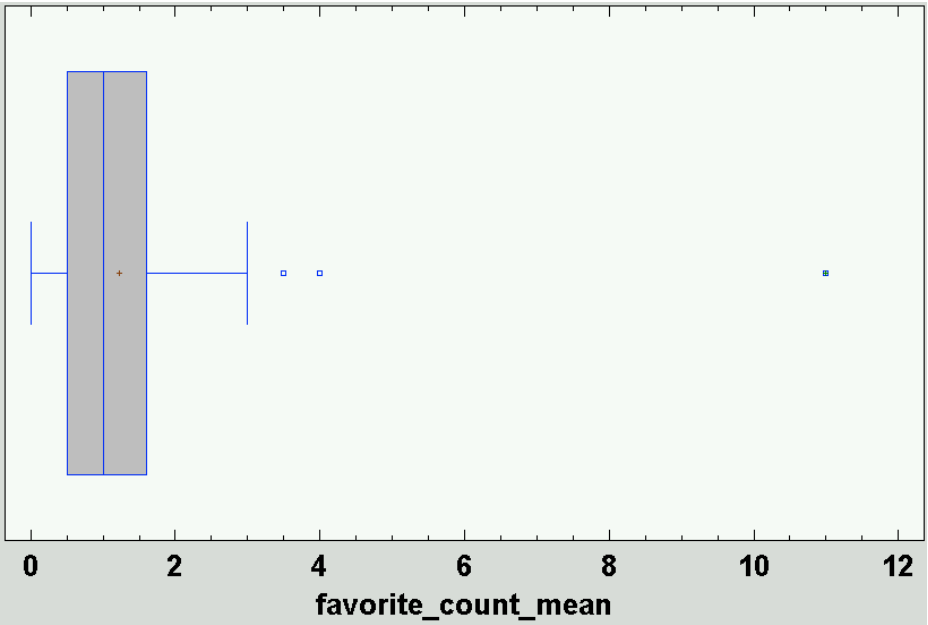


Figura 342. Tráileres: Gráfico de Caja y Bigotes para el valor favorite_count_mean

En la Figura 342 solo se observa un valor anómalo de tipo extremo con 11 favoritos de promedio.

g) Nº de tuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 101 valores con un rango de entre 0 y 8.470.

Presenta un sesgo estandarizado de 7,17025 y una curtosis estandarizada de 6,77794. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

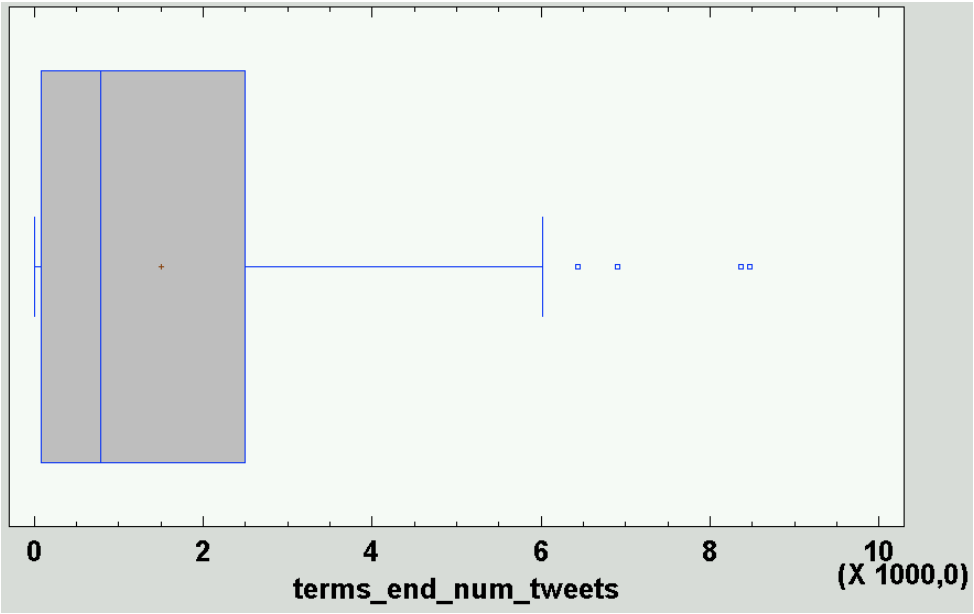


Figura 343. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_num_tweets

En la Figura 343 se puede observar que no existen valores anómalos de tipo extremo.

h) Nº de retuits de la tendencia 14 días después (total)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 697.550.

Presenta un sesgo estandarizado de 21,7633 y una curtosis estandarizada de 62,5373. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

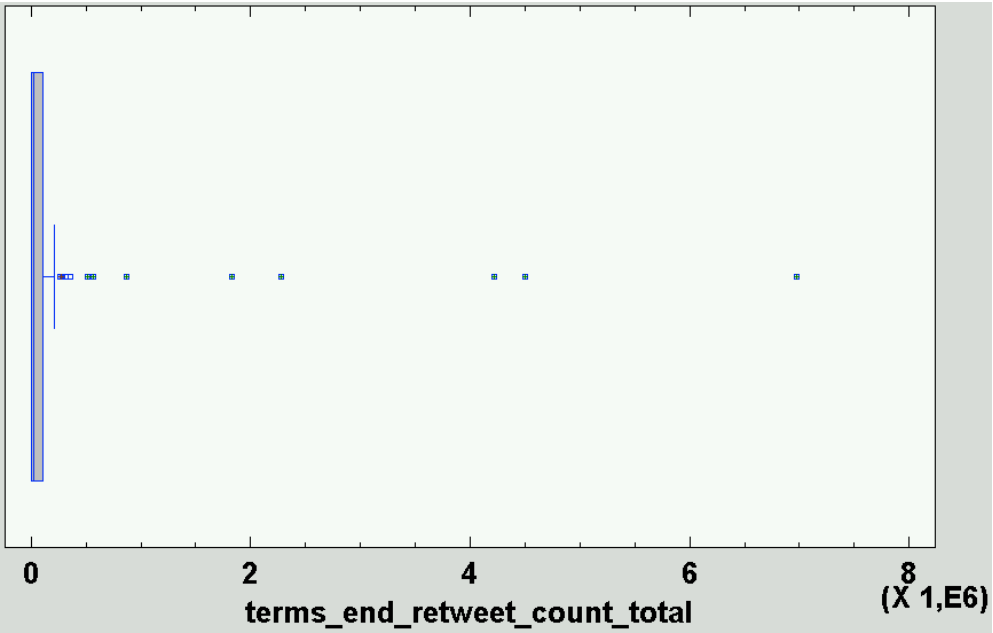


Figura 344. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_total

En la Figura 344 se puede observar que existen valores anómalos de tipo extremo de 511.583 o más.

i) Nº de retuits de la tendencia 14 días después (promedio)

Este escenario de éxito se identifica con la columna terms_end_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 935,42.

Presenta un sesgo estandarizado de 18,5619 y una curtosis estandarizada de 43,7875. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

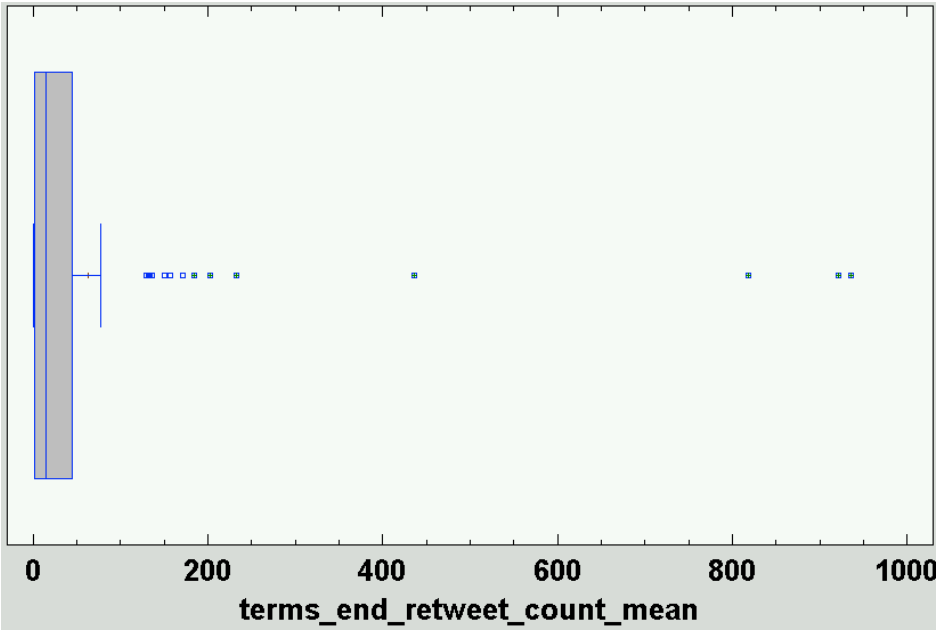


Figura 345. Tráileres: Gráfico de Caja y Bigotes para el valor terms_end_retweet_count_mean

En la Figura 345 se puede observar que existen valores anómalos de tipo extremo de 184,2 o más.

j) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de las variables de éxito. Se pueden observar los siguientes datos de estas:

Tabla 152

Tráileres: Resumen estadístico de las variables de éxito

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
uniquepageviews_total	4.911,84	37,317	180,319
adsense_ecpm_mean	0,00368416	13,3323	26,6034
avgtimeonpage_mean	20.021,9	9,08036	14,771

pageviewspersession_mean	0,948624	19,3283	68,7835
retweet_count_mean	0,333256	6,32462	17,2261
favorite_count_mean	1,71023	18,1545	63,3226
terms_end_num_tweets	3.437.420	7,17025	6,77794
terms_end_retweet_count_total	907.916.000.000	21,7633	62,5373
terms_end_retweet_count_mean	25.275	18,5619	43,7875

Se puede observar en la Tabla 152 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 153

Tráileres: Resumen estadístico de las variables de éxito con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
log(uniquepageviews_total)	0,666503	3,1208	6,29932
log(adsense_ecpm_mean)	0,825455	1,85434	-0,374182
log(avgtimeonpage_mean)	0,919456	-0,798609	-1,14452

log(pageviewspersession_mean)	0,204636	2,17984	4,50298
log(retweet_count_mean)	0,158982	0,0930296	4,01816
log(favorite_count_mean)	0,290905	2,68403	4,41415
log(terms_end_num_tweets)	4,7419	-3,31223	-0,628265
log(terms_end_retweet_count_total)	14,5329	-2,22213	-1,23078
log(terms_end_retweet_count_mean)	4,18972	-1,24858	-0,491845

Todas las variables salvo log(avgtimeonpage_mean) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. Sin embargo, mantienen valores mucho menores, próximos al rango de -2 a +2 y con una dispersión muy parecida, por lo que en este estudio se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

Nos quedamos, por tanto, con las variables que tengan menores sesgo y curtosis estandarizados de su forma original o transformación logarítmica.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 154

Tráileres: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(uniquepageviews_total)	0,737315

En la Tabla 154 se puede ver que la variable tiene el valor-p necesario (mayor o igual que 0,05) para confirmar que sigue una distribución normal, por lo que se añade al modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 155

Tráileres: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	log(retweet_count_mean)
terms_ini_followers_talking_rate	log(favorite_count_mean)
terms_ini_user_num_followers_mean	log(terms_end_num_tweets)
terms_ini_user_num_tweets_mean	log(terms_end_retweet_count_total)
terms_ini_user_age_mean	log(terms_end_retweet_count_mean)
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, por tanto, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio, el número total de retuits de la tendencia 14 días después y el promedio de retuits de la tendencia 14 días después.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

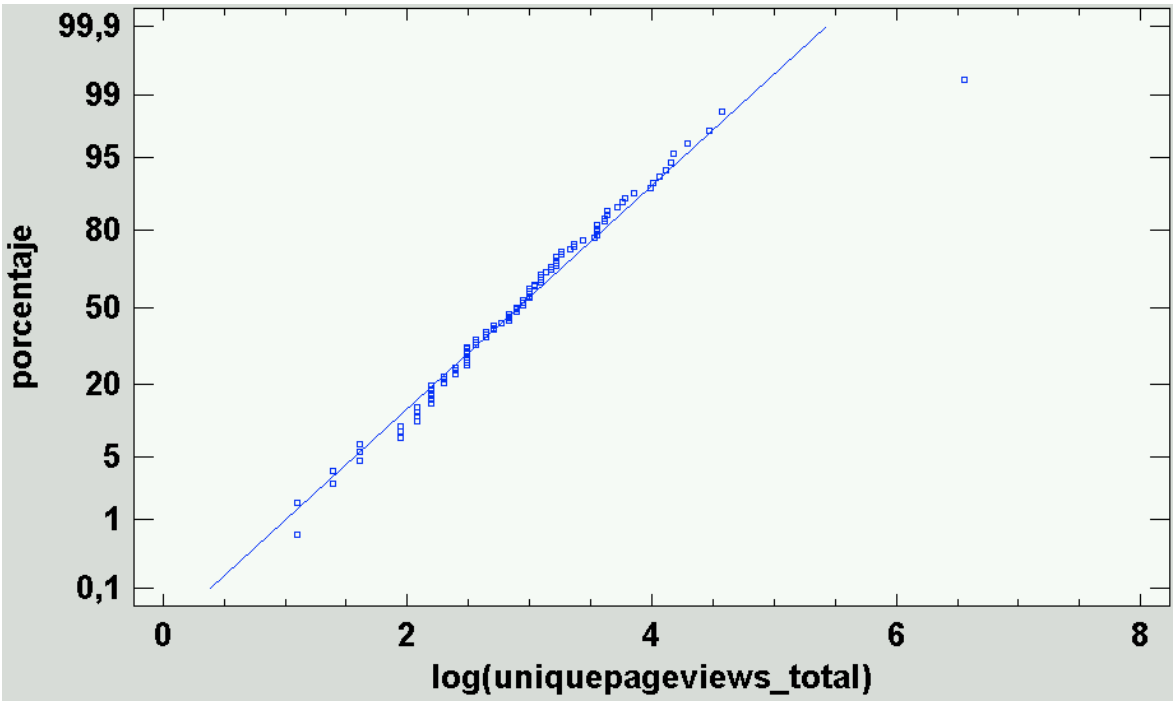


Figura 346. Tráileres: Gráfico de probabilidad normal de la variable log(uniquepageviews_total)

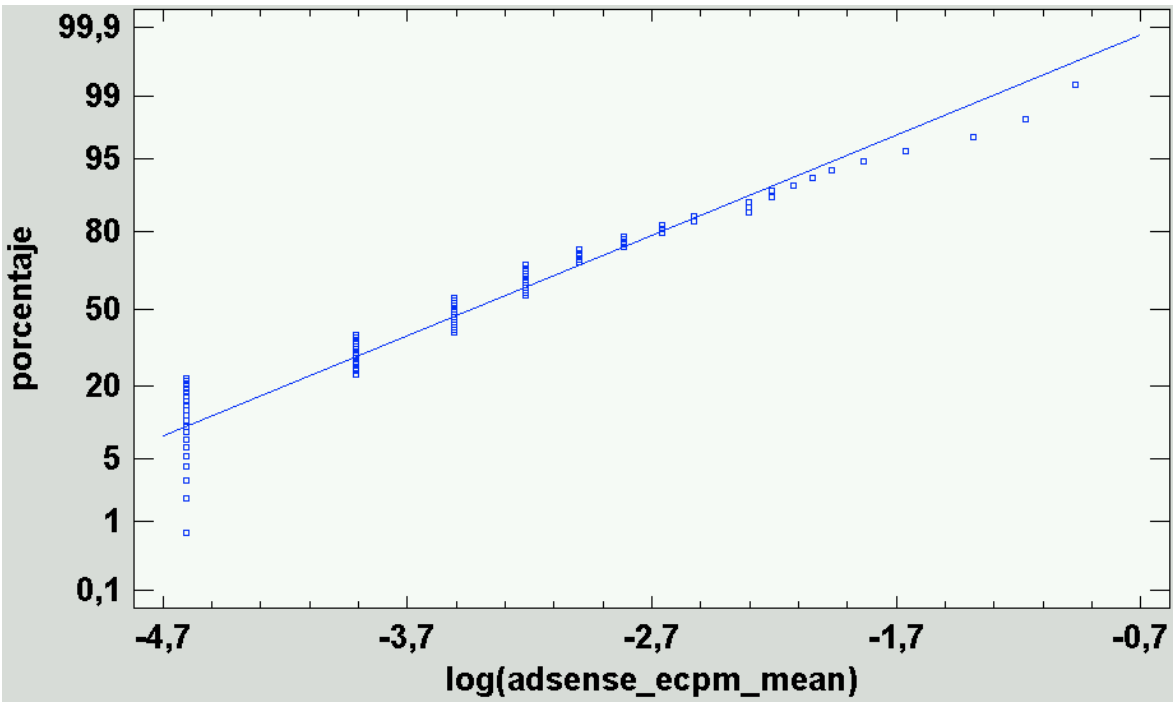


Figura 347. Tráileres: Gráfico de probabilidad normal de la variable log(adsense_ecpm_mean)

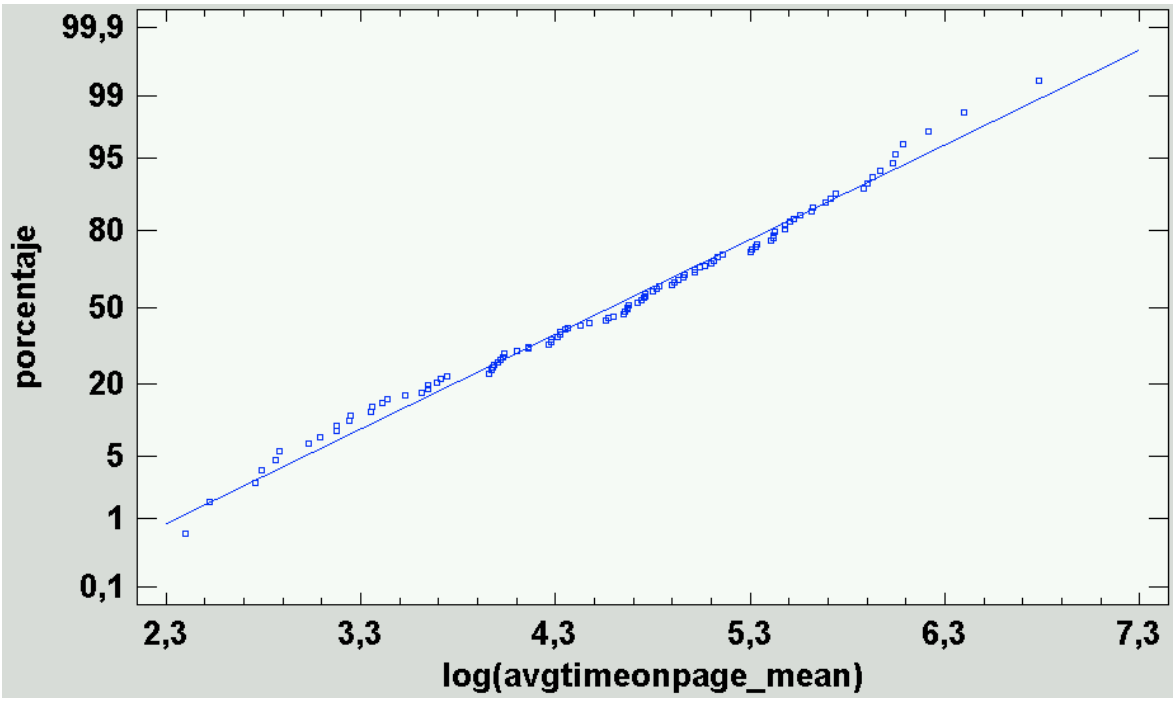


Figura 348. Tráileres: Gráfico de probabilidad normal de la variable log(avgtimeonpage_mean)

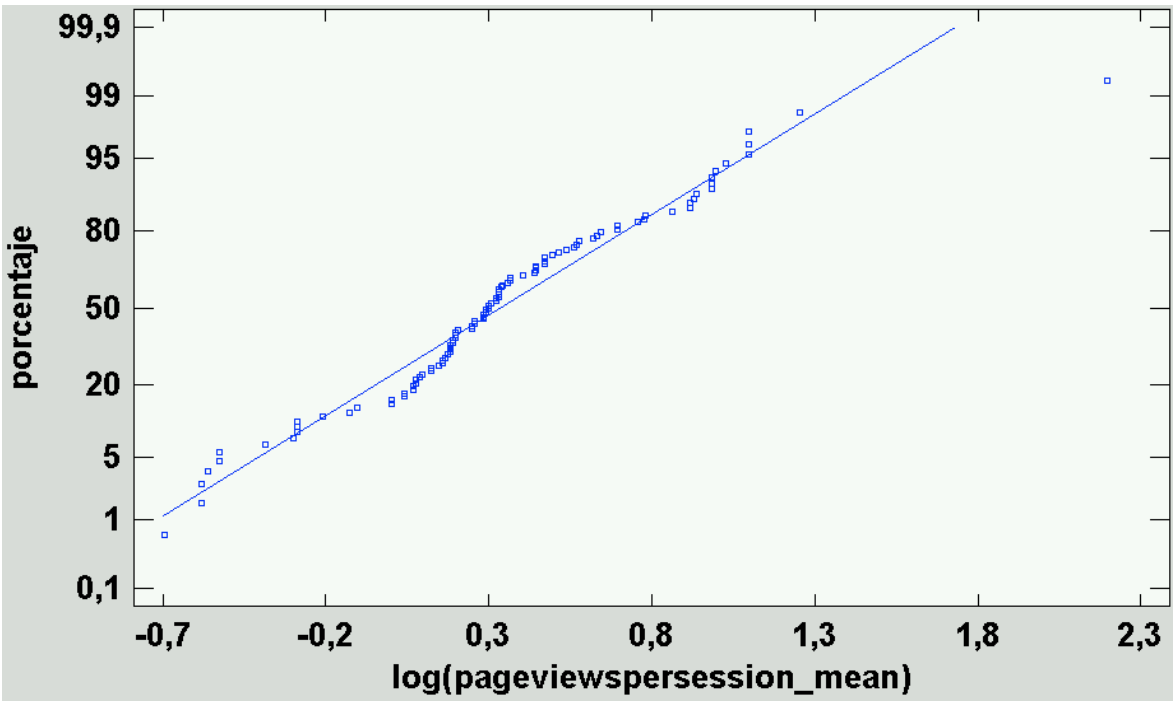


Figura 349. Tráileres: Gráfico de probabilidad normal de la variable log(pageviewspersession_mean)

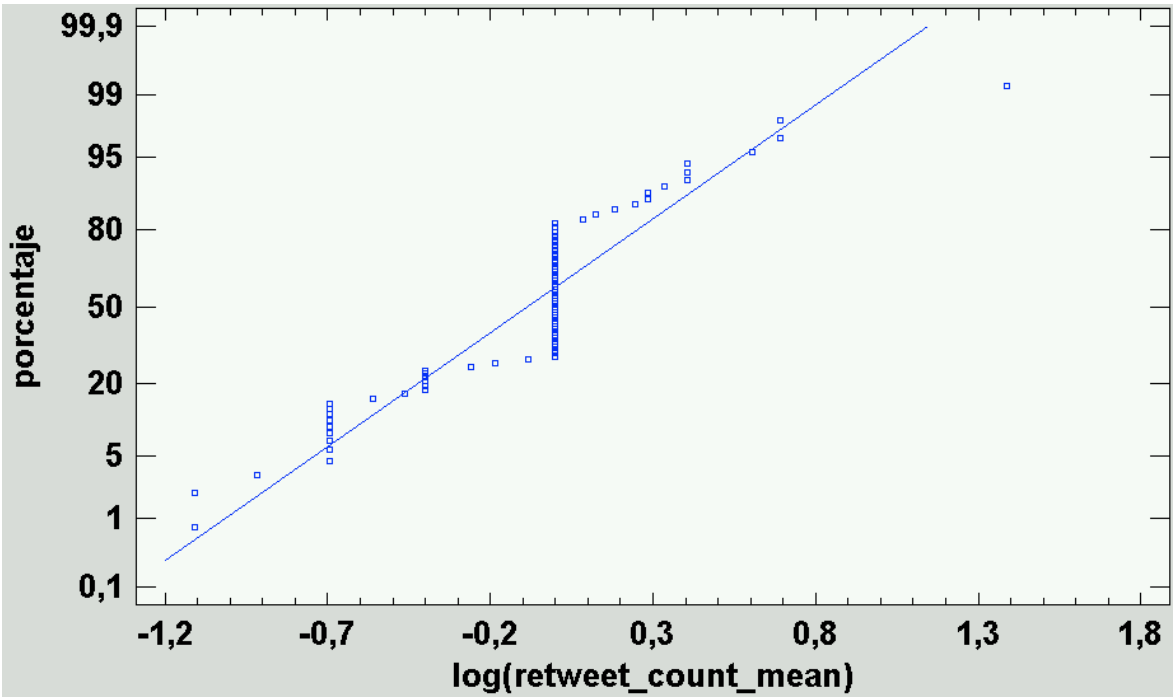


Figura 350. Tráileres: Gráfico de probabilidad normal de la variable log(retweet_count_mean)

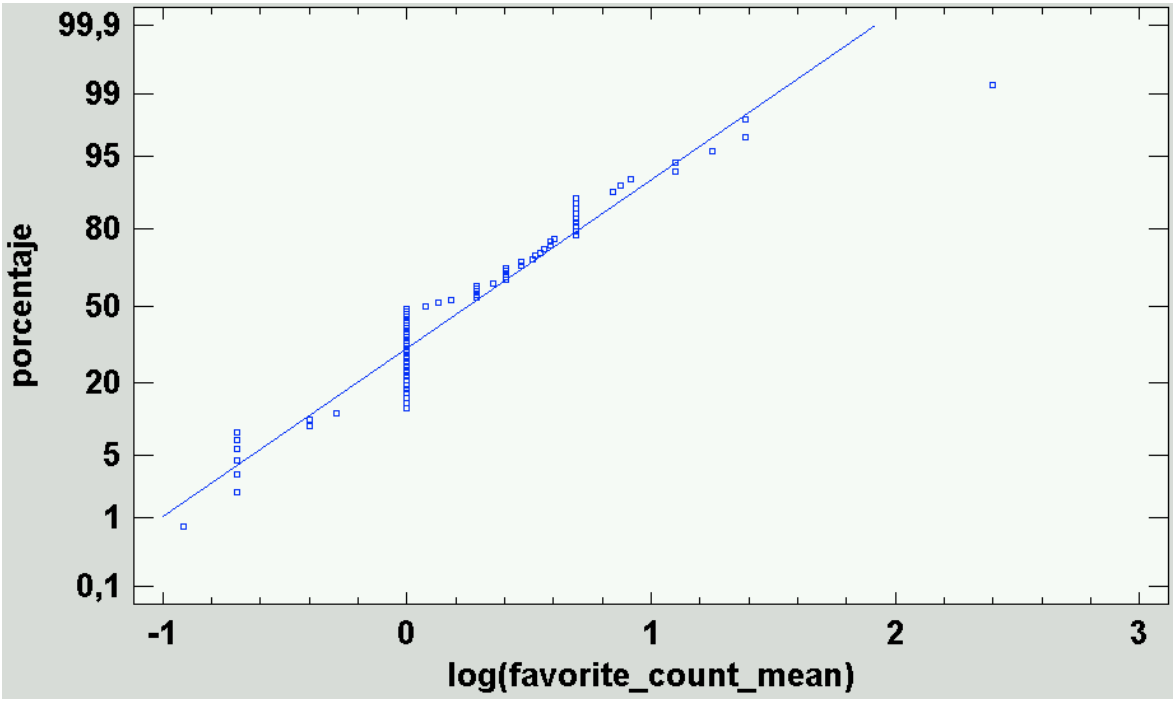


Figura 351. Tráileres: Gráfico de probabilidad normal de la variable log(favorite_count_mean)

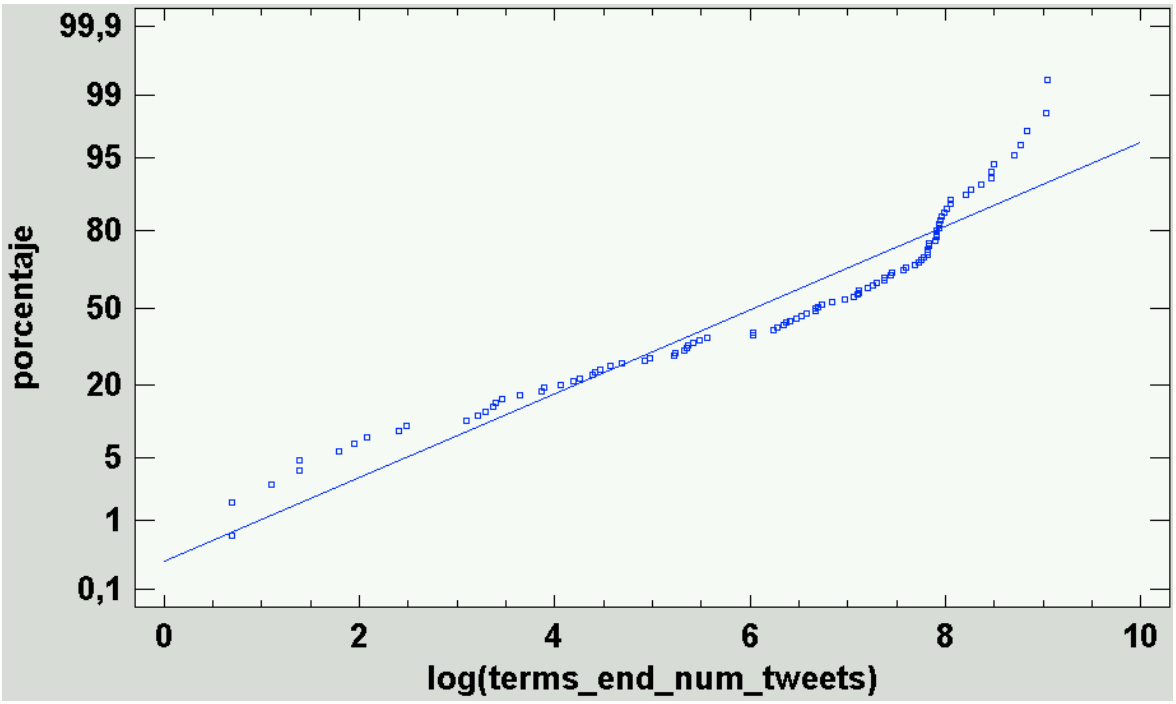


Figura 352. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_num_tweets)

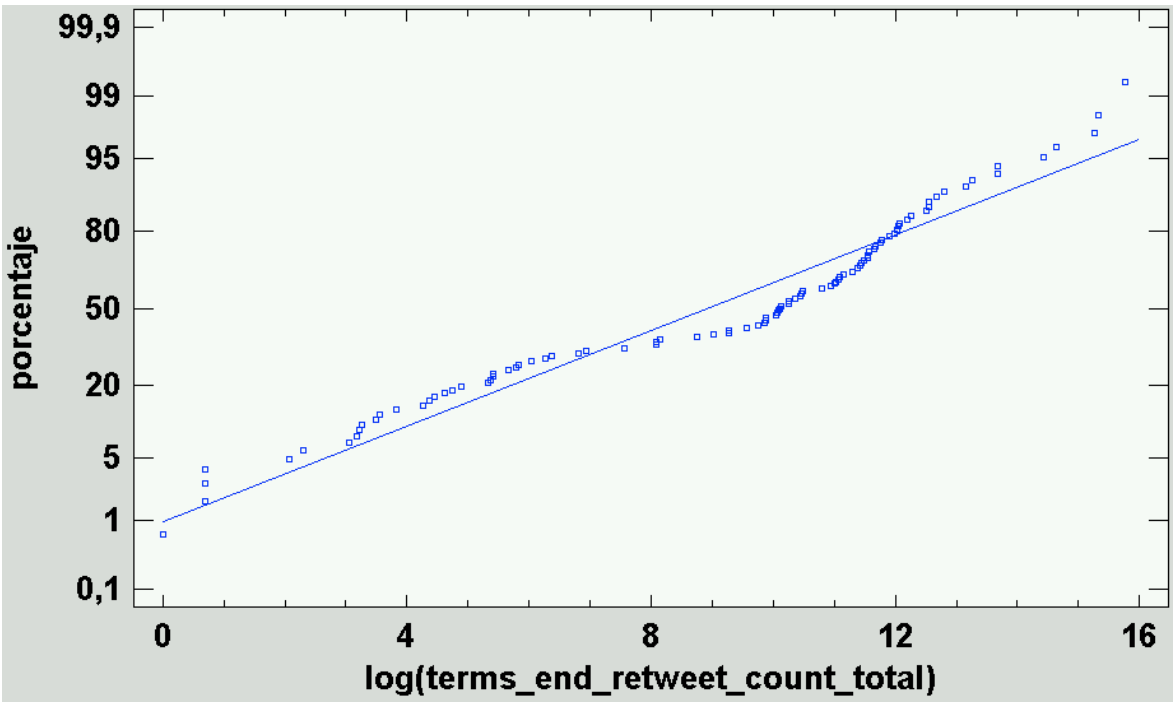


Figura 353. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_total)

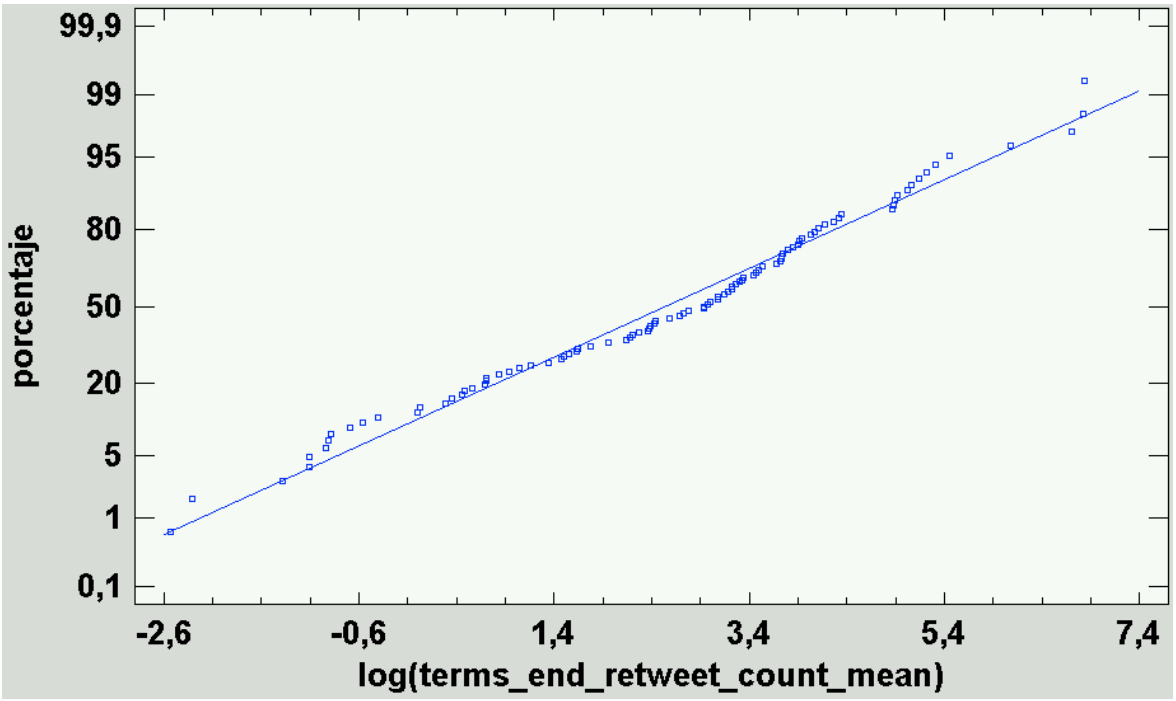


Figura 354. Tráileres: Gráfico de probabilidad normal de la variable log(terms_end_retweet_count_mean)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

k) Filtro de alta correlación

Puesto que se han propuesto más de una variable de éxito, es conveniente evaluar si están asociadas mediante una fuerte correlación. Si fuera ese el caso, el análisis de este estudio se simplificaría ya que las conclusiones de una variable servirían para las que estén fuertemente correlacionadas con ella, lo que permitiría dedicar menos recursos a la predicción y toma de decisiones.

Para ello, es necesario realizar un análisis multivariado de las variables de éxito con su correspondiente transformación logarítmica, cuya matriz de correlaciones Pearson se puede observar en la Figura 355:

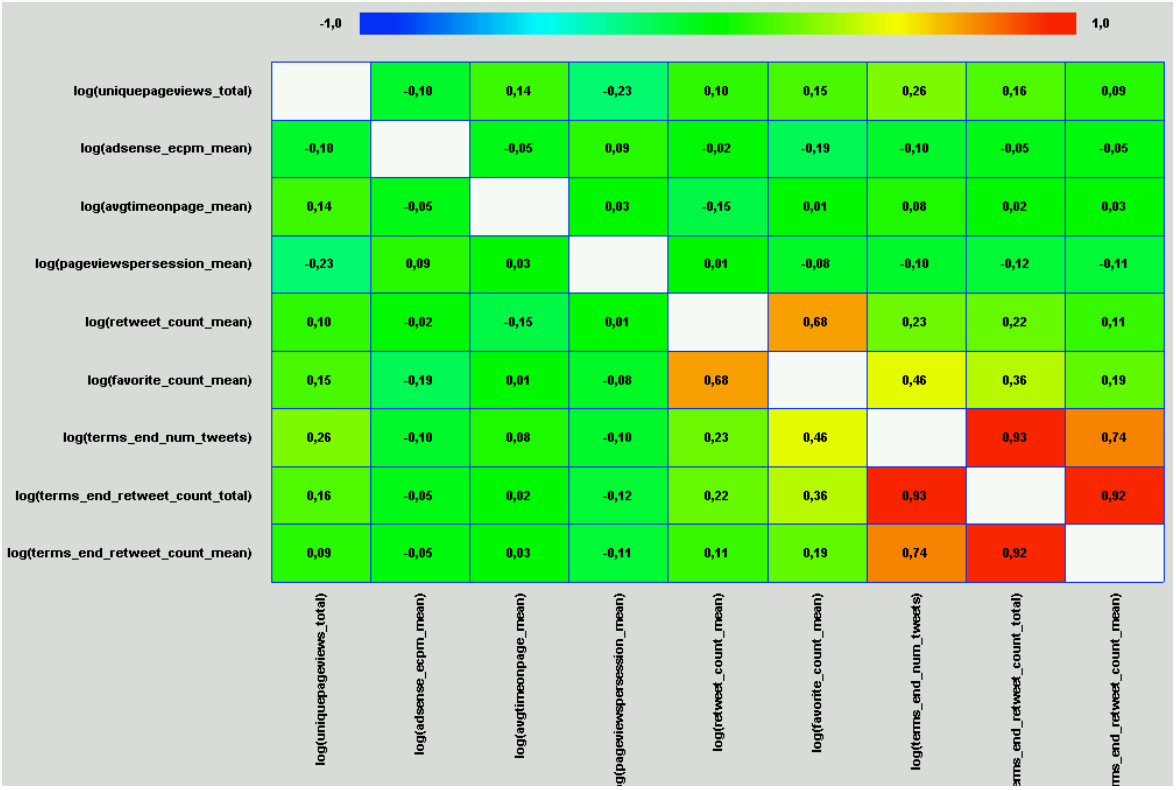


Figura 355. Tráileres: Matriz de correlaciones Pearson entre las variables de éxito transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- log(terms_end_num_tweets) y log(terms_end_retweet_count_total) tienen un coeficiente de correlación de 0,9336 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_end_num_tweets) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,74 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- log(terms_end_retweet_count_total) y log(terms_end_retweet_count_mean) tienen un coeficiente de correlación de 0,916 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige por tanto log(terms_end_retweet_count_mean) por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que log(terms_end_num_tweets) y log(terms_end_retweet_count_total).

La tabla de variables quedaría como sigue:

Tabla 156

Tráileres: Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de éxito.

Variables de predicción	Variables de éxito
terms_ini_num_tweets	log(uniquepageviews_total)
terms_ini_retweet_count_total	log(adsense_ecpm_mean)
terms_ini_retweet_count_mean	log(avgtimeonpage_mean)
terms_ini_favorite_count_total	log(pageviewspersession_mean)
terms_ini_favorite_count_mean	log(retweet_count_mean)
terms_ini_followers_talking_rate	log(favorite_count_mean)
terms_ini_user_num_followers_mean	log(terms_end_retweet_count_mean)
terms_ini_user_num_tweets_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de éxito queda, finalmente, limitada a la transformación logarítmica de: el número total de páginas vistas únicas, el promedio de eCPM de los anuncios de Google AdSense, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de retuits en la cuenta del medio, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

6.1.4.2. Variables de predicción

Se parte de un conjunto de datos con diez características, por lo que es conveniente tratar de reducir la dimensión del conjunto, restableciendo la varianza sin modificar la información relevante de los datos en sí. Esto posibilitará que se reduzca el tiempo y el coste de la computación y facilita la visualización y el análisis de los datos. Además, es una condición necesaria para aplicar la regresión lineal múltiple (Anon., 2017).

a) Número de tuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_num_tweets en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 101 valores con un rango de entre 3 y 9.849.

Presenta un sesgo estandarizado de 6,71346 y una curtosis estandarizada de 5,08282. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

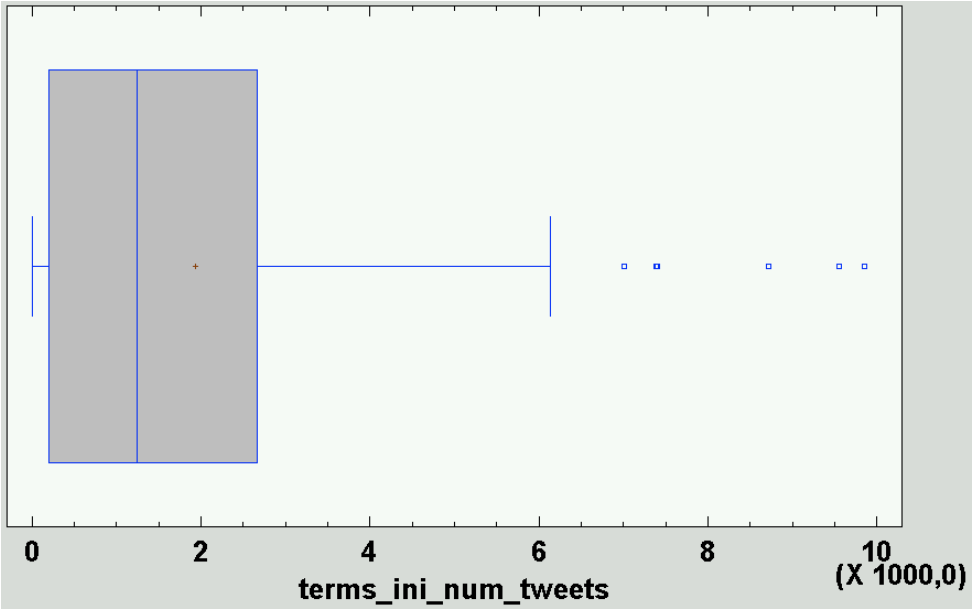


Figura 356. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_num_tweets

En la Figura 356 se puede observar que no existen valores anómalos de tipo extremo.

b) Número de retuits de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 101 valores con un rango de entre 0 y 9.805.270.

Presenta un sesgo estandarizado de 16,5118 y una curtosis estandarizada de 38,4365. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

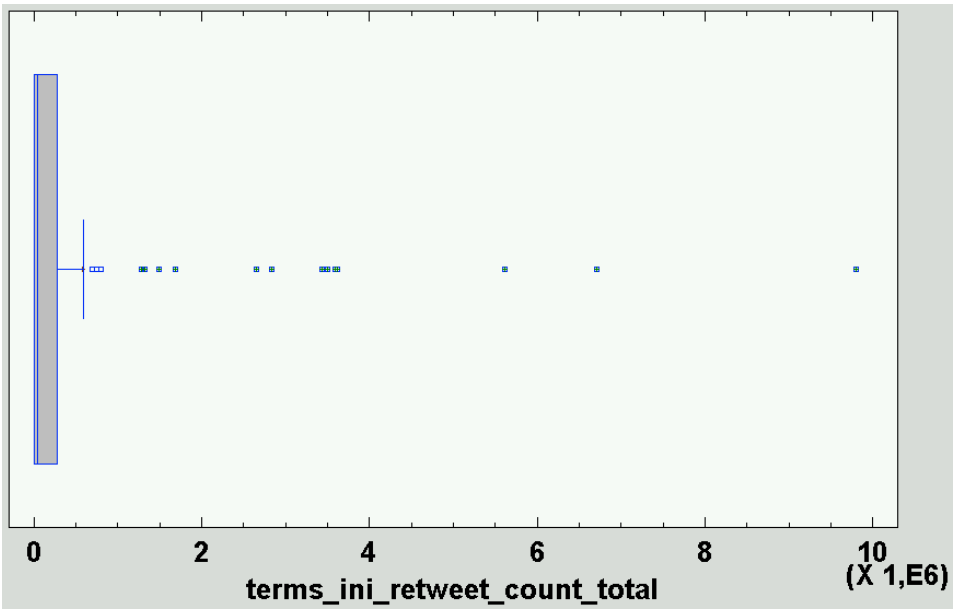


Figura 357. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_total

En la Figura 357 se puede observar que existen valores anómalos de tipo extremo de 1.282.790 o más.

c) Número de retuits de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_retweet_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 6.260,75.

Presenta un sesgo estandarizado de 34,8242 y una curtosis estandarizada de 161,753. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

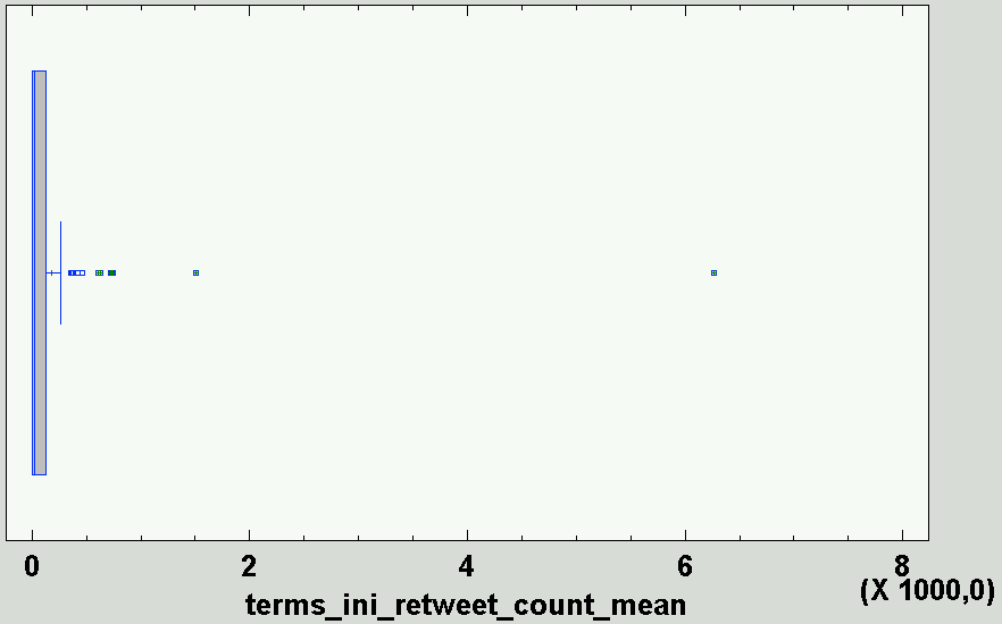


Figura 358. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_retweet_count_mean

En la Figura 358 se puede observar que existen valores anómalos de tipo extremo de 604,16 o más.

d) Número de favoritos de la tendencia inicial (total)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_total en la base de datos. Se trata de una variable aleatoria discreta que toma valores naturales. Consta de 101 valores con un rango de entre 4 y 37.092.

Presenta un sesgo estandarizado de 7,05374 y una curtosis estandarizada de 7,87476. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

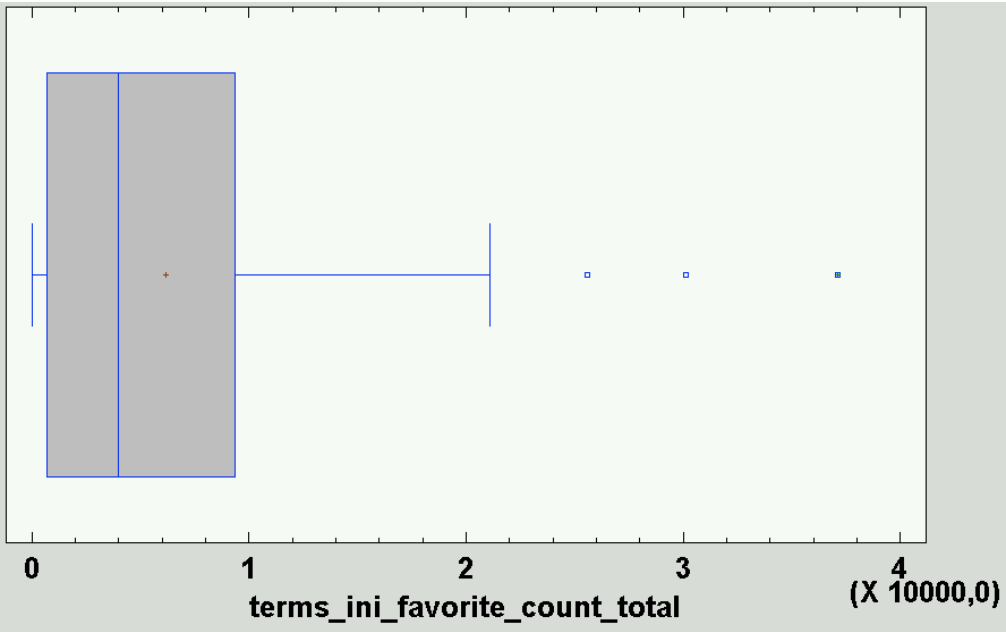


Figura 359. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_total

En la Figura 359 se puede observar que existe un valor anómalo de tipo extremo de 37.092.

e) Número de favoritos de la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_favorite_count_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0,28 y 12,88.

Presenta un sesgo estandarizado de 8,09904 y una curtosis estandarizada de 12,3009. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

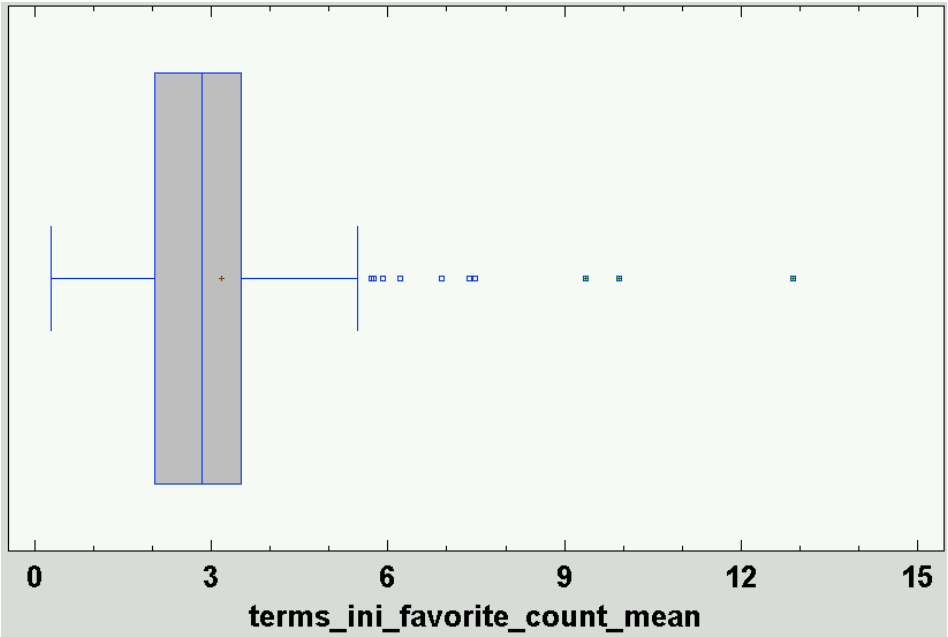


Figura 360. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_favorite_count_mean

En la Figura 360 se puede observar que existen valores anómalos de tipo extremo de 9,35 o más.

f) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_followers_talking_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0 y 0,5.

Presenta un sesgo estandarizado de 18,7582 y una curtosis estandarizada de 56,5774. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

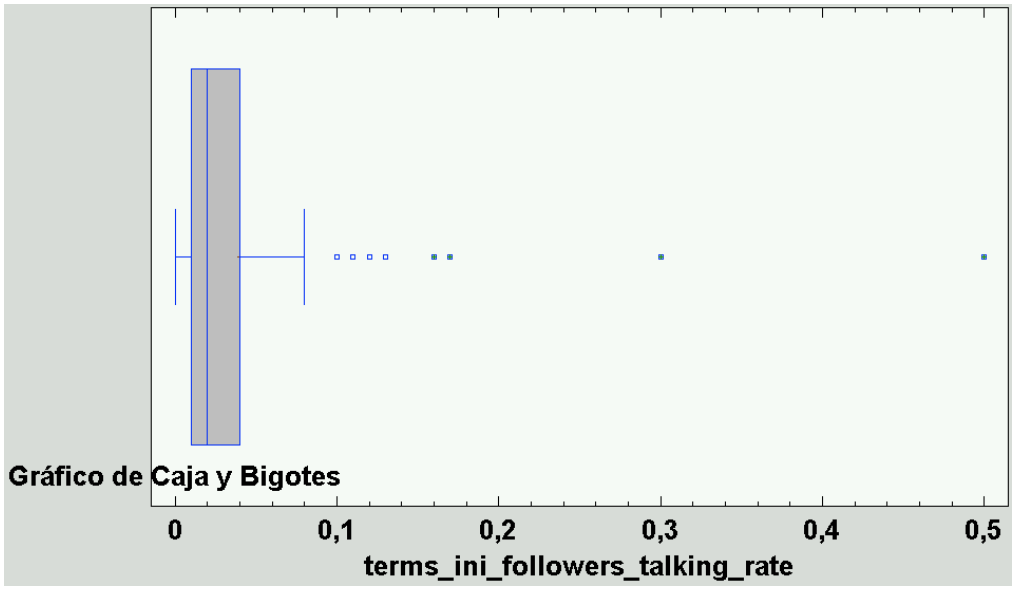


Figura 361. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_followers_talking_rate

En la Figura 361 se puede observar que existen valores anómalos de tipo extremo de 0,16 o más.

g) Número de seguidores de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_followers_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 1.345,26 y 124.117.

Presenta un sesgo estandarizado de 8,60221 y una curtosis estandarizada de 9,30456. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

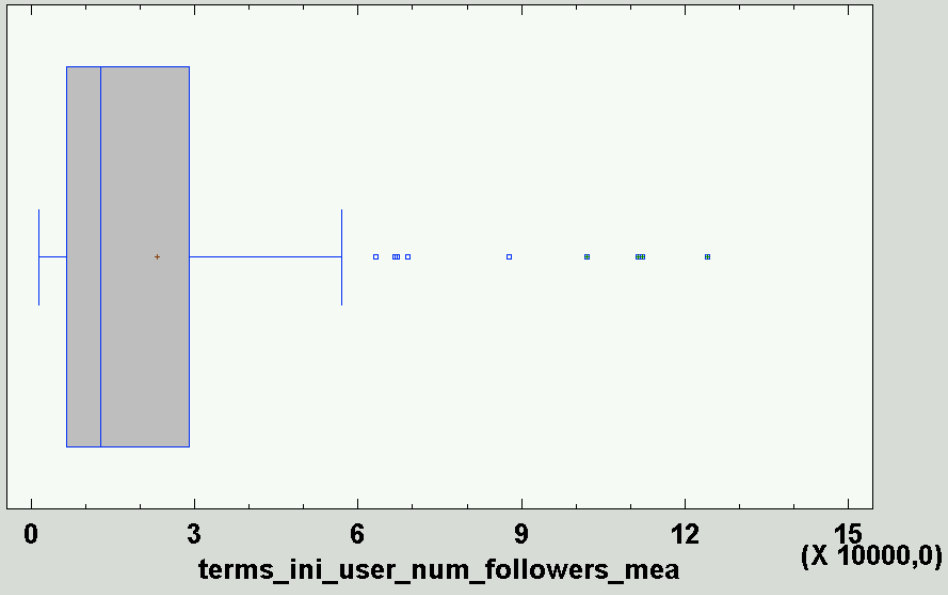


Figura 362. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_followers_mean

En la Figura 362 se puede observar que existen valores anómalos de tipo extremo de 102.013 o más.

h) Número de tuits de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_num_tweets_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 11.646,7 y 238.728.

Presenta un sesgo estandarizado de 14,669 y una curtosis estandarizada de 37,6611. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

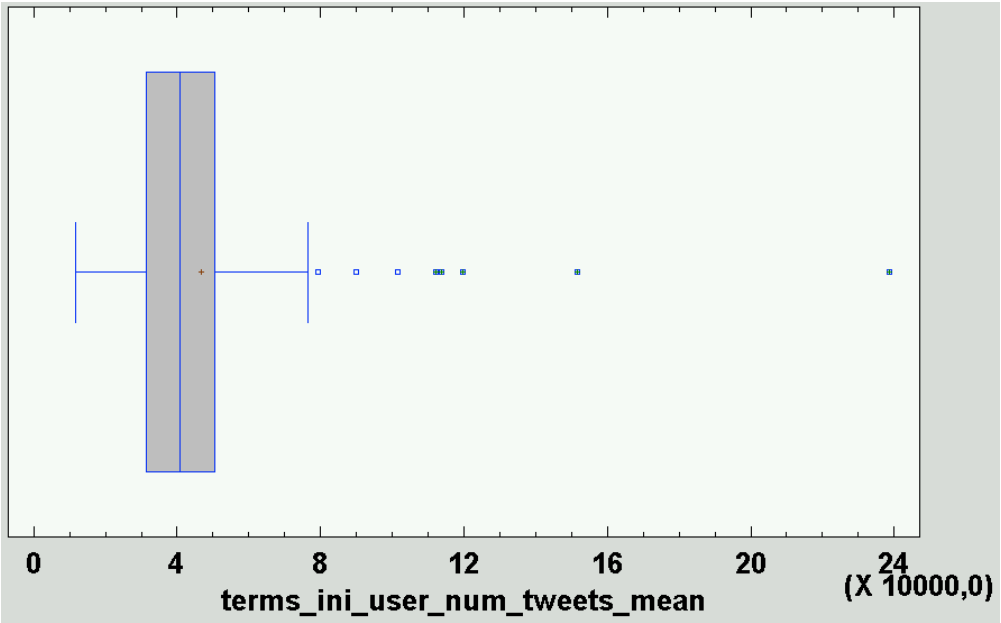


Figura 363. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_num_tweets_mean

En la Figura 363 se puede observar que existen valores anómalos de tipo extremo de 112.112 o más.

i) Edad en número de días de las cuentas de los usuarios que hablan sobre la tendencia inicial (promedio)

Esta variable de predicción se identifica con la columna terms_ini_user_age_mean en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 600,58 y 9.264.

Presenta un sesgo estandarizado de 21,7794 y una curtosis estandarizada de 90,6789. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

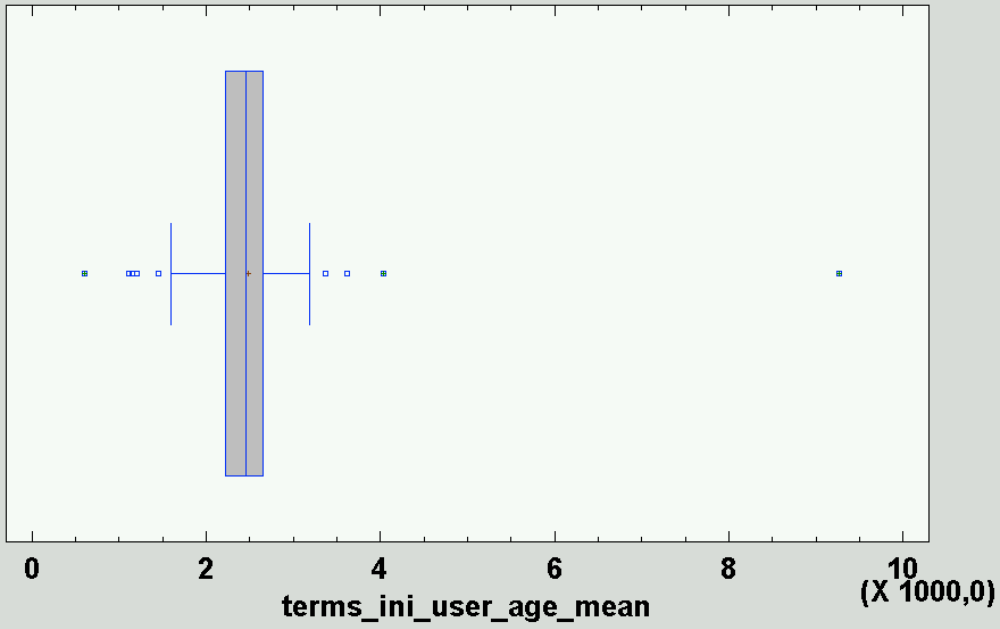


Figura 364. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_user_age_mean

En la Figura 364 se puede observar que existen valores anómalos de tipo extremo de 600,58, y 4.036,62 o más.

j) Ratio de inclusión de URLs en los tuits de la tendencia inicial

Esta variable de predicción se identifica con la columna terms_ini_url_inclusion_rate en la base de datos. Se trata de una variable aleatoria discreta que toma valores reales de dos decimales. Consta de 101 valores con un rango de entre 0,08 y 1.

Presenta un sesgo estandarizado de 3,02151 y una curtosis estandarizada de 1,46713. Puesto que ambos valores estadísticos están fuera del rango de -2 a +2, indican una desviación significativa de la normalidad, es decir, que ambos no se encuentran dentro del rango esperado para datos provenientes de una distribución normal.

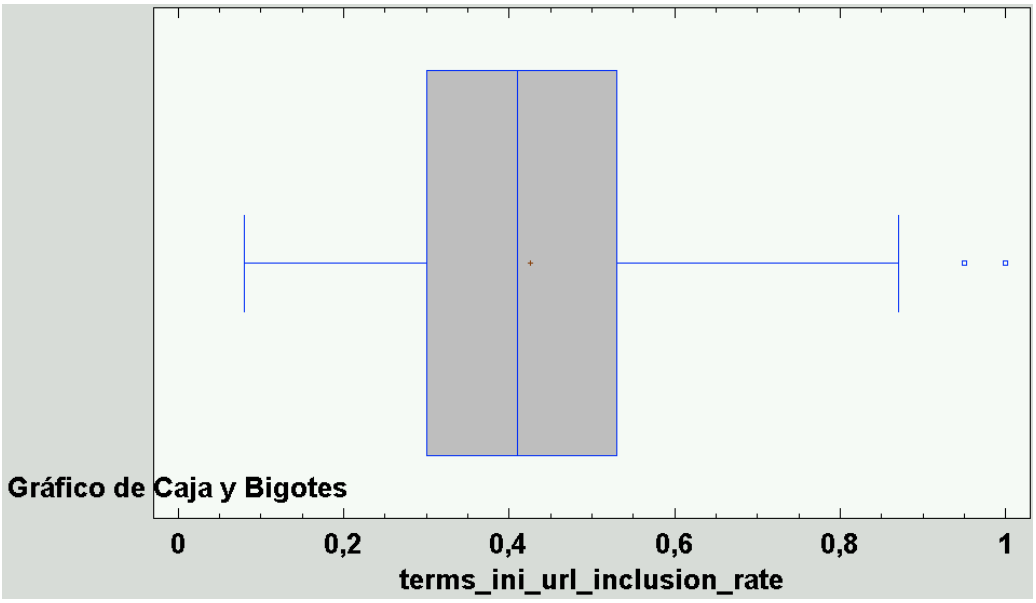


Figura 365. Tráileres: Gráfico de Caja y Bigotes para el valor terms_ini_url_inclusion_rate

En la Figura 365 se puede observar que no existen valores anómalos de tipo extremo.

k) Normalidad y equidistribución de los residuos

Los datos deben seguir una distribución normal para que puedan utilizarse para esta técnica (Barón López & Téllez Montiel, s.f.). Por ello, es necesario analizar las variables que vamos a utilizar en el modelo para comprobar si se distribuyen de modo normal.

Para analizar la normalidad de las variables, se ha realizado un análisis multivariado de todas las variables de predicción y éxito. De esta manera se obtienen todos los datos de las variables del modelo en un mismo análisis. Se pueden observar los siguientes datos de estas:

Tabla 157

Tráileres: Resumen estadístico de las variables de predicción

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada
terms_ini_num_tweets	5.083.710	6,71346	5,08282
terms_ini_retweet_count_total	2.169.690.000.000	16,5118	38,4365

terms_ini_retweet_count_mean	421.544	34,8242	161,753
terms_ini_favorite_count_total	49.796.300	7,05374	7,87476
terms_ini_favorite_count_mean	4,06346	8,09904	12,3009
terms_ini_followers_talking_rate	0,00416396	18,7582	56,5774
terms_ini_user_num_followers_mean	653.079.000	8,60221	9,30456
terms_ini_user_num_tweets_mean	880.103.000	14,669	37,6611
terms_ini_user_age_mean	694.197	21,7794	90,6789
terms_ini_url_inclusion_rate	0,0335951	3,02151	1,46713

Se puede observar en la Tabla 157 que todas las variables presentan un sesgo y una curtosis estandarizados fuera del rango -2 a +2, por lo que indican una desviación significativa de la normalidad.

Para realizar una regresión múltiple, es necesario aproximarnos a la normalidad con las variables que vamos a utilizar. La normalización estadística es una transformación de escala de la distribución de una variable, con el objetivo de hacer comparaciones eliminando los efectos de influencias. Es decir, permite comparar diferentes variables y distintas unidades de medida.

Para ello, en este estudio se ha hecho uso de la transformación más utilizada: el logaritmo (Marín Diazaraque, 2004). Se trata de una transformación muy útil para variaciones muy grandes en los valores, ya que comprime los valores muy altos y aumenta los muy bajos (Chinea, 2008), y especialmente útil para distribuciones con sesgo positivo (hacia la derecha) (Anon., 2015).

Tabla 158

Tráileres: Resumen estadístico de las variables de predicción con transformación logarítmica

Variable	Varianza	Sesgo Estandarizado	Curtosis Estandarizada

log(terms_ini_num_tweets)	3,67038	-3,31232	-0,398357
log(terms_ini_retweet_count_total)	12,713	-2,37865	-0,763882
log(terms_ini_retweet_count_mean)	4,16563	-0,313705	-0,790007
log(terms_ini_favorite_count_total)	4,81967	-4,38542	0,843842
log(terms_ini_favorite_count_mean)	0,427329	-3,09302	3,40637
log(terms_ini_followers_talking_rate)	0,810599	2,39324	0,316208
log(terms_ini_user_num_followers_mean)	1,05146	0,329779	-1,23355
log(terms_ini_user_num_tweets_mean)	0,22794	1,96191	4,33613
log(terms_ini_user_age_mean)	0,0777554	-2,38104	22,9556
log(terms_ini_url_inclusion_rate)	0,221838	-2,87799	2,00212

log(terms_ini_retweet_count_total), log(terms_ini_favorite_count_total), log(terms_ini_favorite_count_mean), log(terms_ini_followers_talking_rate), log(terms_ini_user_num_tweets_mean), log(terms_ini_user_age_mean) y log(terms_ini_url_inclusion_rate) mantienen un sesgo o una curtosis estandarizados fuera del rango de -2 a +2, por lo que no siguen una distribución estrictamente normal. En este estudio, debido a la naturaleza de los datos, se va a considerar que son datos casi normales aquellos que estén en un rango de -5 a +5 en sesgo y curtosis estandarizados.

También se puede apreciar que los valores de varianza son bastante parecidos salvo en el caso de log(terms_ini_retweet_count_total), por lo que se cumple la condición de homocedasticidad menos en esa variable. log(terms_ini_retweet_count_total) se elimina del modelo actual para que dicho requisito se cumpla.

Nos quedamos, por tanto, con las variables que tengan menores sesgo y curtosis estandarizados de su forma original o transformación logarítmica.

Puesto que hay variables con transformación logarítmica con valores de sesgo y curtosis estandarizados fuera del rango -5 a +5, se realiza a continuación una prueba no paramétrica de Kolmogórov-Smirnov a dichas variables para comprobar la bondad de ajuste a una distribución normal (Ruiz Mitjana, 2019). Para ello, se calcula el valor-p de la prueba de Kolmogórov-Smirnov y si es mayor o igual que 0,05, no se puede rechazar que la variable provenga de una distribución normal con un 95% de confianza.

Tabla 159

Tráileres: Valor-p de la prueba de Kolmogórov-Smirnov de las variables de éxito con transformación logarítmica

Variable	Valor-p
log(terms_ini_user_age_mean)	0,0000897383

En la Tabla 159 se puede ver que log(terms_ini_user_age_mean) no supera el valor-p necesario (mayor o igual que 0,05) para confirmar que siguen una distribución normal, por lo que dicha variable no es tenida en cuenta en el modelo.

Por todo ello, y tras quitar las variables con valores de sesgo y curtosis estandarizados fuera de rango -5 a +5 y que no hayan pasado la prueba de Kolmogórov-Smirnov, la tabla queda así:

Tabla 160

Tráileres: Lista de variables de predicción y de éxito tras el estudio de normalidad y equidistribución de residuos de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_num_tweets)	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_total)	log(avgtimeonpage_mean)

log(terms_ini_favorite_count_mean)	log(pageviewspersession_mean)
log(terms_ini_followers_talking_rate)	log(retweet_count_mean)
log(terms_ini_user_num_followers_mean)	log(favorite_count_mean)
log(terms_ini_user_num_tweets_mean)	log(terms_end_retweet_count_mean)
log(terms_ini_url_inclusion_rate)	

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de tuits de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

En este estudio se ha tomado como casi normalidad un rango más amplio de -5 a +5 debido a la naturaleza de los datos, por lo que es adecuado analizar los gráficos de probabilidad normal para contrastar su casi normalidad de manera gráfica:

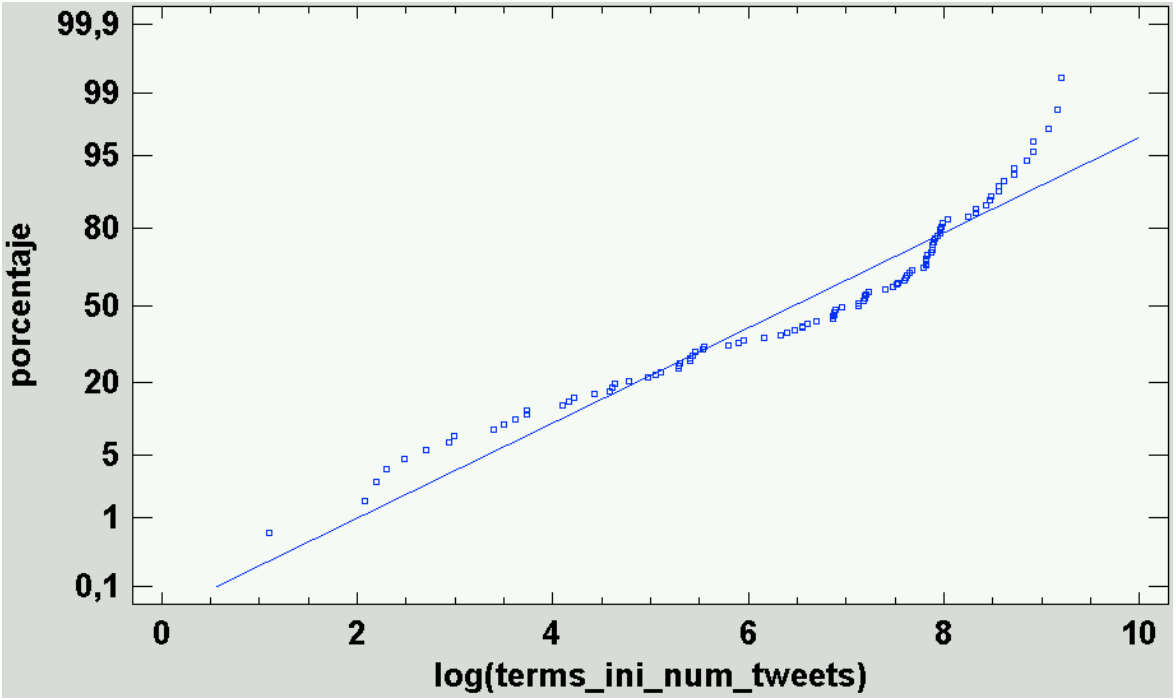


Figura 366. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_num_tweets)

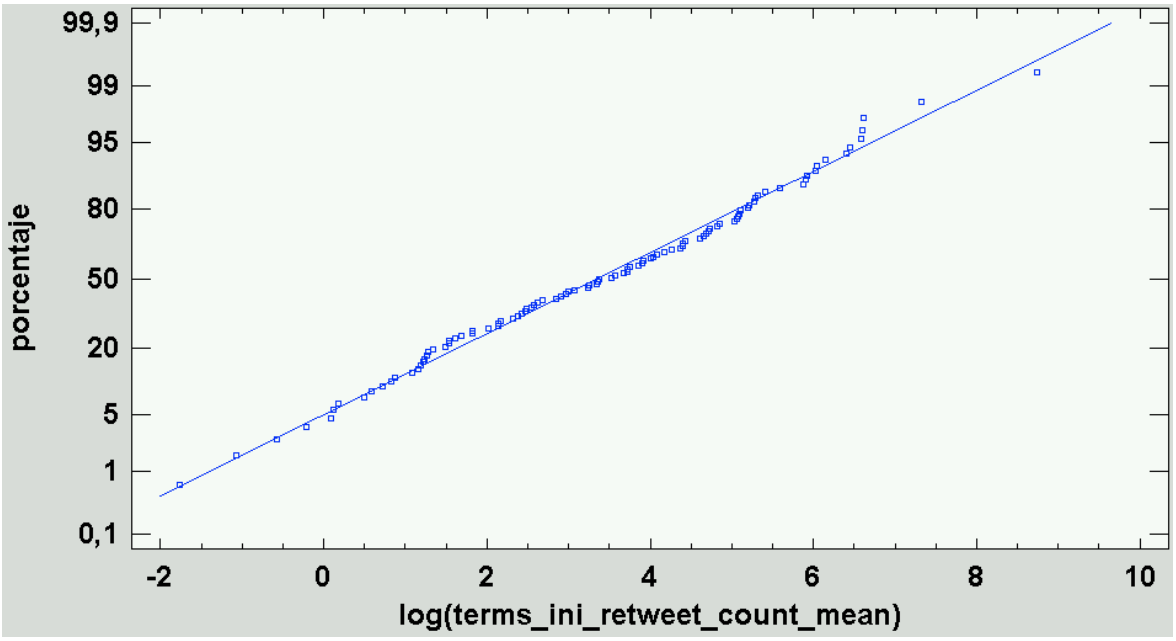


Figura 367. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_retweet_count_mean)

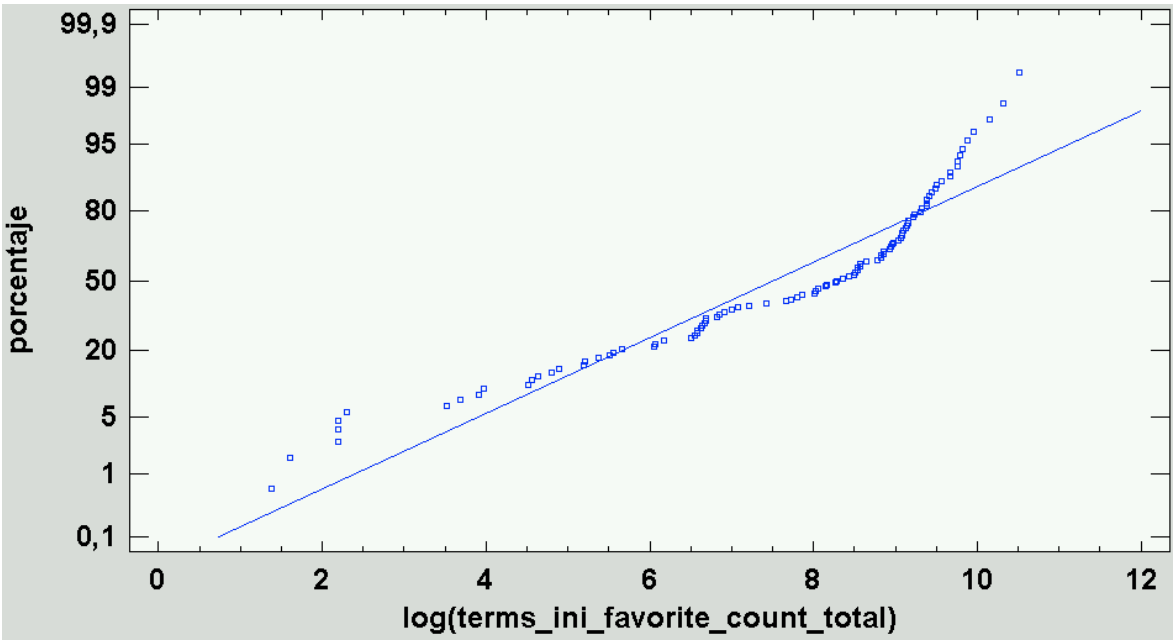


Figura 368. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_total)

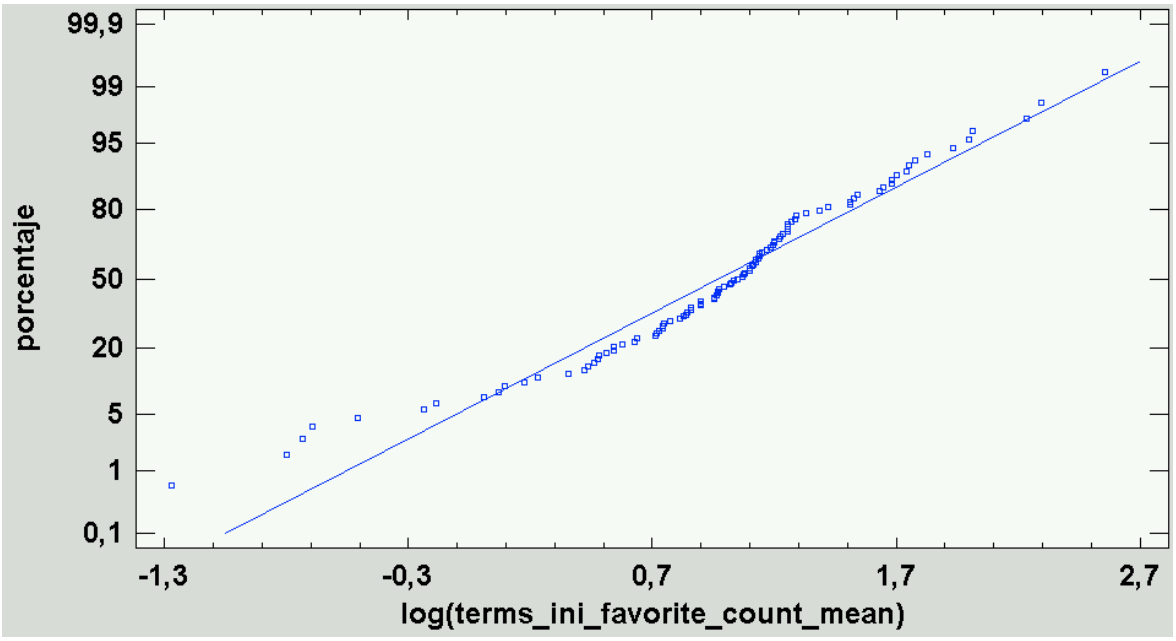


Figura 369. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_favorite_count_mean)

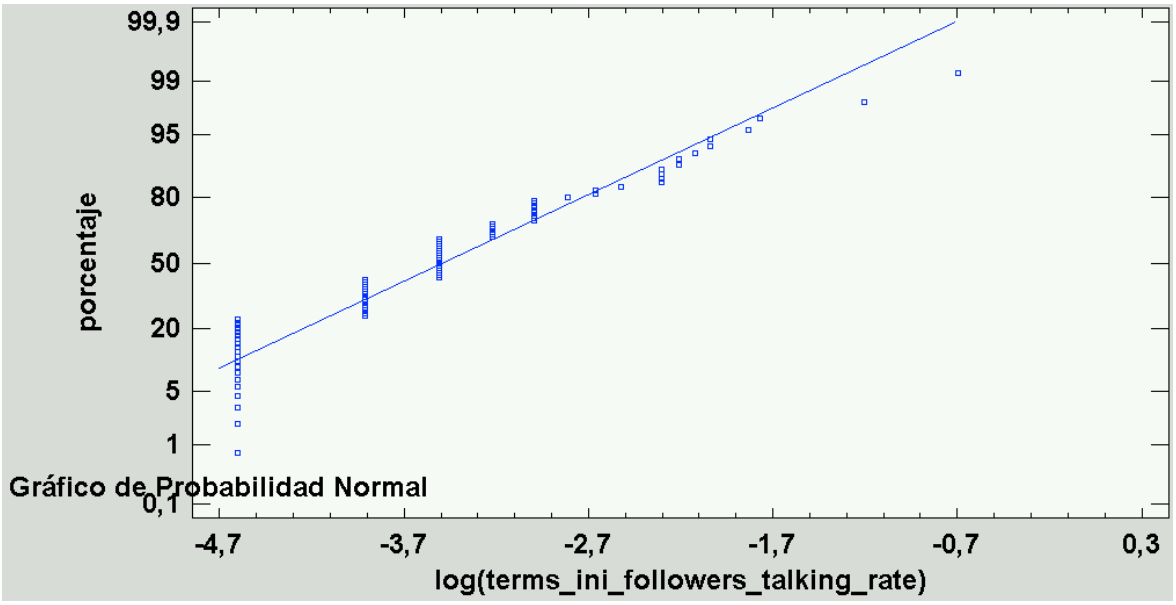


Figura 370. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_followers_talking_rate)

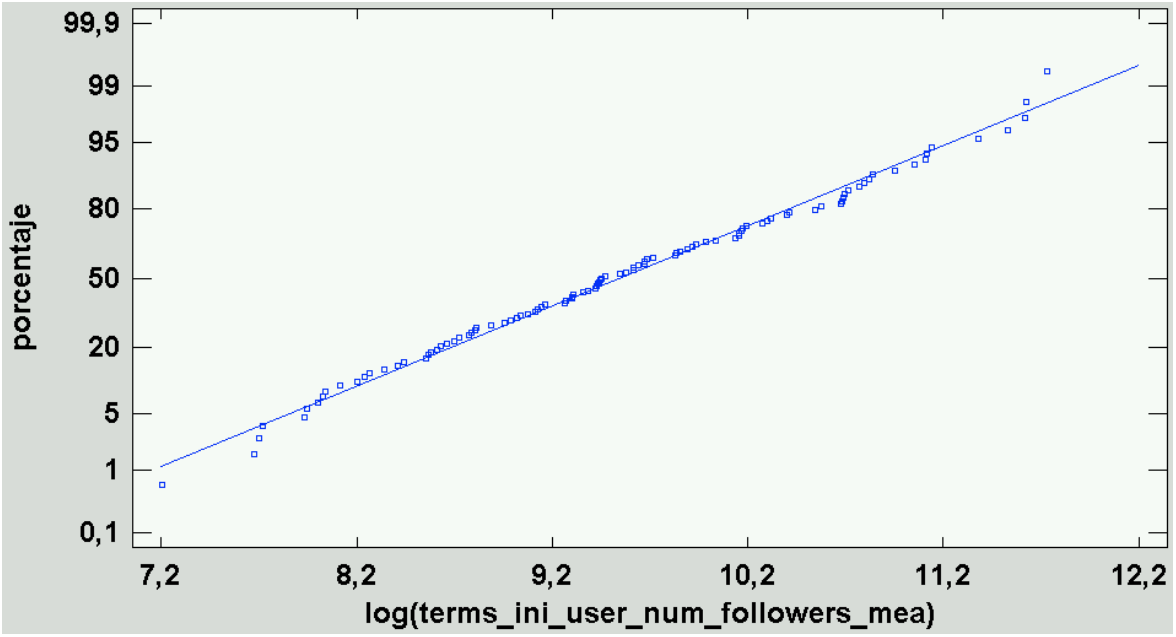


Figura 371. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_followers_mean)

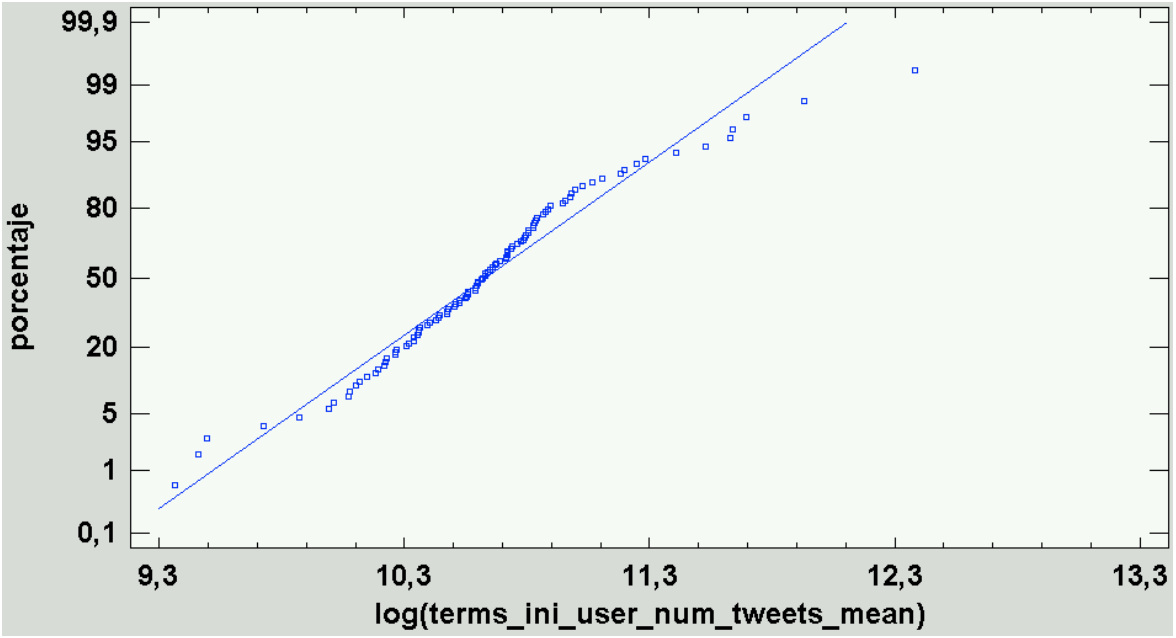


Figura 372. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_user_num_tweets_mean)

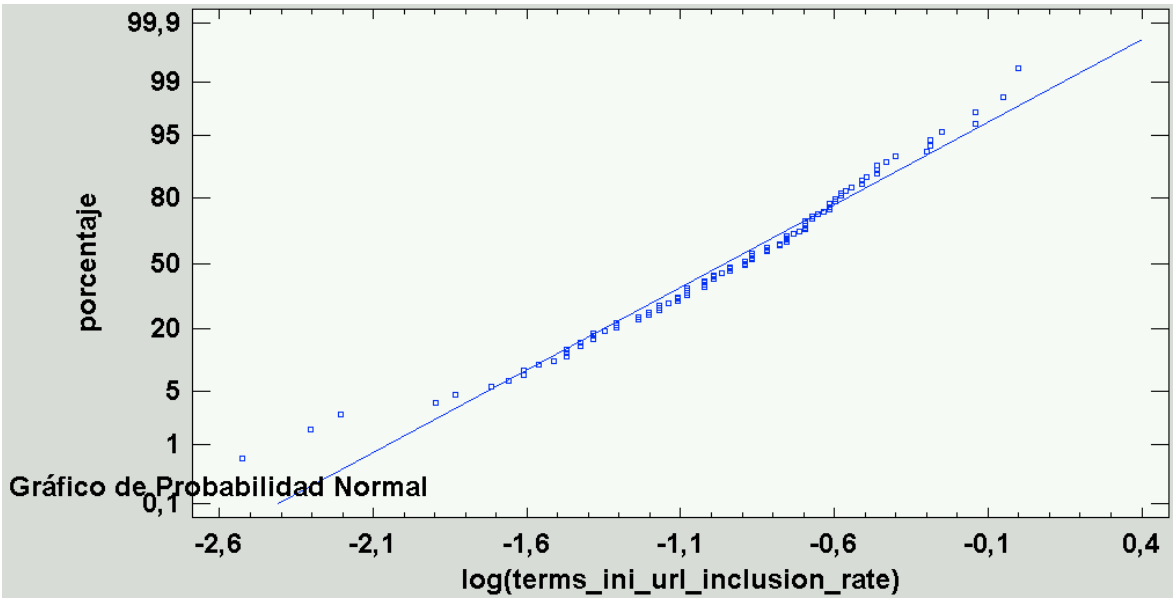


Figura 373. Tráileres: Gráfico de probabilidad normal de la variable log(terms_ini_url_inclusion_rate)

Se puede comprobar en las figuras anteriores que los datos siguen mayormente una distribución casi normal, ya que los puntos quedan muy cercanos a una línea recta.

l) Filtro de alta correlación (colinealidad)

Antes de proceder, es conveniente analizar la correlación que existe entre las variables con las que contamos para el modelo. Puesto que estas han sido normalizadas en el apartado anterior, dicho análisis de correlación Pearson se realizará con su transformación logarítmica.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

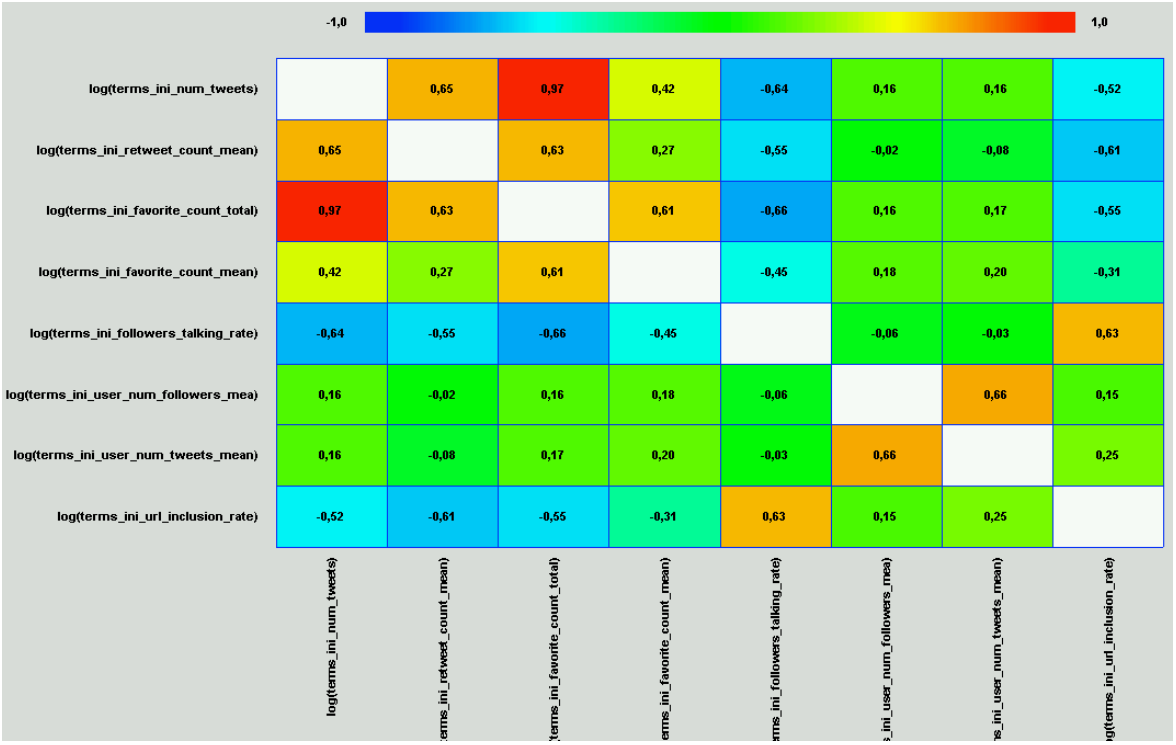


Figura 374. Tráileres: Matriz de correlaciones Pearson entre las variables de predicción transformadas logarítmicamente

Al hacerlo, se han obtenido las siguientes conclusiones:

- log(terms_ini_num_tweets) y log(terms_ini_favorite_count_total) tienen un coeficiente de correlación de 0,9706 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige log(terms_ini_num_tweets) por tener un sesgo y una curtosis estandarizados más cerca de seguir una distribución estrictamente normal que log(terms_ini_favorite_count_total).

La tabla de variables quedaría como sigue:

Tabla 161

Lista de variables de predicción y de éxito tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
log(terms_ini_num_tweets)	log(uniquepageviews_total)

log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	log(retweet_count_mean)
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

La lista de variables de predicción queda, por tanto, limitada a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

m) Análisis de componentes principales (ACP)

A continuación, se aplica el análisis de componentes principales (ACP, o PCA en inglés), una técnica que sirve par describir un conjunto de datos según nuevas variables no correlacionadas llamadas componentes (Dunteman, 1989).

El objetivo es representar los datos de la mejor manera posible a través de mínimos cuadrados, construyendo una transformación lineal según un nuevo sistema de coordenadas para los datos originales. Es decir, se plantea representar la variabilidad de los datos con el menor número de componentes o fórmulas posible, las cuales son combinaciones lineales de las variables originales (Dunteman, 1989).

Si las variables originales están muy correlacionadas entre sí, la mayor parte de la variabilidad se podrá expresar en pocas componentes. Si están totalmente incorrelacionadas, el número de componentes será igual al de las variables y este análisis carecerá de interés.

Las componentes se ordenan según la varianza original siendo la primer componente el que tenga la varianza de mayor tamaño. Cuanto mayor sea su varianza, mayor será la información que aporta esa componente (Amat Rodrigo, 2017).

Al construir la matriz de coeficientes de correlación, es posible una base de vectores propios, cuya transformación lineal es necesaria para mejora la simplicidad e interpretación

que permita tratar de reducir la dimensionalidad de los datos (Dunteman, 1989). Esta reducción se efectuaría seleccionando las componentes principales que más aportan a la varianza e ignorando el resto. Esta selección se produce ordenando las componentes de mayor a menor aportación a la explicación de la variabilidad, y seleccionando tantas como sean necesarias hasta alcanzar un valor propio mayor o igual a 1.

De esta manera, el método ACP condensa la información de múltiples características en unas pocas, ya que se pretende explicar aproximadamente la información con menos valores que los originales.

Realizando el análisis de componentes principales se ha obtenido un total de dos componentes, explicando así el 69,922% de los datos con un valor propio de 1,83068. Las dos componentes tienen la Tabla 162 de pesos, siendo cada peso un valor de entre -1 y 1.

Tabla 162

Tráileres: Tabla de pesos de las componentes

	Componente	*Componente*
	1	2
log(terms_ini_num_tweets)	0,48257	0,107765
log(terms_ini_retweet_count_mean)	0,458367	-0,118577
log(terms_ini_favorite_count_mean)	0,338921	0,214771
log(terms_ini_followers_talking_rate)	-0,485537	0,0074822
log(terms_ini_user_num_followers_mea)	0,0531086	0,643752
log(terms_ini_user_num_tweets_mean)	0,0218076	0,669284
log(terms_ini_url_inclusion_rate)	-0,450684	0,256486

De esta manera, por ejemplo, la primer componente principal tiene la fórmula siguiente, en donde los valores de las variables se han estandarizado restándoles su promedio y dividiéndolos entre su desviación estándar:

0,48257 * log(terms_ini_num_tweets) + 0,458367 * log(terms_ini_retweet_count_mean) + 0,338921 * log(terms_ini_favorite_count_mean) – 0,485537 * log(terms_ini_followers_talking_rate) + 0,0531086 * log(terms_ini_user_num_followers_mean) + 0,0218076 * log(terms_ini_user_num_tweets_mean) – 0,450684 * log(terms_ini_url_inclusion_rate)

Se observa que las variables que aportan positivamente a las dos componentes principales son: el número de tuits, el promedio de favoritos, el promedio de seguidores de los usuarios que participan en la tendencia y el promedio de tuits de los usuarios que participan en la tendencia. El resto sí que aportan un valor negativo en alguna de ellas.

La relación entre las variables y las dos componentes principales se puede ver en la siguiente gráfica, ya que las variables se muestran en dos dimensiones formadas por estas componentes:

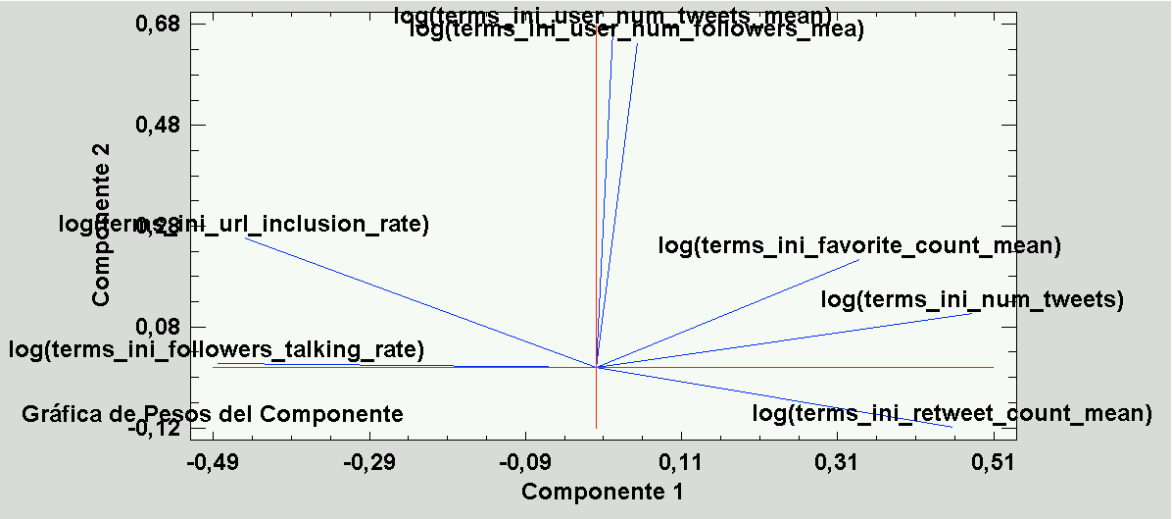


Figura 375. Tráileres: Gráfica de pesos de cada componente principal

En la Tabla 162 se puede comprobar que todas las variables tienen una presencia significativa en alguna de las componentes principales, por lo que no se puede eliminar ninguna de las variables originales mediante esta técnica.

A continuación, se van a analizar todos los artículos que traten sobre tráileres, de manera que se pueda comprobar si las características de los datos y las ecuaciones de predicción varían según la sección a la que pertenezca el artículo.

6.1.4.3. Regresión lineal múltiple

Para estudiar la posible relación entre las variables independientes de predicción de que disponemos y cada variable dependiente de éxito o, dicho de otro modo, para tratar de predecir el cálculo de estas, vamos a realizar un modelo de regresión múltiple.

Para realizar las regresiones múltiples, se cuenta con la Tabla *163* de variables resultante de todos los análisis anteriores:

Tabla 163

Tráileres: Lista final de variables de predicción y de éxito para la regresión lineal múltiple

Variables de predicción	Variables de éxito

log(terms_ini_num_tweets)	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	log(retweet_count_mean)
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

Las variables de predicción responden a la transformación logarítmica de: el promedio de retuits, el promedio de favoritos, el promedio de seguidores de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos aplicados a la tendencia el día de la publicación del artículo.

La lista de variables de éxito está formada por la transformación logarítmica de: el número total de páginas vistas únicas, la duración promedio de la visita, el promedio de páginas vistas por sesión, el promedio de favoritos en la cuenta del medio y el promedio de retuits de la tendencia 14 días después.

a) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión múltiple con la variable dependiente log(uniquepageviews_total). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 164

Tráileres: Valor-P de las variables de la regresión múltiple de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	3,38882	0,1782
log(terms_ini_num_tweets)	0,208386	0,0104

log(terms_ini_retweet_count_mean)	-0,0576398	0,3931
log(terms_ini_favorite_count_mean)	-0,209212	0,2382
log(terms_ini_followers_talking_rate)	-0,0395834	0,8002
log(terms_ini_user_num_followers_mean)	-0,0768118	0,54
log(terms_ini_user_num_tweets_mean)	-0,070531	0,8059
log(terms_ini_url_inclusion_rate)	0,0852412	0,7759
Modelo		0,2179

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 165

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(uniquepageviews_total)

Variable	Estimación	Valor-P
Constante	2,23199	0
log(terms_ini_num_tweets)	0,10475	0,0132
Modelo		0,0132

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(2,23199 + 0,10475 * log(terms_ini_num_tweets))

Para realizar el cálculo de uniquepageviews_total será necesario calcular el exponente de ambos lados de la fórmula, ya que el exponente es la función inversa del logaritmo.

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

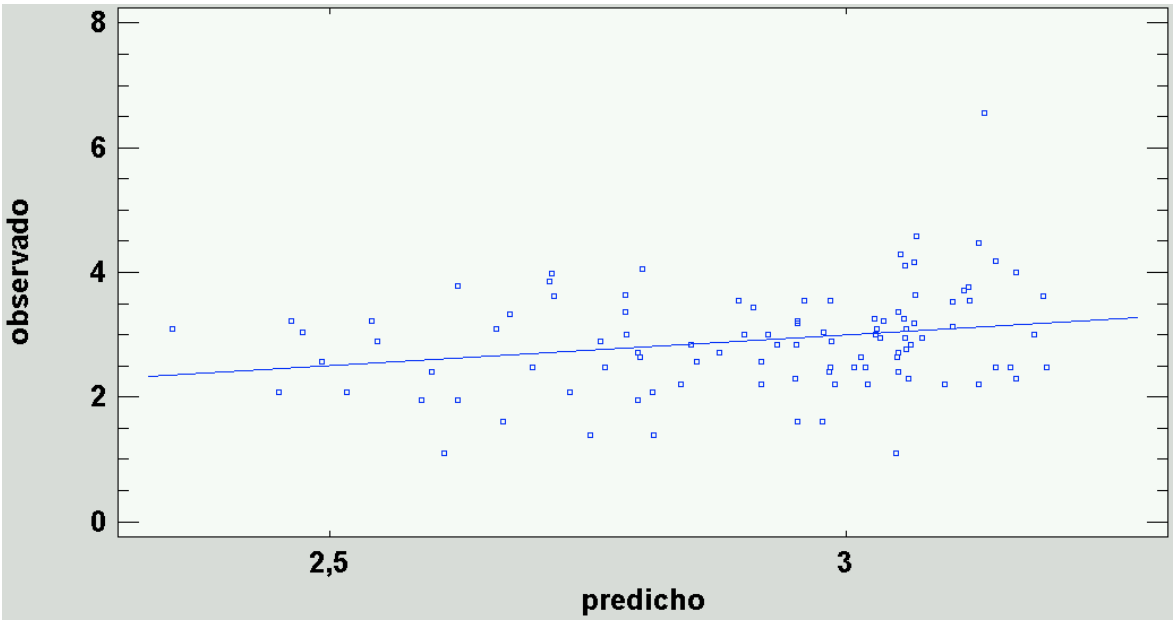


Figura 376. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(uniquepageviews_total) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 6,04245% de la variabilidad de log(uniquepageviews_total), mientras que el R-Cuadrado ajustado indica un 5,09339%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

b) AdSense eCPM (promedio)

Para tratar de predecir el valor de estimación de ingresos de los anuncios por cada 1000 páginas vistas desde los anuncios de Google AdSense es necesario realizar la regresión múltiple con la variable dependiente log(adsense_ecpm_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 166

Tráileres: Valor-P de las variables de la regresión múltiple de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P

Constante	-1,71927	0,5956
log(terms_ini_num_tweets)	0,0265497	0,7646
log(terms_ini_retweet_count_mean)	-0,0582236	0,4501
log(terms_ini_favorite_count_mean)	0,122727	0,5511
log(terms_ini_followers_talking_rate)	0,146408	0,4116
log(terms_ini_user_num_followers_mean)	-0,306232	0,0436
log(terms_ini_user_num_tweets_mean)	0,136519	0,7059
log(terms_ini_url_inclusion_rate)	-0,178493	0,588
Modelo		0,4871

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 167

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(adsense_ecpm_mean)

Variable	Estimación	Valor-P
Constante	-0,886954	0,3323
log(terms_ini_user_num_followers_mean)	-0,268537	0,0062
Modelo		0,0062

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

adsense_ecpm_mean = exp(-0,886954 - 0,268537 * log(terms_ini_user_num_followers_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

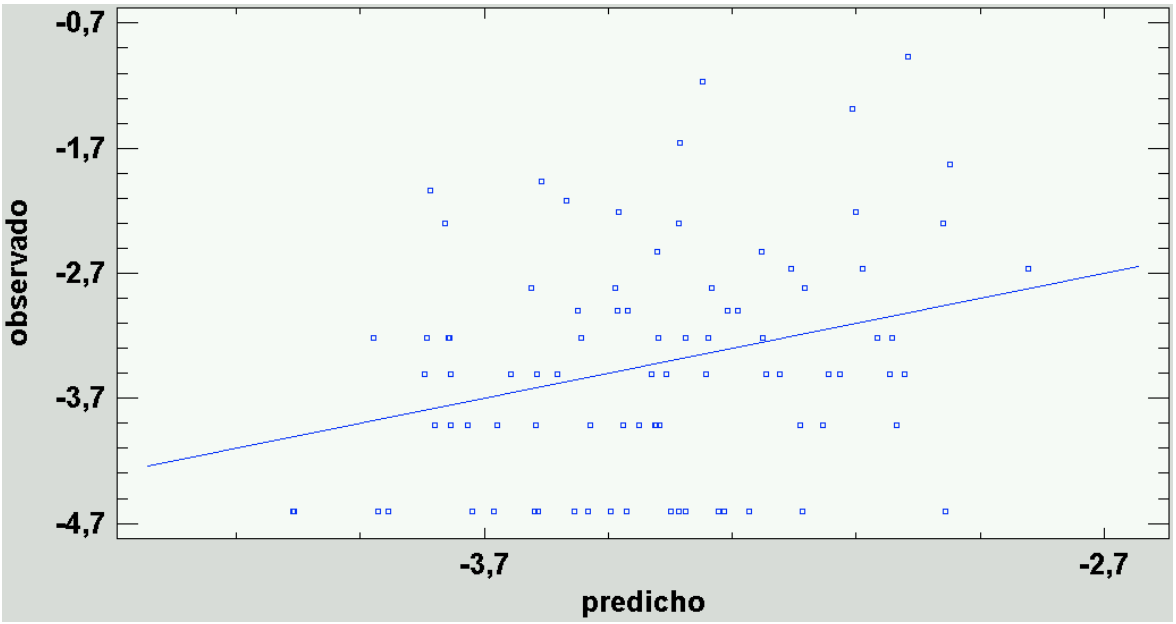


Figura 377. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(adsense_ecpm_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 8,49787% de la variabilidad de log(adsense_ecpm_mean), mientras que el R-Cuadrado ajustado indica un 7,42137%. Este número tan bajo se puede observar en el gráfico anterior, ya que hay una gran cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Duración de la visita (promedio)

Para tratar de predecir el valor de la duración de la visita (promedio) es necesario realizar la regresión múltiple con la variable dependiente log(avgtimeonpage_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 168

Tráileres: Valor-P de las variables de la regresión múltiple de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	2,48977	0,3574
log(terms_ini_num_tweets)	-0,0733675	0,3935
log(terms_ini_retweet_count_mean)	0,0173367	0,8113
log(terms_ini_favorite_count_mean)	-0,440724	0,0229
log(terms_ini_followers_talking_rate)	-0,290539	0,0879
log(terms_ini_user_num_followers_mean)	-0,0172465	0,8983
log(terms_ini_user_num_tweets_mean)	0,165246	0,5937
log(terms_ini_url_inclusion_rate)	-0,379316	0,2419
Modelo		0,135

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 169

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(avgtimeonpage_mean)

Variable	Estimación	Valor-P
Constante	3,8351	0
log(terms_ini_favorite_count_mean)	-0,442721	0,0153
log(terms_ini_followers_talking_rate)	-0,345442	0,006

Modelo	0,0097

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

avgtimeonpage_mean = exp(3,8351 - 0,442721 * log(terms_ini_favorite_count_mean) - 0,345442 * log(terms_ini_followers_talking_rate))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

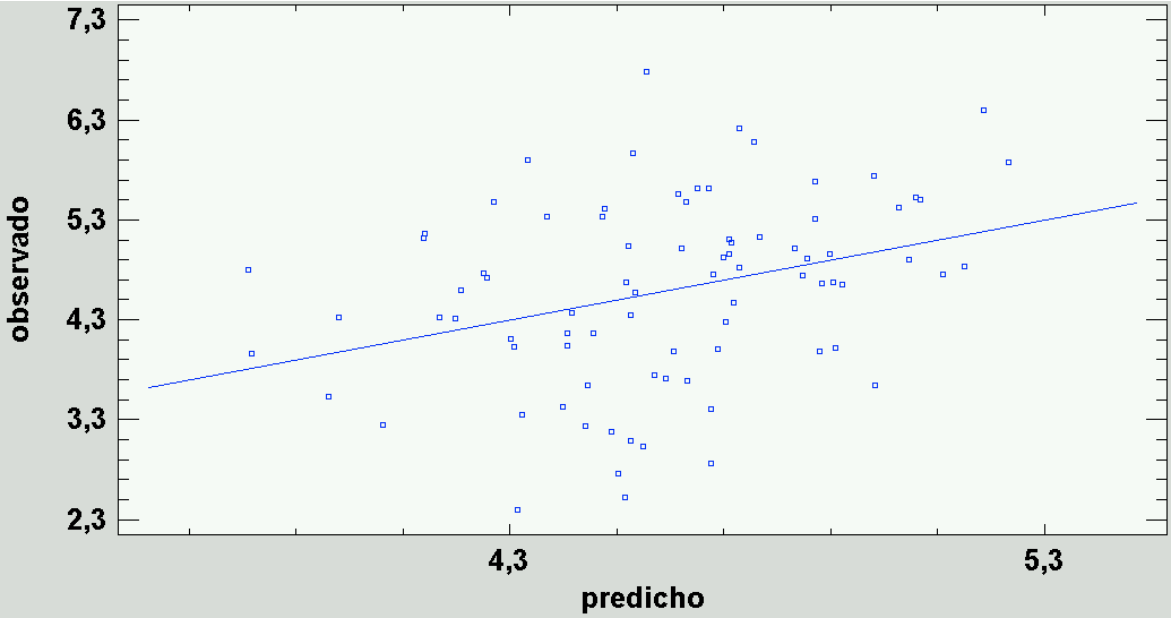


Figura 378. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(avgtimeonpage_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 10,9437% de la variabilidad de log(avgtimeonpage_mean), mientras que el R-Cuadrado ajustado indica un 8,71729%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

d) Páginas vistas por sesión (promedio)

Para tratar de predecir el valor de las páginas vistas por sesión (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(pageviewspersession_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 170

Tráileres: Valor-P de las variables de la regresión múltiple de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	3,21743	0,0172
log(terms_ini_num_tweets)	-0,0064729	0,8779
log(terms_ini_retweet_count_mean)	0,00569641	0,8731
log(terms_ini_favorite_count_mean)	-0,0678105	0,4691
log(terms_ini_followers_talking_rate)	0,167686	0,0458
log(terms_ini_user_num_followers_mean)	-0,0451182	0,4969
log(terms_ini_user_num_tweets_mean)	-0,171732	0,2606
log(terms_ini_url_inclusion_rate)	-0,0470767	0,7665
Modelo		0,0178

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 171

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(pageviewspersession_mean)

Variable	Estimación	Valor-P
Constante	3,828	0,0011
log(terms_ini_followers_talking_rate)	0,175594	0,0015

log(terms_ini_user_num_tweets_mean)	-0,271708	0,0113
Modelo		0,0003

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

pageviewspersession_mean = exp(3,828 + 0,175594 * log(terms_ini_followers_talking_rate) - 0,271708 * log(terms_ini_user_num_tweets_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

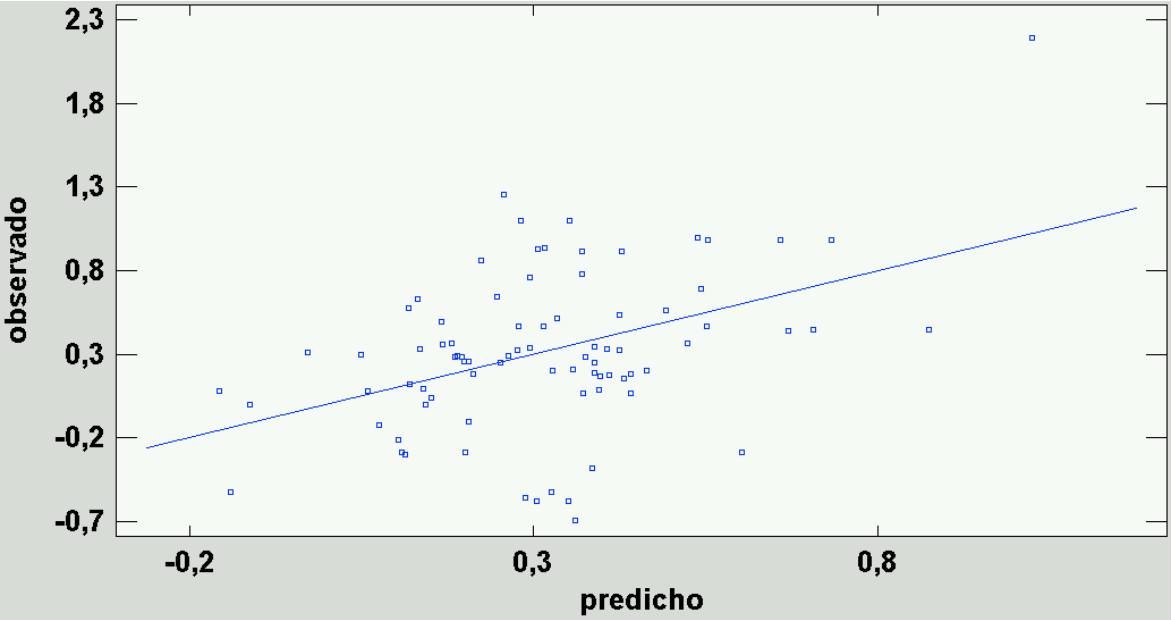


Figura 379. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(pageviewspersession_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 18,3504% de la variabilidad de log(pageviewspersession_mean), mientras que el R-Cuadrado ajustado indica un 16,3091%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

e) Nº de retuits en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de retuits en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 172

Tráileres: Valor-P de las variables de la regresión múltiple de log(retweet_count_mean)

Variable	Estimación	Valor-P
Constante	0,753418	0,5199
log(terms_ini_num_tweets)	0,0792725	0,0401
log(terms_ini_retweet_count_mean)	-0,02736	0,3830
log(terms_ini_favorite_count_mean)	-0,123216	0,1572
log(terms_ini_followers_talking_rate)	0,020401	0,7899
log(terms_ini_user_num_followers_mean)	0,0327939	0,5892
log(terms_ini_user_num_tweets_mean)	-0,12194	0,3713
log(terms_ini_url_inclusion_rate)	0,133996	0,3421
Modelo		0,3878

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 173

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(retweet_count_mean)

Variable	Estimación	Valor-P
Constante	-0,278755	0,0878
log(terms_ini_num_tweets)	0,0582021	0,0246
log(terms_ini_favorite_count_mean)	-0,209238	0,0034
Modelo		0,0075

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

retweet_count_mean = exp(-0,278755 + 0,0582021 * log(terms_ini_num_tweets) - 0,209238 * log(terms_ini_favorite_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

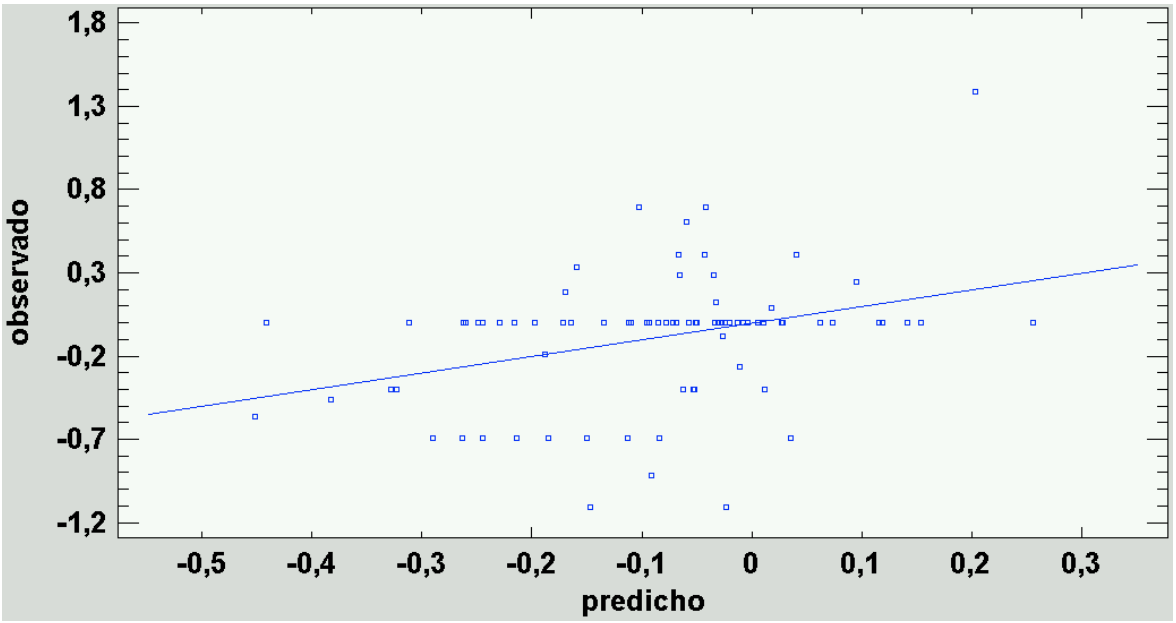


Figura 380. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 11,7912% de la variabilidad de log(retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 9,5294%. Este

número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

f) Nº de favoritos en la cuenta del medio (promedio)

Para tratar de predecir el valor de el número de favoritos en la cuenta del medio (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(favorite_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 174

Tráileres: Valor-P de las variables de la regresión múltiple de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	-1,17035	0,4046
log(terms_ini_num_tweets)	0,119998	0,0089
log(terms_ini_retweet_count_mean)	-0,0252108	0,4915
log(terms_ini_favorite_count_mean)	-0,146945	0,1535
log(terms_ini_followers_talking_rate)	-0,0447799	0,6226
log(terms_ini_user_num_followers_mean)	0,0230955	0,748
log(terms_ini_user_num_tweets_mean)	0,0410282	0,7989
log(terms_ini_url_inclusion_rate)	0,0447851	0,7874
Modelo		0,0969

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 175

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(favorite_count_mean)

Variable	Estimación	Valor-P
Constante	-0,506619	0,0186
log(terms_ini_num_tweets)	0,138805	0,0001
log(terms_ini_favorite_count_mean)	-0,180174	0,0486
Modelo		0,0004

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

favorite_count_mean = exp(-0,506619 + 0,138805 * log(terms_ini_num_tweets) - 0,180174 * log(terms_ini_favorite_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

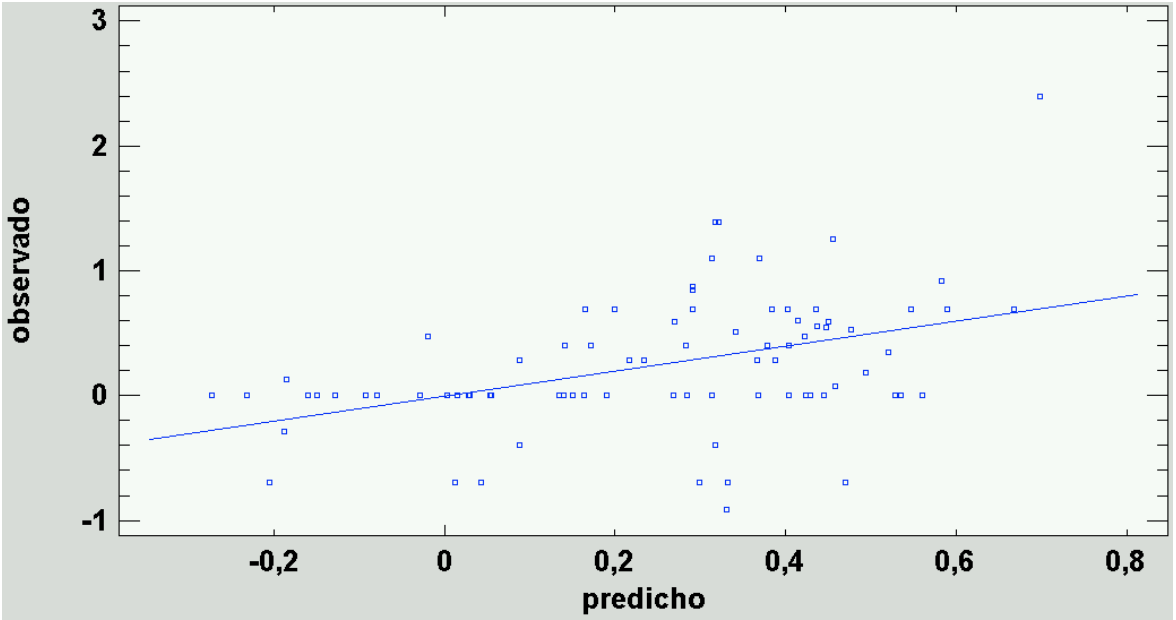


Figura 381. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(favorite_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 18,2644% de la variabilidad de log(favorite_count_mean), mientras que el R-Cuadrado ajustado indica un 16,1687%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

g) Nº de retuits de la tendencia 14 días después (promedio)

Para tratar de predecir el valor de el número de retuits de la tendencia 14 días después (promedio), es necesario realizar la regresión múltiple con la variable dependiente log(terms_end_retweet_count_mean). Se utilizan como variables independientes las disponibles para la predicción, también transformadas logarítmicamente.

Tabla 176

Tráileres: Valor-P de las variables de la regresión múltiple de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	3,14685	0,5013
log(terms_ini_num_tweets)	0,575906	0,0019
log(terms_ini_retweet_count_mean)	0,258214	0,1496
log(terms_ini_favorite_count_mean)	-0,593172	0,0798
log(terms_ini_followers_talking_rate)	-0,33031	0,2442
log(terms_ini_user_num_followers_mean)	-0,00583821	0,9792
log(terms_ini_user_num_tweets_mean)	-0,553056	0,2966
log(terms_ini_url_inclusion_rate)	0,085115	0,8852
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 177

Tráileres: Valor-P de las variables de la regresión múltiple simplificada de log(terms_end_retweet_count_mean)

Variable	Estimación	Valor-P
Constante	-1,86484	0,0114
log(terms_ini_num_tweets)	0,485366	0,0011
log(terms_ini_retweet_count_mean)	0,344228	0,0077
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_mean = exp(-1,86484 + 0,485366 * log(terms_ini_num_tweets) + 0,344228 * log(terms_ini_retweet_count_mean))

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

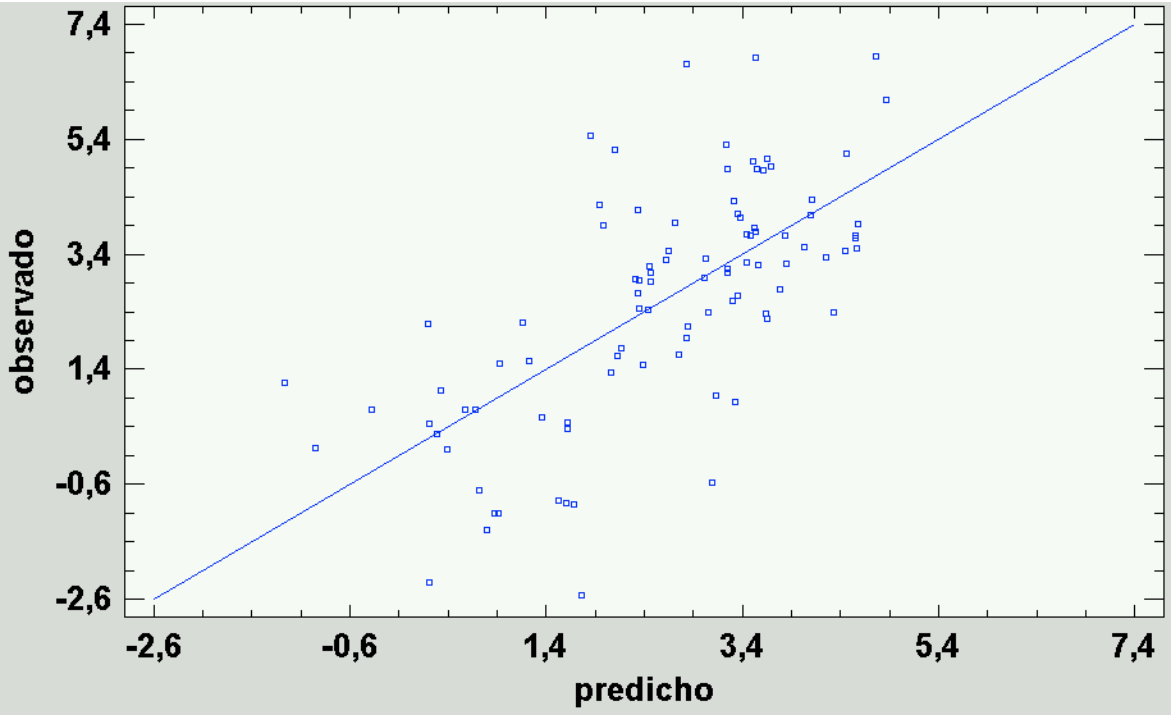


Figura 382. Tráileres: Gráfico de la relación entre la función de predicción de la regresión lineal múltiple y el valor observado de log(terms_end_retweet_count_mean) en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 45,2254% de la variabilidad de log(terms_end_retweet_count_mean), mientras que el R-Cuadrado ajustado indica un 44,0215%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada, pero en general los datos siguen más la línea que en las regresiones lineales múltiples anteriores.

h) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones lineales múltiples de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 178

Tráileres: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones lineales múltiples

Variables de predicción	Variables de éxito
log(terms_ini_num_tweets)	log(uniquepageviews_total)
log(terms_ini_retweet_count_mean)	log(adsense_ecpm_mean)
log(terms_ini_favorite_count_mean)	log(avgtimeonpage_mean)
log(terms_ini_followers_talking_rate)	log(pageviewspersession_mean)
log(terms_ini_user_num_followers_mean)	log(retweet_count_mean)
log(terms_ini_user_num_tweets_mean)	log(favorite_count_mean)
log(terms_ini_url_inclusion_rate)	log(terms_end_retweet_count_mean)

Se puede observar en la

Tabla *178* que todas las variables de predicción salvo log(terms_ini_url_inclusion_rate) participan en alguna de las ecuaciones de predicción, por lo que todas menos la previamente mencionada son necesarias para las variables de éxito elegidas y que se pueden estudiar.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits de la tendencia explica parte de los datos de páginas vistas únicas.
- El promedio de seguidores de los usuarios que participan en la tendencia explica pate de los datos de promedio de eCPM de los anuncios de Google AdSense.
- La ratio de seguidores del medio que hablan de la tendencia explica parte de los datos de la duración de la visita.
- El promedio de favoritos y la ratio de seguidores del medio que hablan de la tendencia explican parte de los datos de la media de páginas vistas por sesión.
- El número de tuits y el promedio de favoritos de la tendencia explican parte de los datos del promedio de retuits en la cuenta del medio.
- El número de tuits y el promedio de favoritos de la tendencia explican parte de los datos del promedio de favoritos en la cuenta del medio.
- El número de tuits y el promedio de retuits de la tendencia explican en parte el promedio de retuits 14 días después.

6.1.4.4. Regresión binomial negativa o de Poisson

A continuación, se tratará de predecir todas las variables de éxito que sean de conteo (enteros y sin números negativos) a partir de todas las variables de predicción que sean independientes entre sí según la regresión binomial negativa o la regresión de Poisson.

a) Filtro de alta correlación (colinealidad)

Las variables que aporten información para tratar de realizar la regresión deben ser independientes, motivo por el cual es necesario hacer un filtro de alta correlación de manera que se asegure que todas aportan información diferente.

El objetivo es detectar y evitar correlaciones fuertes (0,7 o más) entre las variables de predicción para así evitar la colinealidad. Para ello, se ha realizado un análisis multivariado de todas las variables de predicción, obteniendo la siguiente matriz de correlaciones Pearson:

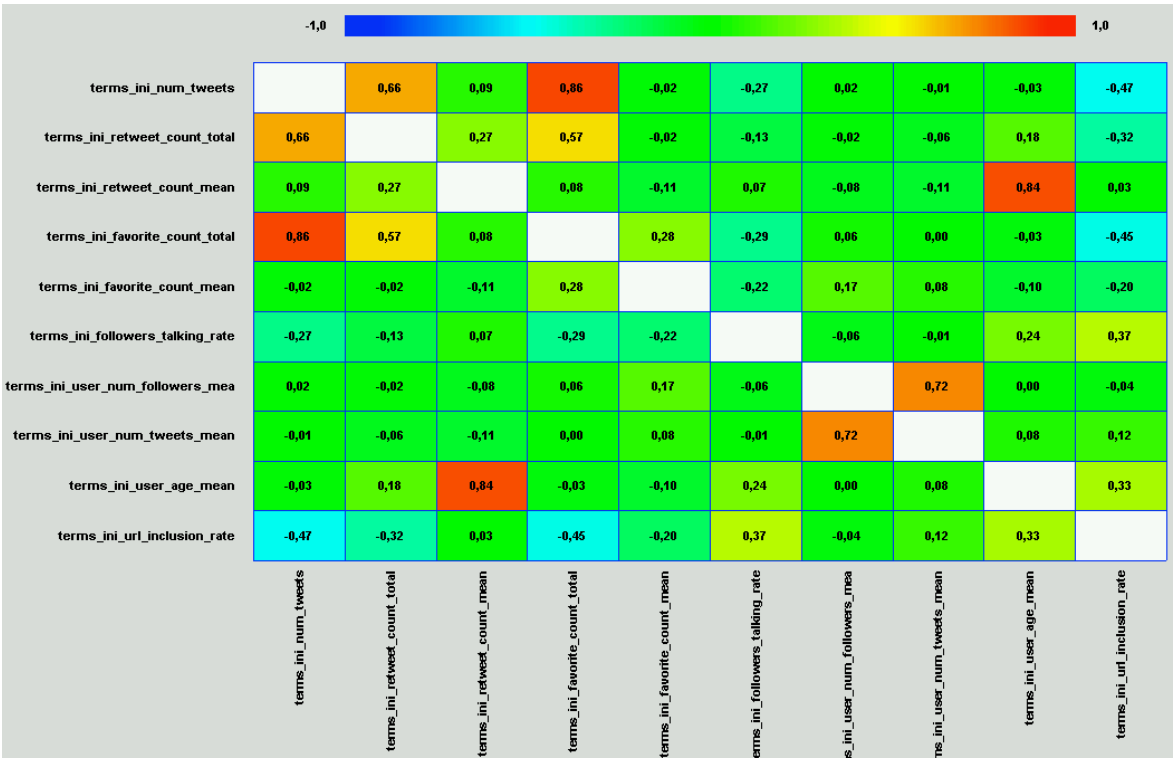


Figura 383. Tráileres: Matriz de correlaciones Pearson entre las variables de predicción

Al hacerlo, se han obtenido las siguientes conclusiones:

- terms_ini_num_tweets y terms_ini_favorite_count_total tienen un coeficiente de correlación de 0,8567 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- tems_ini_retweet_count_mean y terms_ini_user_age_mean tienen un coeficiente de correlación de 0,8404 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.
- tems_ini_user_num_followers_mean y terms_ini_user_num_tweets_mean tienen un coeficiente de correlación de 0,719 y un valor-P cercano a 0, por lo que sí existe una relación estadísticamente significativa (igual o mayor a 0,7) entre ambas variables con un nivel de confianza del 95% o más.

Se elige terms_ini_num_tweets, terms_ini_user_age_mean y tems_ini_user_num_followers_mean por tener un sesgo y una curtosis estandarizados menores, como se puede comprobar en el anexo 6.1.4.2.

La tabla de variables quedaría como sigue:

Tabla 179

Tráileres: Lista de variables de predicción y de éxito para la regresión binomial negativa o de Poisson tras el estudio de correlación de las variables de predicción

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

La lista de variables de predicción queda, por tanto, limitada a: el número total de tuits, el número total de retuits, el promedio de retuits, el promedio de favoritos, la ratio de seguidores del medio que participan, el promedio de seguidores de los usuarios que participan, el promedio de edad en días de la cuenta de los usuarios que participan y la ratio de inclusión de URL en los tuits, todos ellos provenientes de la tendencia el día de la publicación del artículo.

b) Páginas únicas (total)

Para tratar de predecir el valor de las páginas únicas (total), es necesario realizar la regresión con la variable dependiente uniquepageviews_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable uniquepageviews_total, el Chi-cuadrado calculado es 491.184 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 180

Tráileres: Valor-P de las variables de la regresión binomial negativa de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,29272	0
terms_ini_num_tweets	0,000323625	0
terms_ini_retweet_count_total	-0,000000412157	0
terms_ini_favorite_count_mean	-0,0706312	0
terms_ini_followers_talking_rate	-1,47955	0
terms_ini_user_num_followers_mean	-0,00000701167	0
terms_ini_user_age_mean	-0,0000412947	0
terms_ini_url_inclusion_rate	0,181942	1
Modelo		0

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 181

Tráileres: Valor-P de las variables de la regresión binomial negativa simplificada de uniquepageviews_total

Variable	Estimación	Valor-P
Constante	3,35752	0
terms_ini_num_tweets	0,000319464	0

terms_ini_retweet_count_total	-0,000000419462	0
terms_ini_favorite_count_mean	-0,0741463	0
terms_ini_followers_talking_rate	-1,40201	0
terms_ini_user_num_followers_mean	-0,00000698551	0
terms_ini_user_age_mean	-0,0000287853	0,002
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

uniquepageviews_total = exp(3,35752 + 0,000319464 * terms_ini_num_tweets - 0,000000419462 * terms_ini_retweet_count_total - 0,0741463 * terms_ini_favorite_count_mean - 1,40201 * terms_ini_followers_talking_rate - 0,00000698551 * terms_ini_user_num_followers_mean - 0,0000287853 * terms_ini_user_age_mean)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

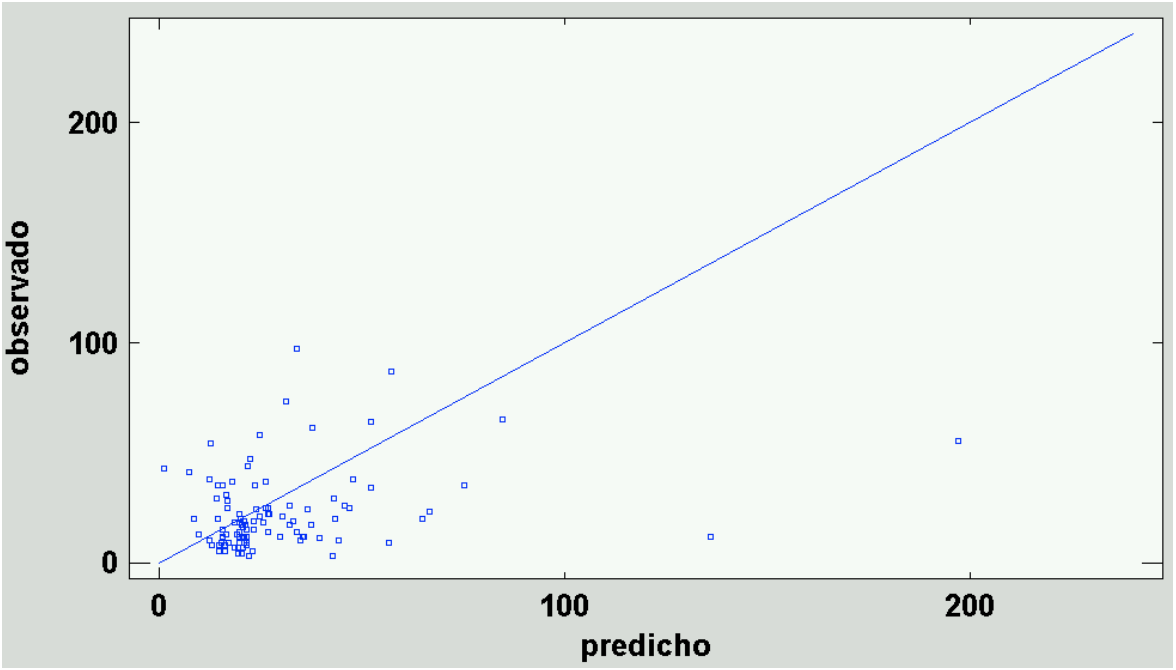


Figura 384. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de uniquepageviews_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 3,83227% de la variabilidad de uniquepageviews_total, mientras que el R-Cuadrado ajustado indica un 3,22807%. Este número bajo se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan de la línea que muestra una predicción acertada.

c) Nº de tuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de tuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_num_tweets. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_num_tweets, el Chi-cuadrado calculado es 343.742.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 182

Tráileres: Valor-P de las variables de la regresión binomial negativa de terms_end_num_tweets

Variable	Estimación	Valor-P

Constante	17,9566	0
terms_ini_num_tweets	0,000340696	0,001
terms_ini_retweet_count_total	-0,0000000163867	1
terms_ini_favorite_count_mean	-0,076762	1
terms_ini_followers_talking_rate	-0,508151	0,9706
terms_ini_user_num_followers_mean	-0,00000756398	0,6968
terms_ini_user_age_mean	-0,000152302	0,6145
terms_ini_url_inclusion_rate	0,740764	1
Modelo		0,0048

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 183

Tráileres: Valor-P de las variables de la regresión binomial negativa simplificada de t erms_end_num_tweets

Variable	Estimación	Valor-P
Constante	17,5667	0
terms_ini_num_tweets	0,000318724	0
Modelo		0

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_num_tweets = exp(17,5667 + 0,000318724 * terms_ini_num_tweets)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

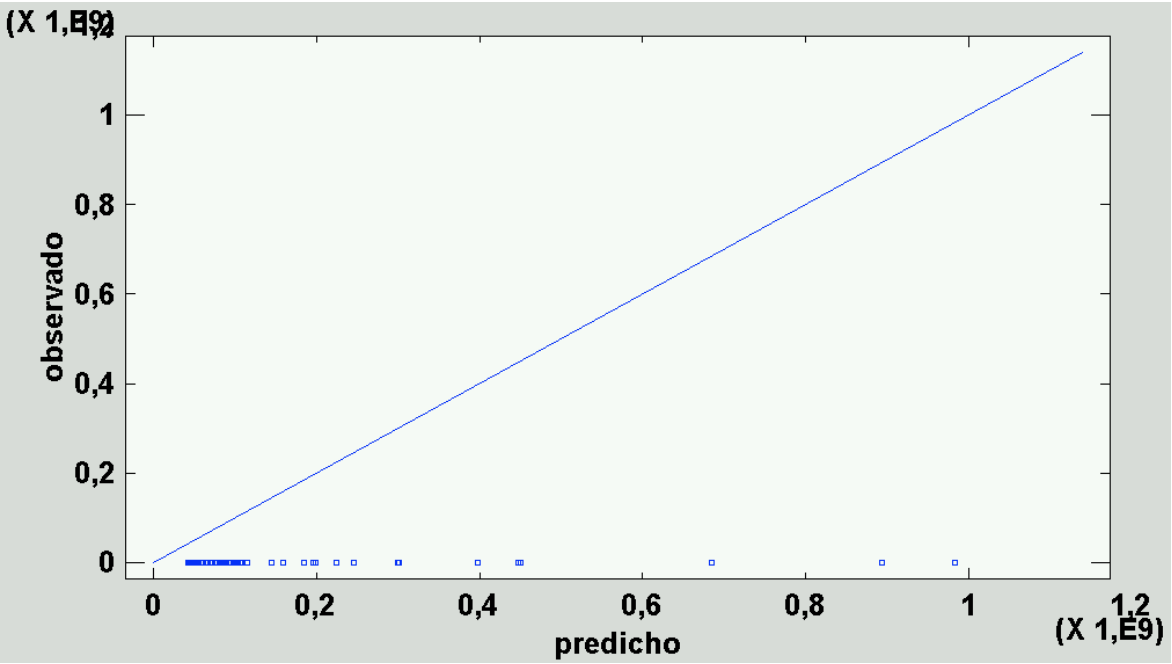


Figura 385. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_num_tweets en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 9,66905% de la variabilidad de terms_end_num_tweets, mientras que el R-Cuadrado ajustado indica un 7,6614%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se acercan a la línea que muestra una predicción acertada.

d) Nº de retuits de la tendencia 14 días después (total)

Para tratar de predecir el valor del número de retuits de la tendencia 14 días después (total), es necesario realizar la regresión con la variable dependiente terms_end_retweet_count_total. Se utilizan como variables independientes las disponibles para la predicción.

Para elegir entre la regresión binomial negativa o la de Poisson, se realiza la prueba de dispersión estadística chi-cuadrado de Pearson, que toma el valor de uno si la varianza es

igual a la media. De esta manera, si la dispersión estadística chi-cuadrado de Pearson está cercana a uno, se utiliza la regresión de Poisson, mientras que, si no es así, se utiliza la binomial negativa. (Meyer, 2018). En el caso de la variable terms_end_retweet_count_total, el Chi-cuadrado calculado es 90.791.600.000.000 con un p-valor cercano a 0, por lo que la binomial negativa es más adecuada con un 95% de confianza.

Tabla 184

Tráileres: Valor-P de las variables de la regresión binomial negativa de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	28,1328	0
terms_ini_num_tweets	0,000371373	1
terms_ini_retweet_count_total	0,000000922259	1
terms_ini_favorite_count_mean	0,60501	0
terms_ini_followers_talking_rate	-61,5525	1
terms_ini_user_num_followers_mean	0,0000297161	1
terms_ini_user_age_mean	-0,00129054	0
terms_ini_url_inclusion_rate	10,7562	1
Modelo		1

El modelo se puede simplificar, ya que incluye variables con Valor-P mayor o igual que 0,05, lo cual indica que esos términos no son estadísticamente significativos con un nivel de confianza del 95% o mayor.

Se realiza, por tanto, un proceso de simplificación del modelo eliminando la variable con un p-valor mayor hasta que todas las presentes en el modelo tengan un p-valor menor que 0,05. Al final, el modelo queda de la siguiente manera:

Tabla 185

Tráileres: Valor-P de las variables de la regresión binomial negativa simplificada de terms_end_retweet_count_total

Variable	Estimación	Valor-P
Constante	31,824	0
terms_ini_retweet_count_total	0,0000002995	0,0002
terms_ini_favorite_count_mean	0,895295	0
terms_ini_followers_talking_rate	-14,7636	0,042
terms_ini_user_num_followers_mean	-0,000113119	0
Modelo		0,0007

El valor-P en la tabla ANOVA del modelo es menor que 0,05, por lo que existe una relación estadísticamente significativa entre las variables con un nivel de confianza del 95%.

La fórmula del modelo ajustado es:

terms_end_retweet_count_total = exp(31,824 + 0,0000002995 * terms_ini_retweet_count_total + 0,895295 * terms_ini_favorite_count_mean - 14,7636 * terms_ini_followers_talking_rate - 0,000113119 * terms_ini_user_num_followers_mean)

La relación entre la función de predicción anterior y el valor observado en esta fase se puede ver en el siguiente gráfico:

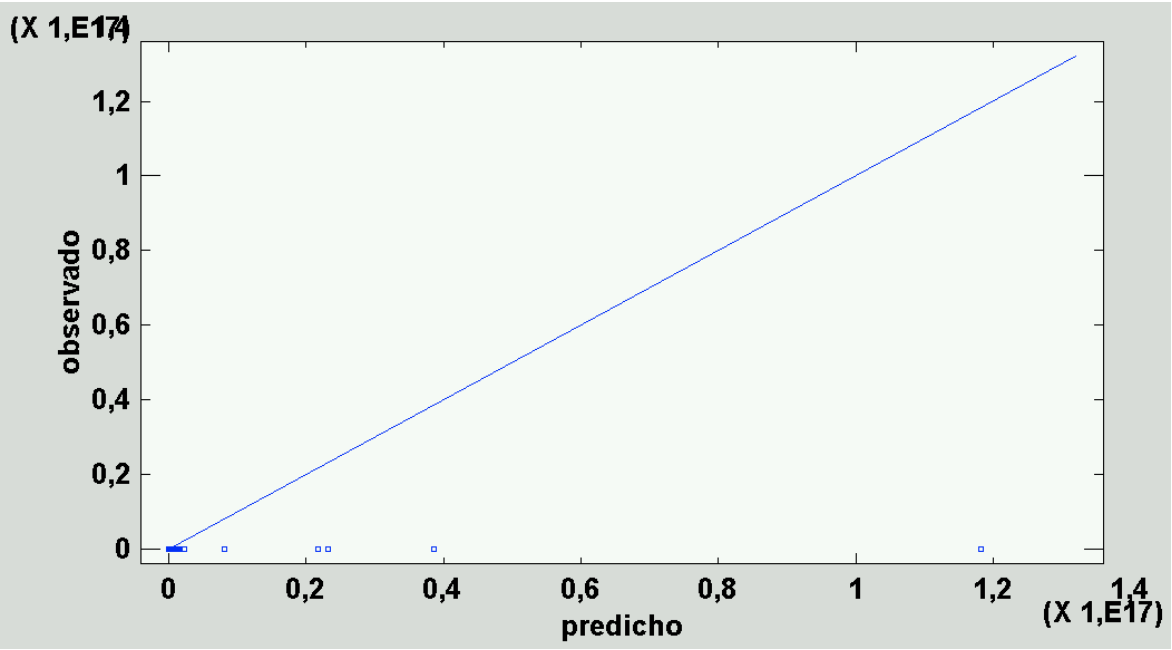


Figura 386. Tráileres: Gráfico de la relación entre la función de predicción según la regresión binomial negativa y el valor observado de terms_end_retweet_count_total en la fase 1

Según el R-Cuadrado, el modelo así ajustado explica el 9,81024% de la variabilidad de terms_end_retweet_count_total, mientras que el R-Cuadrado ajustado indica un 4,71226%. Este número se puede observar en el gráfico anterior, ya que hay una cierta cantidad de valores que se alejan a la línea que muestra una predicción acertada.

e) Resumen de relaciones entre variables de predicción y de éxito

Al acabar la extracción de las ecuaciones de predicción de las regresiones binomiales negativas de cada variable de éxito, se ha obtenido una serie de relaciones que pueden resultar útiles a la hora de dibujar una imagen global de las influencias entre las variables.

Tabla 186

Tráileres: Lista de variables de predicción y de éxito y las relaciones entre estas en las regresiones binomiales negativas

Variables de predicción	Variables de éxito
terms_ini_num_tweets	uniquepageviews_total
terms_ini_retweet_count_total	terms_end_num_tweets
terms_ini_favorite_count_mean	terms_end_retweet_count_total
terms_ini_followers_talking_rate	
terms_ini_user_num_followers_mean	
terms_ini_user_age_mean	
terms_ini_url_inclusion_rate	

Se puede observar en la Tabla *186* que casi todas las variables de predicción participan en alguna de las ecuaciones de predicción, por lo que casi todas son necesarias para las variables de éxito elegidas y que se pueden estudiar. La única que no participa es la ratio de inclusión de URL en los tuits.

Con ello, se pueden extraer las siguientes conclusiones:

- El número de tuits, el número de retuits, el promedio de favoritos, la ratio de seguidores de la cuenta del medio que participan, el número de seguidores de los usuarios que participan y el promedio de la edad en días de los usuarios que participan en la tendencia explican parte de los datos de páginas vistas únicas.
- El número de tuits explica parte de los datos del número de tuits 14 días después.
- El número de retuits, el promedio de favoritos, la ratio de seguidores de la cuenta del medio que participan y el número de seguidores de los usuarios que participan explican en parte el número de retuits 14 días después.

6.2. API de informes de Google Analytics v4

A continuación, se puede ver la documentación oficial sobre la API de informes de Google Analytics v4.

La API de informes de Google Analytics v4 permite el acceso a los datos de los informes de la herramienta de Google Analytics, con URL relativas a analyticsreporting.googleapis.com. La URL de detección de esta API es https://analyticsreporting.googleapis.com/$discovery/rest?version=v4 (Google Developers, s.f.).

El método principal de esta API es batchGet, que devuelve los datos de Analytics con la solicitud HTTP (Google Developers, s.f.):

POST https://analyticsreporting.googleapis.com/v4/reports:batchGet

6.2.1. Cuerpo de la solicitud

El cuerpo de la solicitud contiene datos que presentan la siguiente estructura:

Representación JSON

```
{
  "reportRequests": [
    {
      object(ReportRequest)
    }
  ],
}
```

Nombre del campo	Tipo	Descripción
reportRequests[]	object(Report Request)	Solicitudes. Cada solicitud tiene una respuesta distinta. Puede realizarse un máximo de cinco solicitudes y todas deben tener los mismos

	campos dateRanges, viewId, segments, samplingLevel y cohortGroup.

6.2.2. Cuerpo de la respuesta

Si la solicitud se realiza correctamente, el cuerpo de la respuesta proporciona datos con la siguiente estructura:

Clase de respuesta principal que incluye los informes de la llamada batchGet de la API de informes.

Representación JSON

```
{
  "reports": [
    {
      object(Report)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción
reports[]	object(Report)	Respuestas correspondientes a cada solicitud.

6.2.3. Autorización

Requiere uno de los siguientes alcances de OAuth:

- https://www.googleapis.com/auth/analytics.readonly
- https://www.googleapis.com/auth/analytics

6.2.4. ReportRequest

Clase de solicitud principal que especifica la solicitud de la API de informes.

Representación JSON

```
{
  "viewId": string,
  "dateRanges": [
    {
      object(DateRange)
    }
  ],
  "samplingLevel": enum(Sampling),
  "dimensions": [
    {
      object(Dimension)
    }
  ],
  "dimensionFilterClauses": [
    {
      object(DimensionFilterClause)
    }
  ],
  "metrics": [
    {
      object(Metric)
    }
  ],
  "metricFilterClauses": [
    {
      object(MetricFilterClause)
    }
  ],
  "filtersExpression": string,
  "orderBys": [
    {
      object(OrderBy)
    }
  ],
  "segments": [
    {
      object(Segment)
    }
  ],
```

```
  "pivots": [
    {
      object(Pivot)
    }
  ],
  "cohortGroup": {
    object(CohortGroup)
  },
  "pageToken": string,
  "pageSize": number,
  "includeEmptyRows": boolean,
  "hideTotals": boolean,
  "hideValueRanges": boolean,

}
```

Nombre del campo	Tipo	Descripción
viewId	string	ID de vista de Analytics a partir del cual se obtienen los datos. Todas las clases ReportRequest de un método batchGet deben tener el mismo valor de viewId.
dateRanges[]	object(DateRange)	Periodos de la solicitud. La solicitud puede tener un máximo de dos periodos. La respuesta incluirá un conjunto de valores de métricas para cada combinación de dimensiones de cada periodo indicado en la solicitud. De este modo, si hay dos periodos, se incluirán dos conjuntos de valores de métricas, uno para el periodo original y otro para el segundo periodo. El campo reportRequest.dateRanges no debe especificarse para las solicitudes de cohortes o del valor del ciclo de vida del cliente. Si no se indica ningún periodo, se utilizará el predeterminado (startDate:

		fecha actual - 7 días, endDate: fecha actual - 1 día). Todas las clases ReportRequest de un método batchGet deben tener la misma definición de dateRanges.
samplingLevel	enum(Sampling)	Tamaño de muestra deseado del informe. Si no se especifica ningún valor en el campo samplingLevel, se utilizará el nivel de muestra predeterminado (DEFAULT). Todas las clases ReportRequest de un método batchGet deben tener la misma definición de samplingLevel. Consulta los detalles en la guía de programadores.
dimensions[]	object(Dimension)	Dimensiones solicitadas. Las solicitudes pueden incluir hasta 7 dimensiones.
dimensionFilterClauses[]	object(DimensionFilterClause)	Cláusulas del filtro de dimensiones para filtrar los valores de dimensión. Se combinan de forma lógica con el operador AND. Ten en cuenta que el filtrado se realiza antes de añadir dimensiones, de forma que las métricas devueltas representen el total para las dimensiones relevantes únicamente.
metrics[]	object(Metric)	Métricas solicitadas. En las solicitudes hay que especificar entre 1 y 10 métricas.

metricFilterClauses[]	object(MetricFilterClause)	Cláusulas del filtro de métricas. Se combinan de forma lógica con el operador AND. Los filtros de métricas no se aplican al periodo comparativo, sino solo al primer periodo. Ten en cuenta que el filtrado de las métricas se realiza después de añadir las métricas.
filterExpression	string	Filtros de dimensiones o métricas que acotan los datos que devuelve la solicitud. Para usar el campo filtersExpression, debes indicar la dimensión o la métrica por la que quieras filtrar los datos, seguida de la expresión de filtrado. Por ejemplo, la expresión siguiente selecciona la dimensión ga:browser que empieza por Firefox; ga:browser=~^Firefox. Consulta la referencia de filtros para obtener más información sobre los filtros de dimensiones y métricas.
orderBys[]	object(OrderBy)	Clasificación de las filas de los resultados. Para comparar dos filas, se aplican los elementos del objeto siguiente por orden hasta que se encuentra una diferencia. Se aplica el mismo orden de filas a todos los periodos de los resultados.
segments[]	object(Segment)	Segmentación de los datos que devuelve la solicitud. La definición de un segmento permite analizar un subconjunto de la solicitud del segmento. Una solicitud puede incluir un máximo de cuatro segmentos. Todas las clases ReportRequest de un método batchGet deben tener la misma definición de segments. Las solicitudes que contengan segmentos deben incluir la dimensión ga:segment.

pivots[]	object(Pivot)	Definiciones de la tabla dinámica. Las solicitudes pueden tener un máximo de 2 tablas dinámicas.
cohortGroup	object(CohortGroup)	Grupo de cohortes asociado a la solicitud. Si la solicitud contiene un grupo de cohortes, también debe incluir la dimensión ga:cohort. Todas las clases ReportRequest de un método batchGet deben tener la misma definición de cohortGroup.
pageToken	string	Token de continuación para ir a la página siguiente de los resultados. Si se incluye en la solicitud, se obtendrán las filas después de pageToken. El valor de pageToken debe ser el que se ha devuelto en el parámetro nextPageToken de la respuesta a la solicitud reports.batchGet.
pageSize	number	El tamaño de página sirve para la paginación y especifica el número máximo de filas que devuelve la solicitud. El tamaño de página debe ser >= 0. De forma predeterminada, una consulta devuelve 1000 filas. La API de informes centrales de Analytics devuelve un máximo de 100.000 filas por solicitud, independientemente de las que se hayan solicitado. También puede devolver menos filas de las solicitadas, si hay menos segmentos de dimensión de los previstos. Por ejemplo, hay menos de 300 valores posibles para ga:country, por lo que, al segmentar solo por el país, no se pueden obtener más de 300 filas, aunque se haya configurado pageSize en un valor más alto.

includeEmptyRows	boolean	Si se configura en "false", la respuesta no incluye ninguna fila si el valor de todas las métricas obtenidas es igual a cero. El valor predeterminado es "false", lo que significa que se excluyen estas filas.
hideTotals	boolean	Si se configura en "true", se oculta el total de todas las métricas de las filas coincidentes en ambos periodos. El valor predeterminado es "false" y muestra los totales.
hideValueRanges	boolean	Si se configura en "true", se ocultan los valores máximo y mínimo de todas las filas coincidentes. El valor predeterminado es "false" y muestra los intervalos de valores.

6.2.5. DateRange

Conjunto de días contiguos: startDate, startDate + 1 día, …, endDate. Las fechas de inicio y de finalización se indican en el formato de fecha YYYY-MM-DD que establece la norma ISO8601.

Representación JSON

```
{
  "startDate": string,
  "endDate": string,

}
```

Nombre del campo	Tipo	Descripción
startDate	string	Fecha de inicio de la consulta en formato YYYY-MM-DD.
endDate	string	Fecha de finalización de la consulta en formato YYYY-MM-DD.

6.2.6. Muestreo

Valores del nivel de muestreo.

Valor de enumeración	Descripción
SAMPLING_UNSPECIFIED	Si no se especifica ningún valor en el campo samplingLevel, se utiliza el nivel predeterminado (DEFAULT).
DEFAULT	Devuelve una respuesta con un tamaño de muestra que equilibra velocidad y precisión.
SMALL	Devuelve una respuesta rápida con un tamaño de muestra menor.
LARGE	Devuelve una respuesta más exacta con un tamaño de muestra grande, pero puede provocar que la respuesta sea más lenta.

6.2.7. Dimensión

Las dimensiones son atributos de los datos. Por ejemplo, la dimensión ga:city indica la ciudad (como "Madrid" o "Nueva York") desde la que se origina una sesión.

Representación JSON

```
{
  "name": string,
  "histogramBuckets": [
    string
  ],

}
```

Nombre del campo	Tipo	Descripción
name	string	Nombre de la dimensión que se debe obtener. Por ejemplo: ga:browser.

histogramBuckets[]	string (int64 format)	Si el campo no está vacío, colocamos los valores de la dimensión situados en grupos después de la cadena en int64. Los valores de la dimensión que no constituyen una representación en cadena de un valor entero se convierten a cero. El orden de los valores del grupo debe ser ascendente. Cada grupo se cierra por el extremo inferior y se abre por el extremo superior. El "primer" grupo incluye todos los valores inferiores al primer límite y el "último" grupo incluye todos los valores hasta el infinito. Los valores de dimensiones que se encuentran dentro de un grupo se transforman en un nuevo valor de dimensión. Por ejemplo, si uno proporciona una lista de "0, 1, 3, 4, 7", devolvemos los grupos siguientes: ● grupo n.º 1: valores < 0, valor de dimensión "<0" ● grupo n.º 2: valores en [0,1), valor de dimensión "0" ● grupo n.º 3: valores en [1,3), valor de dimensión "1-2" ● grupo n.º 4: valores en [3,4), valor de dimensión "3" ● grupo n.º 5: valores en [4,7), valor de dimensión "4-6" ● grupo n.º 6: valores >= 7, valor de dimensión "7+" NOTA: Si se aplica la mutación de histograma en alguna dimensión, y se utiliza dicha dimensión para ordenar los elementos, es recomendable usar el tipo de orden HISTOGRAM_BUCKET. De lo contrario, los

valores de la dimensión se clasificarán por orden lexicográfico. Por ejemplo, el orden lexicográfico ascendente es:

"<50", "1001+", "121-1000", "50-120"

Y el orden ascendente de HISTOGRAM_BUCKET es:

"<50", "50-120", "121-1000", "1001+"

El cliente debe solicitar explícitamente "orderType": "HISTOGRAM_BUCKET" para una dimensión con mutación de histograma.

6.2.8. DimensionFilterClause

Grupo de filtros de dimensión. Configura el valor del operador de modo que especifique la forma lógica en que se combinan los filtros.

Representación JSON

```
{
  "operator": enum(FilterLogicalOperator),
  "filters": [
    {
      object(DimensionFilter)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción
operator	enum(FilterLogicalOperator)	Operador para combinar varios filtros de dimensión. Si no se especifica ningún valor, se utiliza OR.
filters[]	object(DimensionFilter)	Conjunto repetido de filtros. Se combinan de forma lógica según el operador especificado.

6.2.9. FilterLogicalOperator

Lógica que se utiliza para combinar los filtros.

Valor de enumeración	Descripción
OPERATOR_UNSPECIFIED	Operador no especificado. Se utiliza OR.

OR	Operador lógico OR.
AND	Operador lógico AND.

6.2.10. DimensionFilter

El filtro de dimensión especifica las opciones de filtrado de una dimensión.

Representación JSON

```
{
  "dimensionName": string,
  "not": boolean,
  "operator": enum(Operator),
  "expressions": [
    string
  ],
  "caseSensitive": boolean,

}
```

Nombre del campo	Tipo	Descripción
dimensionName	string	Dimensión por la que se filtran los resultados. Un campo DimensionFilter debe incluir una dimensión.
not	boolean	Operador lógico NOT. Si este operador booleano se configura como "true", los valores de

		dimensión coincidentes se excluyen del informe. El valor predeterminado es "false".
operator	enum(Operator)	Forma en que se relaciona la dimensión con la expresión. El valor predeterminado es REGEXP.
expressions[]	string	Cadenas o expresión regular con las que deben coincidir los resultados. Solo se usa el primer valor de la lista para la comparación, salvo si el operador es IN_LIST, en cuyo caso se usa toda la lista para filtrar las dimensiones como se explica en la descripción del operador IN_LIST.
caseSensitive	boolean	Valor que indica si debe tenerse en cuenta el uso de mayúsculas y minúsculas a la hora de obtener los resultados. El valor predeterminado es "false".

6.2.11. Operador

Distintos tipos de concordancia admitidos.

Valor de enumeración	Descripción
OPERATOR_UNSPECIFIED	Si no se especifica el tipo de concordancia, se utiliza REGEXP.
REGEXP	La expresión de concordancia se trata como una expresión regular. No todos los tipos de concordancia se tratan como expresiones regulares.

BEGINS_WITH	Se obtiene el valor que empieza con la expresión de concordancia especificada.
ENDS_WITH	Se obtiene el valor que finaliza con la expresión de concordancia especificada.
PARTIAL	Concordancia de cadena secundaria.
EXACT	El valor debe coincidir exactamente con la expresión de concordancia.
NUMERIC_EQUAL	Filtros de comparación de enteros. No distinguen entre mayúsculas y minúsculas y la expresión debe ser una cadena que representa un número entero. Condiciones de error: ● Si la expresión no es un valor int64 válido, el cliente obtendrá un error. ● Las dimensiones introducidas que no sean valores int64 válidos no coincidirán nunca con el filtro.
NUMERIC_GREATER_THAN	Comprueba si la dimensión es numéricamente mayor que la expresión de concordancia. Lee la descripción de NUMERIC_EQUALS para conocer las restricciones.
NUMERIC_LESS_THAN	Comprobar si la dimensión es numéricamente menor que la expresión de concordancia. Leer la descripción de NUMERIC_EQUALS para conocer las restricciones.

IN_LIST	Esta opción se usa para especificar un filtro de dimensión cuya expresión puede incluir cualquier valor de una lista de valores determinada. Esto evita tener que evaluar varios filtros de dimensión de concordancia exacta unidos con el operador OR en cada fila de la respuesta. Por ejemplo: expressions: ["A", "B", "C"] Las filas de la respuesta cuyas dimensiones tienen el valor A, B o C coinciden con este DimensionFilter.

6.2.12. Métrica

Las métricas son mediciones cuantitativas. Por ejemplo, la métrica ga:users indica el total de usuarios para el periodo solicitado.

Representación JSON

```
{
  "expression": string,
  "alias": string,
  "formattingType": enum(MetricType),

}
```

Nombre del campo	Tipo	Descripción

expression	string	Expresión de métrica de la solicitud. Una expresión se forma con una o varias métricas y cifras. Operadores admitidos: más (+), menos (-), negación (operación unaria -), dividido entre (/), multiplicado por (*), paréntesis, números cardinales positivos (0-9); puede incluir decimales y un máximo de 1024 caracteres. Ejemplo: ga:totalRefunds/ga:users. En la mayoría de los casos, la expresión de métrica es solo el nombre de una métrica, como ga:users. Si se añaden métricas MetricType mixtas, como CURRENCY + PERCENTAGE, se obtendrán resultados inesperados.
alias	string	Un alias en la expresión de métrica es un nombre alternativo para la expresión. El alias puede usarse para filtrar y ordenar los resultados. Este campo es opcional y resulta útil si la expresión no está formada por una única métrica, sino que es una expresión compleja que no puede usarse para filtrar y ordenar los resultados. El alias también se utiliza en el encabezado de columna de la respuesta.
formattingType	enum(MetricType)	Especifica cómo debe formatearse la expresión de la métrica. Por ejemplo: INTEGER.

6.2.13. MetricType

Tipos de métricas.

Valor de enumeración	Descripción

METRIC_TYPE_UNSPECIFIED	No se ha especificado el tipo de métrica.
INTEGER	El tipo de métrica es un entero.
FLOAT	El tipo de métrica es un número de punto flotante.
CURRENCY	El tipo de métrica es una moneda.
PERCENT	El tipo de métrica es un porcentaje.
TIME	El tipo de métrica es una hora expresada en el formato HH:MM:SS.

6.2.14. MetricFilterClause

Representa un grupo de filtros de métrica. Configura el valor del operador de modo que especifique la forma lógica en que se combinan los filtros.

Representación JSON

```
{
  "operator": enum(FilterLogicalOperator),
  "filters": [
    {
      object(MetricFilter)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción
operator	enum(FilterLogicalOperator)	Operador para combinar varios filtros de métricas. Si no se especifica ningún valor, se utiliza OR.
filters[]	object(MetricFilter)	Conjunto repetido de filtros. Se combinan de forma lógica según el operador especificado.

6.2.15. MetricFilter

MetricFilter especifica el filtro de una métrica.

Representación JSON

```
{
  "metricName": string,
  "not": boolean,
  "operator": enum(Operator),
  "comparisonValue": string,

}
```

Nombre del campo	Tipo	Descripción
metricName	string	Métrica por la que se filtrarán los resultados. Un campo metricFilter debe incluir un nombre de métrica. Este nombre puede ser un alias que se

		haya definido anteriormente como métrica o bien una expresión de métrica.
not	boolean	Operador lógico NOT. Si este operador booleano se configura como "true", los valores de métrica coincidentes se excluyen del informe. El valor predeterminado es "false".
operator	enum(Operator)	Indica si la métrica es igual (EQUAL), inferior (LESS_THAN) o superior (GREATER_THAN) al valor de comparación (comparisonValue). El valor predeterminado es EQUAL. Si el operador es IS_MISSING, se comprueba si no hay ninguna métrica y se ignora el valor de comparación (comparisonValue).
comparisonValue	string	Valor de comparación.

6.2.16. Operador

Distintas opciones de tipos de comparación.

Valor de enumeración	Descripción
OPERATOR_UNSPECIFIED	Si no se especifica el operador, se usa EQUAL.

EQUAL	El valor de la métrica debe ser exactamente igual al valor de comparación.
LESS_THAN	El valor de la métrica debe ser inferior al valor de comparación.
GREATER_THAN	El valor de la métrica debe ser superior al valor de comparación.
IS_MISSING	Comprueba si no hay ninguna métrica. No tiene en cuenta el valor de comparación (comparisonValue).

6.2.17. OrderBy

Especifica las opciones para ordenar los resultados.

Representación JSON

```
{
  "fieldName": string,
  "orderType": enum(OrderType),
  "sortOrder": enum(SortOrder),

}
```

Nombre del campo	Tipo	Descripción
fieldName	string	Campo por el que se ordenarán los resultados. El orden predeterminado es ascendente. Ejemplo: ga:browser. Recuerda que solo puedes

		especificar un campo. Por ejemplo, ga:browser, ga:city no es válido.
orderType	enum(OrderType)	Tipo de orden. El tipo de orden predeterminado es VALUE.
sortOrder	enum(SortOrder)	Orden según el cual se clasificarán los resultados.

6.2.18. OrderType

El objeto OrderType controla cómo se determina el orden de clasificación.

Valor de enumeración	Descripción
ORDER_TYPE_UNSPECIFIED	Si no se especifica el tipo de orden, la clasificación se basará en el valor.
VALUE	El orden de clasificación se basa en el valor de la columna seleccionada; solo se tiene en cuenta el primer periodo.
DELTA	El orden de clasificación se basa en la diferencia de valores de la columna seleccionada entre los dos primeros periodos. Solo se puede usar si hay exactamente dos periodos.
SMART	El orden de clasificación se basa en el valor ponderado de la columna seleccionada. Si la columna tiene el formato n/d, el valor ponderado de esta relación será (n + totals.n)/(d + totals.d). Solo se puede usar con las métricas que representan relaciones.

HISTOGRAM_BUCKET	El tipo de orden de histograma solo se puede aplicar a las columnas de dimensiones que no tienen grupos de histograma vacíos.
DIMENSION_AS_INTEGER	Si las dimensiones son números de longitud fija, puede usarse el orden normal. DIMENSION_AS_INTEGER puede usarse si las dimensions son números de longitud variable.

6.2.19. SortOrder

Orden de clasificación.

Valor de enumeración	Descripción
SORT_ORDER_UNSPECIFIED	Si no se especifica el orden de clasificación, se utiliza el orden predeterminado ascendente.
ASCENDING	Orden ascendente. El campo se ordena en sentido ascendente.
DESCENDING	Orden descendente. El campo se ordena en sentido descendente.

6.2.20. Segmento

Definición del segmento en los informes que deban segmentarse. Un segmento es un subconjunto de datos de Analytics. Por ejemplo, de entre todos los usuarios que tienes, un segmento podría estar formado por usuarios de un país o una ciudad concretos.

Representación JSON

```
{

  // Union field dynamicOrById can be only one of the following:
  "dynamicSegment": {
    object(DynamicSegment)
  },
  "segmentId": string,
  // End of list of possible types for union field dynamicOrById.

}
```

Nombre del campo	Tipo	Descripción
Campo de unión dynamicOrById. El segmento puede definirse dinámicamente con DynamicSegment o con un ID de un segmento integrado o personalizado. El campo dynamicOrById solo puede ser uno de los siguientes:		
dynamicSegment	object(DynamicSegment)	Definición de segmento dinámico en la solicitud.
segmentId	string	ID de un segmento integrado o personalizado. Por ejemplo: gaid::-3.

6.2.21. DynamicSegment

Definición de segmento dinámico para definir el segmento de la solicitud. Un segmento puede seleccionar usuarios, sesiones, o ambos elementos.

Representación JSON

```
{
  "name": string,
  "userSegment": {
    object(SegmentDefinition)
  },
  "sessionSegment": {
    object(SegmentDefinition)
  },

}
```

Nombre del campo	Tipo	Descripción
name	string	Nombre del segmento dinámico.
userSegment	object(SegmentDefinition)	Segmento de usuario para seleccionar a los usuarios que se incluirán en el segmento.
sessionSegment	object(SegmentDefinition)	Segmento de sesión para seleccionar las sesiones que se incluirán en el segmento.

6.2.22. SegmentDefinition

SegmentDefinition define el segmento como un conjunto de SegmentFilters combinados mediante el operador lógico AND.

Representación JSON

```
{
  "segmentFilters": [
    {
      object(SegmentFilter)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción
segmentFilters []	object(SegmentFilter)	Un segmento se define mediante un conjunto de filtros de segmento que se combinan con el operador lógico AND.

6.2.23. SegmentFilter

SegmentFilter define el segmento como un segmento simple o de secuencia. Una condición de segmento simple incluye condiciones de dimensiones y métricas para seleccionar las sesiones o los usuarios. Una condición de segmento de secuencia puede usarse para seleccionar usuarios o sesiones en función de una serie de condiciones secuenciales.

Representación JSON

```
{
  "not": boolean,

  // Union field simpleOrSequence can be only one of the following:
  "simpleSegment": {
    object(SimpleSegment)
  },
  "sequenceSegment": {
    object(SequenceSegment)
  },
  // End of list of possible types for union field simpleOrSequence.

}
```

Nombre del campo	Tipo	Descripción

not	boolean	Si el valor es "true", los resultados coinciden con el complemento del segmento simple o de secuencia. Por ejemplo, para obtener todas las visitas que no proceden de "Nueva York", podemos definir el segmento de este modo:
		"sessionSegment": {
		"segmentFilters": [{
		"simpleSegment" :{
		"orFiltersForSegment": [{
		"segmentFilterClauses":[{
		"dimensionFilter": {
		"dimensionName": "ga:city",
		"expressions": ["New York"]
		}

```
        }]
      }]
    },
    "not": "True"
  }]
},
```

Campo de unión simpleOrSequence. Es la definición de un segmento simple o de secuencia. El campo simpleOrSequence solo puede ser uno de los siguientes:

simpleSegment	object(SimpleSegment)	Las condiciones de segmento simple constan de una o varias condiciones de dimensión o métrica que se pueden combinar.
sequenceSegment	object(SequenceSegment)	Las condiciones de secuencia constan de uno o varios pasos, donde cada paso se define mediante una o varias condiciones de dimensión o métrica. Se pueden combinar varios pasos con operadores de secuencia especiales.

6.2.24. SimpleSegment

Las condiciones de segmento simple constan de una o varias condiciones de dimensión o métrica que se pueden combinar.

Representación JSON

```
{
  "orFiltersForSegment": [
    {
      object(OrFiltersForSegment)
```

```
    }
  ],
}
```

Nombre del campo	Tipo	Descripción
orFiltersForSegment[]	object(OrFiltersForSegment)	Lista de grupos de filtros de segmento combinados con el operador lógico AND.

6.2.25. OrFiltersForSegment

Lista de filtros de segmento del grupo OR combinados con el operador lógico OR.

Representación JSON

```
{
  "segmentFilterClauses": [
    {
      object(SegmentFilterClause)
    }
  ],
}
```

Nombre del campo	Tipo	Descripción
segmentFilterClauses[]	object(SegmentFilterClause)	Lista de filtros de segmento combinados con el operador OR.

6.2.26. SegmentFilterClause

Cláusula de filtro que se usa en una definición de segmento. Puede ser un filtro de métrica o de dimensión.

Representación JSON

```
{
  "not": boolean,

  // Union field dimensionOrMetricFilter can be only one of the following:
  "dimensionFilter": {
    object(SegmentDimensionFilter)
  },
  "metricFilter": {
    object(SegmentMetricFilter)
  },
  // End of list of possible types for union field dimensionOrMetricFilter.

}
```

Nombre del campo	Tipo	Descripción
not	boolean	Coincide con el complemento (!) del filtro.
Campo de unión dimensionOrMetricFilter. Filtro de dimensión o de métrica. El campo dimensionOrMetricFilter solo puede ser uno de estos:		
dimensionFilter	object(SegmentDimensionFilter)	Filtro de dimensión para la definición del segmento.

metricFilter	object(SegmentMetricFilter)	Filtro de métrica para la definición del segmento.

6.2.27. SegmentDimensionFilter

El filtro de dimensión especifica las opciones de filtrado de una dimensión.

Representación JSON

```
{
  "dimensionName": string,
  "operator": enum(Operator),
  "caseSensitive": boolean,
  "expressions": [
    string
  ],
  "minComparisonValue": string,
  "maxComparisonValue": string,

}
```

Nombre del campo	Tipo	Descripción
dimensionName	string	Nombre de la dimensión para la que se aplica el filtro.
operator	enum(Operator)	Operador que se usará para relacionar la dimensión con las expresiones.

caseSensitive	boolean	Indica si la concordancia debe distinguir entre mayúsculas y minúsculas. No se tiene en cuenta con el operador IN_LIST.
expressions[]	string	Lista de expresiones. Solo el primer elemento se usa para todos los operadores.
minComparisonValue	string	Valores de comparación mínimos para el tipo de concordancia BETWEEN.
maxComparisonValue	string	Valores de comparación máximos para el tipo de concordancia BETWEEN.

6.2.28. Operador

Distintos tipos de concordancia admitidos.

Valor de enumeración	Descripción
OPERATOR_UNSPECIFIED	Si no se especifica el tipo de concordancia, se usa REGEXP.
REGEXP	La expresión de concordancia se trata como una expresión regular. Los demás tipos de concordancia no se tratan como expresiones regulares.
BEGINS_WITH	Se obtiene el valor que empieza con la expresión de concordancia especificada.

ENDS_WITH	Se obtiene el valor que finaliza con la expresión de concordancia especificada.
PARTIAL	Concordancia de cadena secundaria.
EXACT	El valor debe coincidir exactamente con la expresión de concordancia.
IN_LIST	Esta opción se usa para especificar un filtro de dimensión cuya expresión puede incluir cualquier valor de una lista de valores determinada. Esto evita tener que evaluar varios filtros de dimensión de concordancia exacta unidos con el operador OR en cada fila de la respuesta. Por ejemplo: expressions: ["A", "B", "C"] Las filas de la respuesta cuyas dimensiones tienen el valor A, B o C coinciden con este DimensionFilter.
NUMERIC_LESS_THAN	Filtros de comparación de enteros. No distinguen entre mayúsculas y minúsculas y la expresión debe ser una cadena que representa un número entero. Condiciones de error: ● Si la expresión no es un valor int64 válido, el cliente obtendrá un error. ● Las dimensiones introducidas que no sean valores int64 válidos no coincidirán nunca con el filtro. Comprueba si la dimensión es numéricamente menor que la expresión de concordancia.

NUMERIC_GREATER_THAN	Comprueba si la dimensión es numéricamente mayor que la expresión de concordancia.
NUMERIC_BETWEEN	Comprueba si la dimensión se encuentra numéricamente entre el valor mínimo y máximo de la expresión de concordancia (no se incluyen los límites).

6.2.29. SegmentMetricFilter

Filtro de métrica que se usa en una cláusula de filtro de segmento.

Representación JSON

```
{
  "scope": enum(Scope),
  "metricName": string,
  "operator": enum(Operator),
  "comparisonValue": string,
  "maxComparisonValue": string,

}
```

Nombre del campo	Tipo	Descripción
scope	enum(Scope)	El alcance de una métrica indica el nivel en que se define dicha métrica. El alcance de métrica especificado debe ser igual o mayor que su alcance principal definido en el modelo de datos. El alcance principal se define si el segmento selecciona usuarios o sesiones.

metricName	string	Métrica por la que se filtrarán los resultados. Un campo metricFilter debe incluir un nombre de métrica.
operator	enum(Operator)	Especifica la operación que debe realizarse para comparar la métrica. El valor predeterminado es EQUAL.
comparisonValue	string	Valor de comparación. Si el operador es BETWEEN, este valor se trata como el valor de comparación mínimo.
maxComparisonValue	string	El valor de comparación máximo solo se usa para el operador BETWEEN.

6.2.30. Alcance

El alcance de una métrica indica el nivel en el que se define dicha métrica: PRODUCT, HIT, SESSION o USER. Los valores de métrica también se pueden registrar en alcances superiores a su alcance principal. Por ejemplo, ga:pageviews y ga:transactions se pueden registrar en los niveles SESSION y USER con solo agregarlos para cada hit que se produce en esas sesiones o para dichos usuarios.

Valor de enumeración	Descripción
UNSPECIFIED_SCOPE	Si no se especifica el alcance, se usa el alcance predeterminado USER o SESSION en función de si el segmento intenta seleccionar usuarios o sesiones.

PRODUCT	Alcance de producto.
HIT	Alcance de hit.
SESSION	Alcance de sesión.
USER	Alcance de usuario.

6.2.31. Operador

Distintas opciones de tipos de comparación.

Valor de enumeración	Descripción
UNSPECIFIED_OPERATOR	Si no se especifica ningún operador, se usa el operador LESS_THAN.
LESS_THAN	Comprueba si el valor de la métrica es inferior al valor de comparación.
GREATER_THAN	Comprueba si el valor de la métrica es superior al valor de comparación.
EQUAL	Operador igual.

BETWEEN	En el operador BETWEEN, los valores mínimo y máximo son exclusivos. Se usará LT y GT para la comparación.

6.2.32. SequenceSegment

Las condiciones de secuencia constan de uno o varios pasos, donde cada paso se define mediante una o varias condiciones de dimensión o métrica. Se pueden combinar varios pasos con operadores de secuencia especiales.

Representación JSON

```
{
  "segmentSequenceSteps": [
    {
      object(SegmentSequenceStep)
    }
  ],
  "firstStepShouldMatchFirstHit": boolean,

}
```

Nombre del campo	Tipo	Descripción
segmentSequenceSteps[]	object(SegmentSequenceStep)	Lista de pasos de la secuencia.
firstStepShouldMatchFirstHit	boolean	Si se define, la condición del primer paso debe coincidir con el primer hit del visitante (dentro del periodo).

6.2.33. SegmentSequenceStep

Definición de una secuencia de segmento.

Representación JSON

```
{
  "orFiltersForSegment": [
    {
      object(OrFiltersForSegment)
    }
  ],
  "matchType": enum(MatchType),

}
```

Nombre del campo	Tipo	Descripción
orFiltersForSegment[]	object(OrFiltersForSegment)	Una secuencia se especifica con una lista de filtros agrupados con OR que se combinan con el operador AND.
matchType	enum(MatchType)	Especifica si el paso es inmediatamente anterior al próximo paso o si puede producirse en cualquier momento antes de este.

6.2.34. MatchType

Tipo de concordancia de la secuencia.

Valor de enumeración	Descripción
UNSPECIFIED_MATCH_TYPE	Si no se especifica el tipo de concordancia se usa el valor PRECEDES.
PRECEDES	El operador indica que el paso anterior precede al paso siguiente.
IMMEDIATELY_PRECEDES	El operador indica que el paso anterior precede inmediatamente al paso siguiente.

6.2.35. Tabla dinámica

La Tabla dinámica describe la sección dinámica de la solicitud. Permite reorganizar la información de la tabla en algunos informes moviendo los datos a otra dimensión.

Representación JSON

```
{
  "dimensions": [
    {
      object(Dimension)
    }
  ],
  "dimensionFilterClauses": [
    {
      object(DimensionFilterClause)
    }
  ],
  "metrics": [
    {
      object(Metric)
    }
  ],
  "startGroup": number,
  "maxGroupCount": number,

}
```

Nombre del campo	Tipo	Descripción
dimensions[]	object(Dimension)	Lista de dimensiones que se muestran como columnas dinámicas. Las dimensiones dinámicas pueden tener hasta 4 dimensiones y forman parte de la restricción que establece el número máximo de dimensiones permitidas en la solicitud.

dimensionFilterClauses[]	object(DimensionFilterClause)	Los campos DimensionFilterClauses se combinan de forma lógica con el operador AND: solo los datos que se indican en estos campos DimensionFilterClauses afectan a los valores de esta sección dinámica. Pueden usarse filtros de dimensiones para limitar las columnas que se muestran en la sección dinámica. Por ejemplo, si la dimensión solicitada en la sección dinámica es ga:browser y se especifican filtros de clave para limitar ga:browser solo a "IE" o "Firefox", solo se mostrarán estos dos navegadores como columnas.
metrics[]	object(Metric)	Métricas dinámicas. Forman parte de la restricción que establece el número máximo de métricas permitidas en la solicitud.
startGroup	number	Si se solicitan k métricas, la respuesta incluirá varias columnas de datos múltiples de k del informe. Por ejemplo, si creas una columna dinámica de la dimensión ga:browser, obtendrás k columnas para "Firefox", k columnas para "IE", k columnas para "Chrome", etc. La clasificación de los grupos de columnas está determinada por el orden descendente del "total" de los primeros k valores. Las uniones se rompen con la clasificación lexicográfica de la primera dimensión dinámica, después con la clasificación lexicográfica de la segunda

		dimensión dinámica y así sucesivamente. Por ejemplo, si los totales del primer valor de Firefox, IE y Chrome eran 8, 2 y 8, respectivamente, el orden de las columnas será Chrome, Firefox, IE. Los siguientes valores permiten elegir los grupos de k columnas que se incluirán en la respuesta.
maxGroupCount	number	Especifica el número máximo de grupos que deben devolverse. El valor predeterminado es 10 y el valor máximo es 1000.

6.2.36. CohortGroup

Define un grupo de cohortes. Por ejemplo:

```
"cohortGroup": {
  "cohorts": [{
    "name": "cohort 1",
    "type": "FIRST_VISIT_DATE",
    "dateRange": { "startDate": "2015-08-01", "endDate": "2015-08-01" }
  },{
    "name": "cohort 2"
     "type": "FIRST_VISIT_DATE"
```

```
    "dateRange": { "startDate": "2015-07-01", "endDate": "2015-07-01" }

  }]

}
```

Representación JSON

```
{
  "cohorts": [
    {
      object(Cohort)
    }
  ],
  "lifetimeValue": boolean,

}
```

Nombre del campo	Tipo	Descripción
cohorts[]	object(Cohort)	Definición de la cohorte.

lifetimeValue	boolean	Habilita el valor del tiempo de vida del cliente (TVC). El VCV mide el valor del ciclo de vida del cliente de los usuarios que se han obtenido a través de distintos canales. Consulta los informes Análisis de cohortes y Valor del tiempo de vida del cliente. Si el valor del tiempo de vida del cliente es "false": ● Los valores de métrica son similares a los valores del informe de cohortes de la interfaz web. ● Los periodos de la definición de las cohortes deben coincidir con la semana y el mes naturales. Por ejemplo, si se solicita ga:cohortNthWeek, el valor de startDate en la definición de la cohorte debe ser domingo y el valor de endDate debe ser el sábado siguiente; para ga:cohortNthMonth, el valor de startDate debe ser el primer día del mes y el valor de endDate debe ser el último día del mes. Si el valor del ciclo de vida del cliente es "true": ● Los valores de la métrica corresponden a los valores del informe Valor del ciclo de vida del cliente de la interfaz web. ● El informe Valor del ciclo de vida del cliente indica la evolución del valor del usuario (Ingresos) y de sus interacciones (vistas de aplicación, consecuciones de objetivo, sesiones y duración de las sesiones) durante los 90 días posteriores a su adquisición.

- Las métricas se calculan como valor medio acumulado por usuario y por incremento de tiempo.
- No es necesario que los periodos de la definición de la cohorte coincidan con la semana y el mes naturales.
- El valor de viewId debe ser un ID de vista de aplicación.

6.2.37. Cohorte

Define una cohorte. Una cohorte es un grupo de usuarios que comparten una característica común. Por ejemplo, todos los usuarios con la misma fecha de adquisición pertenecen a la misma cohorte.

Representación JSON

```
{
  "name": string,
  "type": enum(Type),
  "dateRange": {
    object(DateRange)
  },

}
```

Nombre del campo	Tipo	Descripción

name	string	Nombre exclusivo para la cohorte. Si no se define ningún nombre, se generará uno automáticamente con los valores de cohorte_[1234…].
type	enum(Type)	Tipo de cohorte. Por ahora, el único tipo admitido es FIRST_VISIT_DATE. Si no se especifica ningún valor en este campo, se usa el tipo de cohorte FIRST_VISIT_DATE.
dateRange	object(DateRange)	Este campo se usa para la cohorte FIRST_VISIT_DATE. La cohorte selecciona los usuarios cuya fecha de la primera visita se encuentra entre la fecha de inicio y la fecha de finalización definidas en el campo DateRange. Los periodos deben coincidir con las solicitudes de cohorte. Si la solicitud contiene ga:cohortNthDay, el periodo será exactamente de un día. Si contiene ga:cohortNthWeek, el periodo deberá coincidir con el límite de la semana (que empieza un domingo y acaba el sábado siguiente). Si contiene ga:cohortNthMonth, el período debe coincidir con el mes (que empieza el primer día y acaba el último día del mes). Las solicitudes de VCV no presentan estas restricciones. No es necesario indicar ningún periodo en el campo reportsRequest.dateRanges.

6.2.38. Tipo

Tipo de cohorte.

Valor de enumeración	Descripción
UNSPECIFIED_COHORT_TYPE	Si no se especifica ningún valor, se utiliza FIRST_VISIT_DATE.
FIRST_VISIT_DATE	Cohortes seleccionadas en función de la fecha de la primera visita.

6.2.39. Informe

Respuesta de datos a la solicitud.

Representación JSON

```
{
  "columnHeader": {
    object(ColumnHeader)
  },
  "data": {
    object(ReportData)
  },
  "nextPageToken": string,

}
```

Nombre del campo	Tipo	Descripción
columnHeader	object(ColumnHeader)	Encabezados de columna.

data	object(Report Data)	Datos de la respuesta.
nextPageToken	string	Token de página para recuperar la siguiente página de resultados de la lista.

6.2.40. ColumnHeader

Encabezados de columna.

Representación JSON

```
{
  "dimensions": [
    string
  ],
  "metricHeader": {
    object(MetricHeader)
  },
}
```

Nombre del campo	Tipo	Descripción
dimensions[]	string	Nombres de las dimensiones de la respuesta.
metricHeader	object(MetricHeader)	Encabezados de las métricas de la respuesta.

6.2.41. MetricHeader

Encabezados de las métricas.

Representación JSON

```
{
  "metricHeaderEntries": [
    {
      object(MetricHeaderEntry)
    }
  ],
  "pivotHeaders": [
    {
      object(PivotHeader)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción
metricHeaderEntries[]	object(MetricHeaderEntry)	Encabezados de las métricas de la respuesta.
pivotHeaders[]	object(PivotHeader)	Encabezados de las tablas dinámicas de la respuesta.

6.2.42. MetricHeaderEntry

Encabezado de las métricas.

Representación JSON

```
{
  "name": string,
  "type": enum(MetricType),

}
```

Nombre del campo	Tipo	Descripción
name	string	Nombre del encabezado.
type	enum(MetricType)	Tipo de métrica, por ejemplo, INTEGER.

6.2.43. PivotHeader

Encabezados de cada una de las secciones dinámicas definidas en la solicitud.

Representación JSON

```
{
  "pivotHeaderEntries": [
    {
      object(PivotHeaderEntry)
    }
  ],
  "totalPivotGroupsCount": number,

}
```

Nombre del campo	Tipo	Descripción

pivotHeaderEntries[]	object(PivotHeaderEntry)	Encabezado de una sola sección dinámica.
totalPivotGroupsCount	number	Número total de grupos de la sección dinámica.

6.2.44. PivotHeaderEntry

Encabezados de cada una de las columnas de métricas correspondientes a las métricas solicitadas en la sección dinámica de la respuesta.

Representación JSON

```
{
  "dimensionNames": [
    string
  ],
  "dimensionValues": [
    string
  ],
  "metric": {
    object(MetricHeaderEntry)
  },

}
```

Nombre del campo	Tipo	Descripción
dimensionNames[]	string	Nombre de las dimensiones en la respuesta dinámica.

dimensionValues[]	string	Valores de las dimensiones en la respuesta dinámica.
metric	object(MetricHeaderEntry)	Encabezado de la métrica en la respuesta dinámica.

6.2.45. ReportData

Parte del informe que incluye los datos.

Representación JSON

```
{
  "rows": [
    {
      object(ReportRow)
    }
  ],
  "totals": [
    {
      object(DateRangeValues)
    }
  ],
  "rowCount": number,
  "minimums": [
    {
      object(DateRangeValues)
    }
  ],
  "maximums": [
    {
      object(DateRangeValues)
    }
  ],
  "samplesReadCounts": [
    string
  ],
  "samplingSpaceSizes": [
    string
  ],
  "isDataGolden": boolean,

}
```

Nombre del campo	Tipo	Descripción
rows[]	object(ReportRow)	Hay un campo ReportRow para cada combinación exclusiva de dimensiones.

totals[]	object(DateRangeValues)	Todos los formatos de valor solicitados en el conjunto de todas las columnas de cada periodo solicitado obtienen un total. El total de un formato de valor se obtiene calculando primero el total de métricas indicado en el formato de valor y después evaluando el formato de valor como expresión escalar. Por ejemplo, para obtener los "totales" de 3 / (ga:sessions + 2) calculamos 3 / ((sum of all relevant ga:sessions) + 2). Los totales se calculan antes de la paginación.
rowCount	number	Total de filas que coinciden con la consulta.
minimums[]	object(DateRangeValues)	Valores mínimo y máximo de todas las filas que coinciden con la consulta. Ambos campos están vacíos si el valor hideValueRanges de la solicitud es "false" o si el valor de rowCount es cero.
maximums[]	object(DateRangeValues)	Valores mínimo y máximo de todas las filas que coinciden con la consulta. Ambos campos están vacíos si el valor hideValueRanges de la solicitud es "false" o si el valor de rowCount es cero.
samplesReadCounts[]	string (int64 format)	Si los resultados se muestrean, se devuelve el número total de muestras leídas, con una entrada por periodo. Si los resultados no se muestrean, este campo no se define. Consulta los detalles en la guía de programadores.

samplingSpaceSizes[]	string (int64 format)	Si los resultados se muestrean, se devuelve el número total de muestras presentes, con una entrada por periodo. Si los resultados no se muestrean, este campo no se define. Consulta los detalles en la guía de programadores.
isDataGolden	boolean	Indica si los datos de la respuesta a la solicitud son valiosos o no. Los datos son valiosos cuando la misma solicitud se realiza con posterioridad y no obtiene nuevos resultados.

6.2.46. ReportRow

Fila del informe.

Representación JSON

```
{
  "dimensions": [
    string
  ],
  "metrics": [
    {
      object(DateRangeValues)
    }
  ],

}
```

Nombre del campo	Tipo	Descripción

dimensions[]	string	Lista de dimensiones solicitadas.
metrics[]	object(DateRangeValues)	Lista de métricas de cada periodo solicitado.

6.2.47. DateRangeValues

Se usa para devolver una lista de métricas para una única combinación de periodo y dimensión.

Representación JSON

```
{
  "values": [
    string
  ],
  "pivotValueRegions": [
    {
      object(PivotValueRegion)
    }
  ],
}
```

Nombre del campo	Tipo	Descripción
values[]	string	Cada valor corresponde a una métrica de la solicitud.

pivotValueRegions[]	object(PivotValueRegion)	Valores de cada sección dinámica.

6.2.48. PivotValueRegion

Valores de métrica de la sección dinámica.

Representación JSON

```
{
  "values": [
    string
  ],

}
```

Nombre del campo	Tipo	Descripción
values[]	string	Valores de las métricas en cada una de las secciones dinámicas.

6.2.49. Parámetros de consulta estándar

Los parámetros de consulta siguientes se pueden utilizar con todos los métodos y recursos de la API (Google Developers, s.f.).

Los parámetros de consulta que se usan en todas las operaciones de la versión 4 de la API de informes de Analytics se muestran en la tabla siguiente.

Notas (sobre las claves de la API y los tokens de autenticación):

1. El parámetro key es obligatorio en todas las solicitudes, a menos que proporciones un token de OAuth 2.0 en la solicitud.

2. Debes enviar un token de autorización en todas las solicitudes que requieran el ámbito de OAuth. OAuth 2.0 es el único protocolo de autorización admitido.
3. Puedes indicar un token de OAuth 2.0 con cualquier solicitud de dos formas:
    - Utilizando el parámetro de consulta access_token así: ?access_token=oauth2-token
    - Utilizando el encabezado de HTTP Authorization así: Authorization: Bearer oauth2-token

Todos los parámetros son opcionales, a menos que se indique lo contrario.

Parámetro	Significado	Notas
access_token	Token de OAuth 2.0 para el usuario actual.	● Una forma posible de proporcionar un token de OAuth 2.0.
alt	Formato de datos de la respuesta.	● Valores válidos: json, media y proto ● Valor predeterminado: json
fields	Selector que especifica un subconjunto de campos para incluirlos en la respuesta.	● Para obtener más información, consulta la sección de respuesta parcial en el documento Consejos para el rendimiento. ● Se utiliza para mejorar el rendimiento.
key	Clave de la API (OBLIGATORIA*)	● *Es obligatorio, salvo si proporciona un token de OAuth 2.0. ● La clave de la API identifica tu proyecto y te proporciona acceso a la API, a la cuota y a los informes. ● Puedes obtener la clave de la API de tu proyecto en la consola de la API de Google.
prettyPrint	Muestra la respuesta con sangrados y saltos de línea.	● Devuelve la respuesta con un formato legible si el valor es true. ● Valor predeterminado: true. ● Cuando el valor es false, puede reducir el tamaño de la carga útil de la respuesta, lo que puede conllevar una mejora del rendimiento en algunos entornos.

quotaUser	Cadena arbitraria que identifica de forma exclusiva a un usuario.	• Te permite aplicar cuotas por usuario desde una aplicación orientada al servidor, incluso en aquellos casos en los que no se conoce la dirección IP del usuario. Esto puede ocurrir, por ejemplo, con aplicaciones que ejecutan tareas cron en App Engine en nombre de un usuario. • Puedes elegir cualquier cadena arbitraria que identifique de forma exclusiva a un usuario, pero se limita a 40 caracteres. • Obtén más información sobre cómo limitar el uso de la API.

6.2.50. Límites y cuotas en las solicitudes a la API

Las cuotas siguientes son aplicables a todas las API de informes como, por ejemplo, la versión 3 de la API de informes centrales, la versión 4 de la API de informes de Analytics, la versión 3 de la API de informes en tiempo real o la versión 3 de la API de informes de embudos multicanal (Google Developers, s.f.):

- 10.000 solicitudes por vista (perfil) al día; no se puede incrementar
- 10 solicitudes simultáneas por vista (perfil); no se puede incrementar

Nota: Aunque se aúne varias solicitudes de las API de informes en una única solicitud, es decir, en bloque, no se puede hacer más solicitudes de las que admiten las cuotas.

Si la solicitud que se realiza a una API de informes falla y aparece el código de respuesta 500 o 503, se puede volver a enviarla. Google Analytics permite que se produzcan:

- 10 solicitudes fallidas por proyecto y perfil cada hora
- 50 solicitudes fallidas por proyecto y perfil cada día

Si el número de solicitudes fallidas supera estas cuotas, aparecerá el error siguiente: “Quota Error: The number of recent failed writes is too high.”

Si se supera la cuota de solicitud a una API de Google Analytics, esta devuelve el código de error 403 o 429 y un mensaje indicando que la cuenta ha superado la cuota.

6.2.51. Respuestas de error

Si una solicitud de la API de administración se realiza correctamente, la API devuelve el código de estado 200. Si se produce un error en una solicitud, la API devuelve un código de estado HTTP y un motivo en la respuesta según el tipo de error. Además, el cuerpo de

la respuesta contiene una descripción detallada de lo que ha provocado el error. A continuación, se presenta un ejemplo de una respuesta de error (Google Developers, s.f.):


```
{
 "error": {
  "errors": [
   {
    "domain": "global",
    "reason": "invalidParameter",
    "message": "Invalid value '-1' for max-results. Value must be within the range: [1, 1000]",
    "locationType": "parameter",
    "location": "max-results"
   }
  ],
  "code": 400,
  "message": "Invalid value '-1' for max-results. Value must be within the range: [1, 1000]"
 }
}
```


6.2.51.1. Tabla de errores

Código	Motivo	Descripción	Acción recomendada
400	invalidParameter	Indica que un parámetro de solicitud tiene un valor no válido. En los campos locationType y location de la respuesta de error se	No vuelvas a intentar la acción sin antes corregir el problema. Debes proporcionar un valor válido para el parámetro

		proporciona información sobre dicho valor.	especificado en la respuesta de error.
400	badRequest	Indica que la consulta no es válida. Por ejemplo, falta el ID superior o bien la combinación de dimensiones o métricas solicitadas no es válida.	No vuelvas a intentar la acción sin antes corregir el problema. Debes realizar cambios en la consulta de la API para que funcione.
401	invalidCredentials	Indica que el token de autenticación no es válido o ha caducado.	No vuelvas a intentar la acción sin antes corregir el problema. Debes obtener un nuevo token de autenticación.
403	insufficientPermissions	Indica que el usuario no tiene suficientes permisos para la entidad especificada en la consulta.	No vuelvas a intentar la acción sin antes corregir el problema. Debes obtener permisos suficientes para realizar la operación en la entidad especificada.
403	dailyLimitExceeded	Indica que el usuario ha superado la cuota diaria (por proyecto o por vista [perfil]).	No vuelvas a intentar la acción sin antes corregir el problema. Has agotado tu cuota diaria. Consulta Límites y cuotas en las solicitudes de API.

403	userRateLimitExceeded	Indica que se ha superado el límite de Consultas por usuario por cada 100 segundos. El valor predeterminado en la consola de APIs de Google es de 100 consultas cada 100 segundos por usuario. Puedes aumentar este límite de la consola hasta un máximo de 1000.	Vuelve a intentarlo con un retardo exponencial. Debes reducir la frecuencia con la que envías las solicitudes.
403	rateLimitExceeded	Indica que se han superado los límites de frecuencia de consultas por cada 100 segundos del proyecto.	Vuelve a intentarlo con un retardo exponencial. Debes reducir la frecuencia con la que envías las solicitudes.
403	quotaExceeded	Indica que se han alcanzado las 10 solicitudes simultáneas por vista (perfil) en la API de informes centrales.	Vuelve a intentarlo con un retardo exponencial. Debes esperar a que termine al menos una solicitud en curso correspondiente a esta vista (perfil).
500	internalServerError	Se ha producido un error interno inesperado del servidor.	No reintentes esta consulta más de una vez.
503	backendError	El servidor ha devuelto un error.	No vuelvas a intentar esta consulta más de una vez.

6.2.52. Informes de Google Ad Manager

Google Ad Manager es una completa plataforma de publicación de anuncios alojada que optimiza la gestión de anuncios, tanto si los publicas en sitios web como en páginas web para dispositivos móviles, aplicaciones móviles, juegos o una combinación de estos.

Google Ad Manager utiliza las mismas métricas que AdSense, por lo que, si ya se conocen esas herramientas, se encontrará un entorno familiar. Si no se es un experto en métricas, no hay problema. A continuación, se presenta un resumen rápido de las diez métricas esenciales que se puede utilizar en esta plataforma.

1. Impresiones de GAM: cada vez que se muestra un anuncio individual en el sitio web, se registra una impresión de anuncio de Ad Manager. Por ejemplo, si una página que contiene dos bloques de anuncios se visualiza una vez, se contabilizan dos impresiones.
2. Cobertura de GAM: la cobertura es el porcentaje de las solicitudes de anuncio que han devuelto al menos un anuncio. Normalmente, la cobertura permite identificar sitios web en los que Ad Manager no puede ofrecer anuncios segmentados. (Impresiones de anuncio/Total de solicitudes de anuncios) * 100.
3. Número de páginas vistas con obtención de ingresos de GAM: mide el número total de páginas vistas en tu propiedad que se han mostrado con un anuncio desde tu cuenta de Ad Manager vinculada. Nota: Una sola página puede tener varios bloques de anuncios.
4. Impresiones de GAM/sesión: indica la proporción de impresiones de anuncios de Ad Manager en relación con las sesiones de Analytics (Impresiones de anuncio/Sesiones de Analytics).
5. % de impresiones visibles de GAM: representa el porcentaje de las impresiones de anuncio que eran visibles. Se considera que una impresión es visible cuando aparece en el navegador de un usuario y este ha podido verla.
6. Clics de GAM: indica el número de veces que se ha hecho clic en los anuncios de Ad Manager de tu sitio web.
7. CTR de GAM: indica el porcentaje de impresiones de Ad Manager que han generado un clic en un anuncio.
8. Ingresos de GAM: es una estimación del total de ingresos publicitarios según las impresiones publicadas.
9. Ingresos de GAM/1000 sesiones: expresa el total de ingresos estimados procedentes de los anuncios de Ad Manager por cada 1000 sesiones de Analytics.

Esta métrica se basa en las sesiones en tu sitio web y no en las impresiones de anuncio.

10. eCPM de GAM: es el coste estimado por mil páginas vistas. Indica los ingresos de Ad Manager que se obtienen por cada 1000 páginas vistas.

6.3. Listado completo de dimensiones y métricas de la API de informes de Google Analytics

A continuación, se puede ver un listado completo de las dimensiones y métricas de la API de informes de Google Analytics (Google Developers, s.f.).

6.3.1. Usuario

6.3.1.1. Dimensiones

- Tipo de usuario ga:userType
- Contador de Sesiones ga:sessionCount
- Días desde la última sesión ga:daysSinceLastSession
- Valor definido por el usuario ga:userDefinedValue
- Cubo de usuario ga:userBucket

6.3.1.2. Métricas

- Usuarios ga:users
- Usuarios nuevos ga:newUsers
- % De nuevas sesiones ga:percentNewSessions
- Usuarios activos durante 1 día ga:1dayUsers
- Usuarios activos durante 7 días ga:7dayUsers
- Usuarios activos durante 14 días ga:14dayUsers
- Usuarios activos durante 28 días ga:28dayUsers
- Usuarios activos de 30 días ga:30dayUsers
- Número de sesiones por usuario ga:sessionsPerUser

6.3.2. Sesión

6.3.2.1. Dimensiones

- Duración de la sesión ga:sessionDurationBucket

6.3.2.2. Métricas

- Sesiones ga:sessions
- Rebotes ga:bounces

- Porcentaje de rebote ga:bounceRate
- Duración de la sesión ga:sessionDuration
- Promedio Duración de la sesión ga:avgSessionDuration
- Combinaciones de dimensiones únicas ga:uniqueDimensionCombinations
- Visitas ga:hits

6.3.3. Fuentes de tráfico

6.3.3.1. Dimensiones

- Ruta de referencia ga:referralPath
- Remitente completo ga:fullReferrer
- Campaña ga:campaign
- Fuente ga:source
- Medio ga:medium
- Fuente / medio ga:sourceMedium
- Palabra clave ga:keyword
- Contenido del anuncio ga:adContent
- Red social ga:socialNetwork
- Referencia de fuente social ga:hasSocialSourceReferral
- Código de campaña ga:campaignCode

6.3.3.2. Métricas

- Búsquedas orgánicas ga:organicSearches

6.3.4. Adwords

6.3.4.1. Dimensiones

- Google Ads: grupo de anuncios ga:adGroup
- Google Ads: espacio publicitario ga:adSlot
- Red de distribución de anuncios ga:adDistributionNetwork
- Tipo de coincidencia de consulta ga:adMatchType
- Tipo de concordancia de palabra clave ga:adKeywordMatchType
- Consulta de busqueda ga:adMatchedQuery
- Dominio de colocación ga:adPlacementDomain
- URL de ubicación ga:adPlacementUrl

- Formato de anuncio ga:adFormat
- Tipo de orientación ga:adTargetingType
- Tipo de colocación ga:adTargetingOption
- URL visible ga:adDisplayUrl
- URL de destino ga:adDestinationUrl
- ID de cliente de Google Ads ga:adwordsCustomerID
- ID de campaña de Google Ads ga:adwordsCampaignID
- ID del grupo de anuncios de Google Ads ga:adwordsAdGroupID
- ID de creatividad de Google Ads ga:adwordsCreativeID
- ID de criterios de Google Ads ga:adwordsCriteriaID
- Consultar recuento de palabras ga:adQueryWordCount
- Anuncio de video TrueView ga:isTrueViewVideoAd

6.3.4.2. Métricas

- Impresiones ga:impressions
- Clics ga:adClicks
- Costo ga:adCost
- CPM ga:CPM
- CPC ga:CPC
- CTR ga:CTR
- Costo por transacción ga:costPerTransaction
- Coste por conversión de objetivo ga:costPerGoalConversion
- Costo por conversión ga:costPerConversion
- RPC ga:RPC
- ROAS ga:ROAS

6.3.5. Conversiones de objetivos

6.3.5.1. Dimensiones

- Lugar de finalización del objetivo ga:goalCompletionLocation
- Paso anterior del objetivo - 1 ga:goalPreviousStep1
- Paso anterior del objetivo - 2 ga:goalPreviousStep2
- Paso anterior del objetivo - 3 ga:goalPreviousStep3

6.3.5.2. Métricas

- Inicios del objetivo XX ga:goalXXStarts
- Inicio de objetivos ga:goalStartsAll
- Consecuciones del objetivo XX ga:goalXXCompletions
- Consecuciones de objetivos ga:goalCompletionsAll
- Valor del objetivo XX ga:goalXXValue
- Valor del objetivo ga:goalValueAll
- Valor de objetivo por sesión ga:goalValuePerSession
- Tasa de conversión del objetivo XX ga:goalXXConversionRate
- Tasa de conversión de objetivos ga:goalConversionRateAll
- Embudos de conversión abandonados del objetivo XX ga:goalXXAbandons
- Embudos Abandonados ga:goalAbandonsAll
- Tasa de abandono del objetivo XX ga:goalXXAbandonRate
- Tasa de abandono total ga:goalAbandonRateAll

6.3.6. Plataforma o dispositivo

6.3.6.1. Dimensiones

- Navegador ga:browser
- Versión del navegador ga:browserVersion
- Sistema operativo ga:operatingSystem
- Versión del sistema operativo ga:operatingSystemVersion
- Marca de dispositivo móvil ga:mobileDeviceBranding
- Modelo de dispositivo móvil ga:mobileDeviceModel
- Selector de entrada móvil ga:mobileInputSelector
- Información del dispositivo móvil ga:mobileDeviceInfo
- Nombre de marketing del dispositivo móvil ga:mobileDeviceMarketingName
- Categoría de dispositivo ga:deviceCategory
- Tamaño del navegador ga:browserSize
- Fuente de datos ga:dataSource

6.3.7. Red geográfica

6.3.7.1. Dimensiones

- Continente ga:continent
- Subcontinente ga:subContinent
- País ga:country
- Región ga:region
- Metro ga:metro
- Ciudad ga:city
- Latitud ga:latitude
- Longitud ga:longitude
- Dominio de Red ga:networkDomain
- Proveedor de servicio ga:networkLocation
- ID de ciudad ga:cityId
- ID de continente ga:continentId
- Código ISO del país ga:countryIsoCode
- ID del metro ga:metroId
- ID de región ga:regionId
- Código ISO de la región ga:regionIsoCode
- Código de subcontinente ga:subContinentCode

6.3.8. Sistema

6.3.8.1. Dimensiones

- Versión flash ga:flashVersion
- Soporte de Java ga:javaEnabled
- Idioma ga:language
- Colores de pantalla ga:screenColors
- Nombre de visualización de propiedad de origen ga:sourcePropertyDisplayName
- ID de seguimiento de propiedad de origen ga:sourcePropertyTrackingId
- Resolución de la pantalla ga:screenResolution

6.3.9. Seguimiento de página

6.3.9.1. Dimensiones

- Nombre de host ga:hostname
- Página ga:pagePath
- Nivel de ruta de página 1 ga:pagePathLevel1
- Nivel de ruta de página 2 ga:pagePathLevel2
- Nivel de ruta de página 3 ga:pagePathLevel3
- Nivel de ruta de página 4 ga:pagePathLevel4
- Título de la página ga:pageTitle
- Página de destino ga:landingPagePath
- Segunda pagina ga:secondPagePath
- Salir de la página ga:exitPagePath
- Ruta de la página anterior ga:previousPagePath
- Profundidad de página ga:pageDepth

6.3.9.2. Métricas

- Valor de página ga:pageValue
- Entradas ga:entrances
- Entradas / páginas vistas ga:entranceRate
- Vistas de página ga:pageviews
- Páginas / Sesión ga:pageviewsPerSession
- Vistas de página únicas ga:uniquePageviews
- Tiempo en la página ga:timeOnPage
- Promedio Tiempo en la página ga:avgTimeOnPage
- Salidas ga:exits
- % Salida ga:exitRate

6.3.10. Agrupación de contenido

6.3.10.1. Dimensiones

- Página de aterrizaje Grupo XX ga:landingContentGroupXX
- Página anterior Grupo XX ga:previousContentGroupXX
- Grupo de páginas XX ga:contentGroupXX

6.3.10.2. Métricas

- Vistas únicas XX ga:contentGroupUniqueViewsXX

6.3.11. Búsqueda interna

6.3.11.1. Dimensiones

- Estado de búsqueda del sitio ga:searchUsed
- Término de búsqueda ga:searchKeyword
- Palabra clave refinada ga:searchKeywordRefinement
- Categoría de búsqueda del sitio ga:searchCategory
- Página de inicio ga:searchStartPage
- Página de destino ga:searchDestinationPage
- Buscar página de destino ga:searchAfterDestinationPage

6.3.11.2. Métricas

- Páginas vistas de resultados ga:searchResultViews
- Total de búsquedas únicas ga:searchUniques
- Páginas vistas de resultados / búsqueda ga:avgSearchResultViews
- Sesiones con búsqueda ga:searchSessions
- % De sesiones con búsqueda ga:percentSessionsWithSearch
- Profundidad de búsqueda ga:searchDepth
- Promedio Profundidad de búsqueda ga:avgSearchDepth
- Buscar refinamientos ga:searchRefinements
- % De refinamientos de búsqueda ga:percentSearchRefinements
- Tiempo después de la búsqueda ga:searchDuration
- Tiempo después de la búsqueda ga:avgSearchDuration
- Salidas de búsqueda ga:searchExits
- % De salidas de búsqueda ga:searchExitRate
- Tasa de conversión del objetivo XX de búsquedas en el sitio ga:searchGoalXXConversionRate
- Tasa de conversión del objetivo de búsqueda en el sitio ga:searchGoalConversionRateAll
- Valor de objetivo por búsqueda ga:goalValueAllPerSearch

6.3.12. Velocidad del sitio

6.3.12.1. Métricas

- Tiempo de carga de la página (ms) ga:pageLoadTime
- Muestra de carga de página ga:pageLoadSample
- Promedio Tiempo de carga de la página (segundos) ga:avgPageLoadTime
- Tiempo de búsqueda de dominio (ms) ga:domainLookupTime
- Promedio Tiempo de búsqueda de dominio (seg) ga:avgDomainLookupTime
- Tiempo de descarga de la página (ms) ga:pageDownloadTime
- Promedio Tiempo de descarga de la página (segundos) ga:avgPageDownloadTime
- Tiempo de redireccionamiento (ms) ga:redirectionTime
- Promedio Tiempo de redireccionamiento (seg) ga:avgRedirectionTime
- Tiempo de conexión del servidor (ms) ga:serverConnectionTime
- Promedio Tiempo de conexión del servidor (seg) ga:avgServerConnectionTime
- Tiempo de respuesta del servidor (ms) ga:serverResponseTime
- Promedio Tiempo de respuesta del servidor (seg) ga:avgServerResponseTime
- Muestra de métricas de velocidad ga:speedMetricsSample
- Tiempo interactivo del documento (ms) ga:domInteractiveTime
- Promedio Tiempo interactivo del documento (seg) ga:avgDomInteractiveTime
- Tiempo de carga del contenido del documento (ms) ga:domContentLoadedTime
- Promedio Tiempo de carga del contenido del documento (seg) ga:avgDomContentLoadedTime
- Ejemplo de métricas de latencia DOM ga:domLatencyMetricsSample

6.3.13. Seguimiento de aplicaciones

6.3.13.1. Dimensiones

- ID del instalador de la aplicación ga:appInstallerId
- Version de aplicacion ga:appVersion
- Nombre de la aplicación ga:appName
- ID de aplicación ga:appId
- Nombre de pantalla ga:screenName
- Profundidad de pantalla ga:screenDepth
- Pantalla de aterrizaje ga:landingScreenName
- Pantalla de salida ga:exitScreenName

6.3.13.2. Métricas

- Vistas de pantalla ga:screenviews
- Vistas de pantalla únicas ga:uniqueScreenviews
- Pantallas / Sesión ga:screenviewsPerSession
- Tiempo en la pantalla ga:timeOnScreen
- Promedio Tiempo en la pantalla ga:avgScreenviewDuration

6.3.14. Seguimiento de eventos

6.3.14.1. Dimensiones

- Categoría de evento ga:eventCategory
- Acción de evento ga:eventAction
- Etiqueta de evento ga:eventLabel

6.3.14.2. Métricas

- Eventos totales ga:totalEvents
- Eventos únicos ga:uniqueEvents
- Valor del evento ga:eventValue
- Promedio Valor ga:avgEventValue
- Sesiones con evento ga:sessionsWithEvent
- Eventos / Sesión con Evento ga:eventsPerSessionWithEvent

6.3.15. Comercio electrónico

6.3.15.1. Dimensiones

- ID de transacción ga:transactionId
- Afiliación ga:affiliation
- Sesiones a transacción ga:sessionsToTransaction
- Días para la transacción ga:daysToTransaction
- SKU del producto ga:productSku
- Producto ga:productName
- categoria de producto ga:productCategory
- Código de moneda ga:currencyCode
- Opciones de pago ga:checkoutOptions

- Promoción interna creativa ga:internalPromotionCreative
- ID de promoción interna a:internalPromotionId
- Nombre de promoción interna ga:internalPromotionName
- Puesto de promoción interna ga:internalPromotionPosition
- Código de cupón de pedido ga:orderCouponCode
- Producto de marca ga:productBrand
- Categoría de producto (comercio electrónico mejorado) ga:productCategoryHierarchy
- Categoría de producto Nivel XX ga:productCategoryLevelXX
- Código de cupón de producto ga:productCouponCode
- Nombre de la lista de productos ga:productListName
- Posición de la lista de productos ga:productListPosition
- Variante de producto ga:productVariant
- Etapa de compras ga:shoppingStage

6.3.15.2. Métricas

- Actas ga:transactions
- Tasa de conversión de comercio electrónico ga:transactionsPerSession
- Ingresos ga:transactionRevenue
- Promedio Valor del pedido ga:revenuePerTransaction
- Valor por sesión ga:transactionRevenuePerSession
- Envío ga:transactionShipping
- Impuesto ga:transactionTax
- Valor total ga:totalValue
- Cantidad ga:itemQuantity
- Compras únicas ga:uniquePurchases
- Promedio Precio ga:revenuePerItem
- Ingresos por producto ga:itemRevenue
- Promedio Cantidad ga:itemsPerPurchase
- Ingresos locales ga:localTransactionRevenue
- Envíos locales ga:localTransactionShipping
- Impuesto local ga:localTransactionTax
- Ingresos por productos locales ga:localItemRevenue
- Tasa de compra a detalle ga:buyToDetailRate
- Tarifa de carrito a detalle ga:cartToDetailRate

- CTR de promoción interna ga:internalPromotionCTR
- Clics de promoción interna ga:internalPromotionClicks
- Vistas de promoción interna ga:internalPromotionViews
- Cantidad de reembolso del producto local ga:localProductRefundAmount
- Cantidad de reembolso local ga:localRefundAmount
- El producto se agrega al carrito ga:productAddsToCart
- Pagar productos ga:productCheckouts
- Vistas de detalles del producto ga:productDetailViews
- Lista de productos CTR ga:productListCTR
- Clics en la lista de productos ga:productListClicks
- Vistas de la lista de productos ga:productListViews
- Cantidad de reembolso del producto ga:productRefundAmount
- Devoluciones de productos ga:productRefunds
- El producto se elimina del carrito ga:productRemovesFromCart
- Ingresos del producto por compra ga:productRevenuePerPurchase
- Cantidad agregada al carrito ga:quantityAddedToCart
- Cantidad desprotegida ga:quantityCheckedOut
- Cantidad reembolsada ga:quantityRefunded
- Cantidad eliminada del carrito ga:quantityRemovedFromCart
- cantidad devuelta ga:refundAmount
- Ingresos por usuario ga:revenuePerUser
- Reembolsos ga:totalRefunds
- Transacciones por usuario ga:transactionsPerUser

6.3.16. Interacciones sociales

6.3.16.1. Dimensiones

- Red social ga:socialInteractionNetwork
- Acción social ga:socialInteractionAction
- Red social y acción (Hit) ga:socialInteractionNetworkAction
- Entidad social ga:socialInteractionTarget
- Tipo social ga:socialEngagementType

6.3.16.2. Métricas

- Acciones sociales ga:socialInteractions

- Acciones sociales únicas ga:uniqueSocialInteractions
- Acciones por sesión social ga:socialInteractionsPerSession

6.3.17. Tiempos de usuario

6.3.17.1. Dimensiones

- Categoría de tiempo ga:userTimingCategory
- Etiqueta de tiempo ga:userTimingLabel
- Variable de tiempo ga:userTimingVariable

6.3.17.2. Métricas

- Tiempo de usuario (ms) ga:userTimingValue
- Muestra de sincronización del usuario ga:userTimingSample
- Promedio Tiempo de usuario (seg.) ga:avgUserTimingValue

6.3.18. Excepciones

6.3.18.1. Dimensiones

- Descripción de excepción ga:exceptionDescription

6.3.18.2. Métricas

- Excepciones ga:exceptions
- Excepciones / Pantalla ga:exceptionsPerScreenview
- Choques ga:fatalExceptions
- Accidentes / Pantalla ga:fatalExceptionsPerScreenview

6.3.19. Experimentos de contenido

6.3.19.1. Dimensiones

- ID del experimento ga:experimentId
- Variante ga:experimentVariant
- ID de experimento con variante ga:experimentCombination
- Nombre del experimento ga:experimentName

6.3.20. Variables o columnas personalizadas

6.3.20.1. Dimensiones

- Dimensión personalizada XX ga:dimensionXX
- Variable personalizada (clave XX) ga:customVarNameXX
- Variable personalizada (valor XX) ga:customVarValueXX

6.3.20.2. Métricas

- Valor métrico personalizado XX ga:metricXX
- Métrica calculada ga:calcMetric_<NAME>

6.3.21. Tiempo

6.3.21.1. Dimensiones

- Fecha ga:date
- Año ga:year
- Mes del año ga:month
- Semana del año ga:week
- Día del mes ga:day
- Hora ga:hour
- Minuto ga:minute
- Índice de mes ga:nthMonth
- Índice de semana ga:nthWeek
- Índice del día ga:nthDay
- Índice de minutos ga:nthMinute
- Día de la semana ga:dayOfWeek
- Nombre del día de la semana ga:dayOfWeekName
- Hora del día ga:dateHour
- Fecha Hora y Minuto ga:dateHourMinute
- Mes del año ga:yearMonth
- Semana del año ga:yearWeek
- Semana ISO del año ga:isoWeek
- Año ISO ga:isoYear
- Semana ISO del año ISO ga:isoYearIsoWeek

- Índice de horas ga:nthHour

6.3.22. DoubleClick Campaign Manager

6.3.22.1. Dimensiones

- Anuncio CM (modelo GA) ga:dcmClickAd
- ID de anuncio CM (modelo GA) ga:dcmClickAdId
- Tipo de anuncio CM (modelo GA) ga:dcmClickAdType
- ID de tipo de anuncio de CM ga:dcmClickAdTypeId
- Anunciante CM (modelo GA) ga:dcmClickAdvertiser
- ID de anunciante CM (modelo GA) ga:dcmClickAdvertiserId
- Campaña CM (modelo GA) ga:dcmClickCampaign
- ID de campaña CM (modelo GA) ga:dcmClickCampaignId
- ID de creatividad CM (modelo GA) ga:dcmClickCreativeId
- CM Creative (modelo GA) ga:dcmClickCreative
- ID de renderizado CM (modelo GA) ga:dcmClickRenderingId
- Tipo de creatividad CM (modelo GA) ga:dcmClickCreativeType
- ID de tipo de creatividad CM (modelo GA) ga:dcmClickCreativeTypeId
- Versión creativa CM (modelo GA) ga:dcmClickCreativeVersion
- Sitio CM (modelo GA) ga:dcmClickSite
- ID del sitio CM (modelo GA) ga:dcmClickSiteId
- Colocación CM (Modelo GA) ga:dcmClickSitePlacement
- ID de ubicación de CM (modelo GA) ga:dcmClickSitePlacementId
- ID de configuración de CM Floodlight (modelo GA) ga:dcmClickSpotId
- Actividad CM ga:dcmFloodlightActivity
- CM Actividad y Grupo ga:dcmFloodlightActivityAndGroup
- CM Activity Group ga:dcmFloodlightActivityGroup
- ID de grupo de actividad de CM ga:dcmFloodlightActivityGroupId
- ID de actividad de CM ga:dcmFloodlightActivityId
- ID de anunciante de CM ga:dcmFloodlightAdvertiserId
- ID de configuración de CM Floodlight ga:dcmFloodlightSpotId
- Anuncio CM ga:dcmLastEventAd
- ID de anuncio CM (modelo CM) ga:dcmLastEventAdId
- Tipo de anuncio CM (modelo CM) ga:dcmLastEventAdType
- ID de tipo de anuncio CM (modelo CM) ga:dcmLastEventAdTypeId

- Anunciante CM (modelo CM) ga:dcmLastEventAdvertiser
- ID de anunciante CM (modelo CM) ga:dcmLastEventAdvertiserId
- Tipo de atribución CM (modelo CM) ga:dcmLastEventAttributionType
- Campaña CM (Modelo CM) ga:dcmLastEventCampaign
- ID de campaña CM (modelo CM) ga:dcmLastEventCampaignId
- ID de creatividad CM (modelo CM) ga:dcmLastEventCreativeId
- CM Creative (modelo CM) ga:dcmLastEventCreative
- ID de renderizado CM (modelo CM) ga:dcmLastEventRenderingId
- Tipo de creatividad CM (modelo CM) ga:dcmLastEventCreativeType
- ID de tipo de creatividad CM (modelo CM) ga:dcmLastEventCreativeTypeId
- CM Creative Version (Modelo CM) ga:dcmLastEventCreativeVersion
- Sitio CM (Modelo CM) ga:dcmLastEventSite
- ID del sitio CM (modelo CM) ga:dcmLastEventSiteId
- Colocación CM (Modelo CM) ga:dcmLastEventSitePlacement
- ID de ubicación de CM (modelo CM) ga:dcmLastEventSitePlacementId
- ID de configuración de Floodlight CM (modelo CM) ga:dcmLastEventSpotId

6.3.22.2. Métricas

- Conversiones CM ga:dcmFloodlightQuantity
- Ingresos CM ga:dcmFloodlightRevenue
- CM CPC ga:dcmCPC
- CM CTR ga:dcmCTR
- Clics de CM ga:dcmClicks
- Costo CM ga:dcmCost
- Impresiones CM ga:dcmImpressions
- CM ROAS ga:dcmROAS
- CM RPC ga:dcmRPC

6.3.23. Audiencia

6.3.23.1. Dimensiones

- Años ga:userAgeBracket
- Género ga:userGender
- Otra categoría ga:interestOtherCategory
- Categoría de afinidad (alcance) ga:interestAffinityCategory

- Segmento de mercado ga:interestInMarketCategory

6.3.24. Adsense

6.3.24.1. Métricas

- Ingresos de AdSense ga:adsenseRevenue
- Bloques de anuncios de AdSense vistos ga:adsenseAdUnitsViewed
- Impresiones de AdSense ga:adsenseAdsViewed
- Anuncios de AdSense en los que se hizo clic ga:adsenseAdsClicks
- Impresiones de página de AdSense ga:adsensePageImpressions
- CTR de AdSense ga:adsenseCTR
- ECPM de AdSense ga:adsenseECPM
- Salidas de AdSense ga:adsenseExits
- Porcentaje de impresiones visibles de AdSense ga:adsenseViewableImpressionPercent
- Cobertura de AdSense ga:adsenseCoverage

6.3.25. Editor

6.3.25.1. Métricas

- Impresiones de editor ga:totalPublisherImpressions
- Cobertura del editor ga:totalPublisherCoverage
- Vistas de página monetizadas por el editor ga:totalPublisherMonetizedPageviews
- Impresiones del editor / sesión ga:totalPublisherImpressionsPerSession
- % De impresiones visibles del editor ga:totalPublisherViewableImpressionsPercent
- Clics de editor ga:totalPublisherClicks
- CTR de editor ga:totalPublisherCTR
- Ingresos del editor ga:totalPublisherRevenue
- Ingresos del editor / 1000 sesiones ga:totalPublisherRevenuePer1000Sessions
- ECPM del editor ga:totalPublisherECPM

6.3.26. Ad Exchange

6.3.26.1. Métricas

- Impresiones de AdX ga:adxImpressions

- Cobertura de AdX ga:adxCoverage
- Vistas de página monetizadas de AdX ga:adxMonetizedPageviews
- Impresiones de AdX / sesión ga:adxImpressionsPerSession
- Porcentaje de impresiones visibles de AdX ga:adxViewableImpressionsPercent
- Clics de AdX ga:adxClicks
- CTR de AdX ga:adxCTR
- Ingresos de AdX ga:adxRevenue
- Ingresos de AdX / 1000 sesiones ga:adxRevenuePer1000Sessions
- ECPM de AdX ga:adxECPM

6.3.27. Reabastecimiento de DoubleClick para editores

6.3.27.1. Dimensiones

- ID de línea de pedido de GAM ga:dfpLineItemId
- Nombre de elemento de línea GAM ga:dfpLineItemName

6.3.27.2. Métricas

- Impresiones de relleno de GAMg a:backfillImpressions
- Cobertura de relleno de GAM ga:backfillCoverage
- Vistas de página monetizadas de relleno de GAM ga:backfillMonetizedPageviews
- Impresiones de backfill de GAM / sesión ga:backfillImpressionsPerSession
- Relleno de GAM Impresiones visibles% ga:backfillViewableImpressionsPercent
- Clics de relleno de GAM ga:backfillClicks
- CTR de relleno de GAM ga:backfillCTR
- Ingresos de relleno de GAM ga:backfillRevenue
- Ingresos de relleno de GAM / 1000 sesiones ga:backfillRevenuePer1000Sessions
- GAM relleno eCPM ga:backfillECPM

6.3.28. DoubleClick para editores

6.3.28.1. Métricas

- Impresiones GAM ga:dfpImpressions
- Cobertura GAM ga:dfpCoverage
- Vistas de página monetizadas de GAM ga:dfpMonetizedPageviews
- Impresiones de GAM / sesión ga:dfpImpressionsPerSession

- % De impresiones visibles de GAM ga:dfpViewableImpressionsPercent
- Clics de GAM ga:dfpClicks
- GAM CTR ga:dfpCTR
- Ingresos de GAM ga:dfpRevenue
- Ingresos de GAM / 1000 sesiones ga:dfpRevenuePer1000Sessions
- GAM eCPM ga:dfpECPM

6.3.29. Valor del tiempo de vida y cohortes

6.3.29.1. Dimensiones

- Campaña de adquisición ga:acquisitionCampaign
- Medio de adquisición ga:acquisitionMedium
- Fuente de adquisición ga:acquisitionSource
- Fuente de adquisición / Medio ga:acquisitionSourceMedium
- Canal de adquisición ga:acquisitionTrafficChannel
- Grupo ga:cohort
- Día ga:cohortNthDay
- Mes ga:cohortNthMonth
- Semanag a:cohortNthWeek

6.3.29.2. Métricas

- Usuarios ga:cohortActiveUsers
- Vistas de aplicaciones por usuario ga:cohortAppviewsPerUser
- Vistas de aplicaciones por usuario (LTV) ga:cohortAppviewsPerUserWithLifetimeCriteria
- Consecuciones de objetivos por usuario ga:cohortGoalCompletionsPerUser
- Consecuciones de objetivos por usuario (LTV) ga:cohortGoalCompletionsPerUserWithLifetimeCriteria
- Vistas de página por usuario ga:cohortPageviewsPerUser
- Vistas de página por usuario (LTV) ga:cohortPageviewsPerUserWithLifetimeCriteria
- Retención de usuarios ga:cohortRetentionRate
- Ingresos por usuario ga:cohortRevenuePerUser
- Ingresos por usuario (LTV) ga:cohortRevenuePerUserWithLifetimeCriteria
- Duración de la sesión por usuario ga:cohortSessionDurationPerUser

- Duración de la sesión por usuario (LTV) ga:cohortSessionDurationPerUserWithLifetimeCriteria
- Sesiones por usuario ga:cohortSessionsPerUser
- Sesiones por usuario (LTV) ga:cohortSessionsPerUserWithLifetimeCriteria
- Usuarios totales ga:cohortTotalUsers
- Usuarios ga:cohortTotalUsersWithLifetimeCriteria

6.3.30. Agrupación de canales

6.3.30.1. Dimensiones

- Agrupación de canales predeterminada ga:channelGrouping

6.3.31. DoubleClick Bid Manager

6.3.31.1. Dimensiones

- Anunciante DV360 (modelo GA) ga:dbmClickAdvertiser
- ID de anunciante DV360 (modelo GA) ga:dbmClickAdvertiserId
- DV360 Creative ID (modelo GA) ga:dbmClickCreativeId
- DV360 Exchange (Modelo GA) ga:dbmClickExchange
- DV360 ID de intercambio (modelo GA) ga:dbmClickExchangeId
- Orden de inserción de DV360 (modelo GA) ga:dbmClickInsertionOrder
- ID de pedido de inserción DV360 (modelo GA) ga:dbmClickInsertionOrderId
- Nombre de elemento de línea DV360 (modelo GA) ga:dbmClickLineItem
- ID de línea de pedido de DV360 (modelo GA) ga:dbmClickLineItemId
- Sitio DV360 (modelo GA) ga:dbmClickSite
- DV360 ID del sitio (modelo GA) ga:dbmClickSiteId
- Anunciante DV360 (modelo CM) ga:dbmLastEventAdvertiser
- ID de anunciante DV360 (modelo CM) ga:dbmLastEventAdvertiserId
- DV360 Creative ID (modelo CM) ga:dbmLastEventCreativeId
- Intercambio DV360 (Modelo CM) ga:dbmLastEventExchange
- DV360 ID de intercambio (modelo CM) ga:dbmLastEventExchangeId
- Orden de inserción de DV360 (modelo CM) ga:dbmLastEventInsertionOrder
- DV360 ID de pedido de inserción (modelo CM) ga:dbmLastEventInsertionOrderId
- Elemento de línea DV360 (modelo CM) ga:dbmLastEventLineItem
- ID de línea de pedido de DV360 (modelo CM) ga:dbmLastEventLineItemId

- Sitio DV360 (modelo CM) ga:dbmLastEventSite
- DV360 ID del sitio (modelo CM) ga:dbmLastEventSiteId

6.3.31.2. Métricas

- DV360 eCPA ga:dbmCPA
- DV360 eCPC ga:dbmCPC
- DV360 eCPM ga:dbmCPM
- DV360 CTR ga:dbmCTR
- Clics DV360 ga:dbmClicks
- Conversiones DV360 ga:dbmConversions
- DV360 Costo ga:dbmCost
- Impresiones DV360 ga:dbmImpressions
- DV360 ROAS ga:dbmROAS

6.3.32. Búsqueda de DoubleClick

6.3.32.1. Dimensiones

- Grupo de anuncios SA360 ga:dsAdGroup
- ID del grupo de anuncios SA360 ga:dsAdGroupId
- Anunciante SA360 ga:dsAdvertiser
- ID de anunciante SA360 ga:dsAdvertiserId
- Agencia SA360 ga:dsAgency
- ID de la agencia SA360 ga:dsAgencyId
- Campaña SA360 ga:dsCampaign
- ID de campaña SA360 ga:dsCampaignId
- Cuenta del motor SA360 ga:dsEngineAccount
- ID de la cuenta del motor SA360 ga:dsEngineAccountId
- Palabra clave SA360 ga:dsKeyword
- ID de palabra clave SA360 ga:dsKeywordId

6.3.32.2. Métricas

- SA360 CPC ga:dsCPC
- SA360 CTR ga:dsCTR
- Clics SA360 ga:dsClicks
- SA360 Costo ga:dsCost

- Impresiones SA360 ga:dsImpressions
- SA360 Profit ga:dsProfit
- SA360 ROAS ga:dsReturnOnAdSpend
- SA360 RPC ga:dsRevenuePerClick

6.4. Twitter Standard API

A continuación, se puede ver la documentación oficial sobre la Twitter Standard API.

Twitter ofrece tres tipos de APIs con diferentes niveles de acceso a los datos según las necesidades de cada aplicación:

- Estándar: Las API estándar gratuitas son excelentes para comenzar, probar una integración o validar un concepto. Incluye: API básicas y gratuitas, complejidad de consultas básicas y acceso al foro.
- Premium: Las API premium ofrecen acceso escalable a los datos de Twitter para aquellos que buscan crecer, experimentar e innovar Incluye: Acceso escalable a más datos, sandbox gratuito y contratos mensuales flexibles y acceso al foro.
- Empresa: Las API empresariales ofrecen el más alto nivel de acceso y confiabilidad a quienes dependen de los datos de Twitter. Incluye: API de nivel empresarial, paquetes personalizados y contratos anuales y administradores de cuentas y soporte técnico dedicados (Twitter Developers, s.f.).

Tweets	Estándar (gratis)	Premium	Empresa
Publica y participa	✓		
Tweets de búsqueda: 7 días	✓		
Buscar tweets: 30 días		✓	✓
Buscar tweets: archivo completo		✓	✓
Filtrar tweets	✓		✓
Tweets de muestra	✓		✓
Lote de tweets			✓
Mensajes directos	✓		
Cuenta y usuarios	✓	✓	✓
Métrica			✓
API de anuncios	✓		
Herramientas de editor y SDK	✓		

Figura 387. Tabla de características de los tipos de API de Twitter (Twitter Developers, s.f.)

La versión Estándar consiste en REST (Representational State Transfer) APIs y Streaming APIs. La versión Empresa es una suscripción de pago que incluye filtros manguera, búsqueda histórica y APIs de participación para analítica de datos más profunda, con escucha y otras aplicaciones de negocio y empresas. La versión Premium consiste en un pago según su uso y consiste en versiones fiables y asequibles de las APIs de la versión Empresa, permitiendo que el negocio crezca con su uso. Twitter también dispone de otros tipos de APIs como Ads API que obligan a las aplicaciones a ser admitidas para poder utilizarlas.

La API pretende ser un recurso RESTful. Con la excepción de los webhooks Streaming API y Account Activity, las variables de la API de Twitter intentan ajustarse a muchos de los principios de diseño de Representational State Transfer (REST). Las API de Twitter

utilizan el formato de datos JSON para las respuestas (y, en algunos casos, para las solicitudes).

La API está basada en HTTP. La mayoría de los métodos para recuperar datos de la API de Twitter requieren una solicitud GET. Los métodos que envían, cambian o destruyen datos requieren un POST. También se acepta una solicitud DELETE para métodos que destruyen datos. Los métodos API que requieren un método HTTP en particular devolverán un error si no se invocan con el estilo correcto. Los códigos de respuesta HTTP son significativos. Todas las conexiones deben realizarse a través de TLS 1.2.

Los ID de tuit pueden romper Javascript. Se utiliza el campo id_str en lugar de id siempre que esté presente para mantenerse a salvo. Los navegadores web / intérpretes de Javascript / los consumidores de JSON pueden utilizar grandes identificadores basados en números enteros, por lo que se recomienda utilizar la representación de cadena.

Hay límites en la cantidad de llamadas y cambios que se pueden realizar en un día. El uso de la API tiene una tasa limitada , con límites de uso legítimo adicionales basados en cuentas para escribir / crear / eliminar terminales, para proteger Twitter del abuso.

Los parámetros tienen ciertas expectativas. Algunos métodos de API toman parámetros opcionales o obligatorios. Debe tenerse en cuenta al realizar solicitudes con parámetros:

- Los valores de los parámetros deben convertirse a UTF-8 y codificarse en URL .
- El parámetro de página comienza en 1, no en 0.

Cuando se indique, algunos métodos de API devolverán resultados diferentes según los encabezados HTTP enviados por el cliente. Cuando el mismo comportamiento se puede controlar tanto con un parámetro como con un encabezado HTTP, el parámetro tendrá prioridad.

Hay límites de paginación. Los clientes pueden acceder a un máximo teórico de 3200 estados a través de la página y contar los parámetros para los métodos API de REST user_timeline . Otros métodos de línea de tiempo tienen un máximo teórico de 800 estados. Las solicitudes que superen el límite darán como resultado una respuesta con un código de estado de 200 y un resultado vacío en el formato solicitado. Twitter aún mantiene una base de datos de todos los Tuits enviados por un usuario; sin embargo, para garantizar el rendimiento, existe un límite en las llamadas a la API.

Hay bibliotecas de API de Twitter para casi cualquier lenguaje de programación. La comunidad ha creado numerosas bibliotecas API de Twitter (Twitter Developers, s.f.).

6.4.1. Métodos de autentificación

Las API de Twitter manejan enormes cantidades de datos. La forma en que se aseguran de que estos datos están protegidos tanto para los desarrolladores como para los usuarios es mediante la autenticación. Existen algunos métodos de autenticación, cada uno de los cuales se enumera a continuación. La mayoría de los desarrolladores no necesitarán trabajar con los detalles de autenticación, ya que las bibliotecas cliente de Twitter ya implementan el protocolo.

- OAuth 1.0a permite que una aplicación de desarrollador de Twitter autorizada acceda a información de cuenta privada o realice una acción de Twitter en nombre de una cuenta de Twitter.
- OAuth 2.0 Bearer Token permite que una aplicación de desarrollador de Twitter acceda a información disponible públicamente en Twitter.
- Autenticación básica. Muchas de las API empresariales de Twitter requieren el uso de autenticación básica HTTP.

Los métodos más comunes utilizados por Twitter Developer Platform son OAuth 1.0ay OAuth 2.0 Bearer Token.

Algunas diferencias entre los métodos de token de portador de OAuth 1.0ay OAuth 2.0 son:

Caso de uso	**OAuth 1.0a**	**Token de portador de OAuth 2.0**
Buscar Tuits	✓	✓
Extraer cronogramas de usuarios	✓	✓
Obtener datos de tendencias	✓	✓

Publicar, dar “me gusta” o retuiitear un tuitt	✓	
Recuperar la dirección de correo electrónico de un usuario	✓	
Leer o escribir datos de anunciantes	✓	
Claves y tokens requeridos	Claves de API de consumidor + token de acceso y token de acceso secreto	Token de portador
Límites de tarifa	Distinto por usuario y, a veces, aplicación de desarrollador de Twitter	Distinto por aplicación de desarrollador de Twitter

Las claves de API de consumidor y el token de portador de su aplicación, así como su token de acceso personal y el token secreto de acceso se pueden obtener en la sección de aplicaciones para desarrolladores de Twitter que se encuentra en el portal de desarrolladores . Para generar tokens de acceso para un usuario diferente, se deberá utilizar el proceso OAuth de 3 vías (Twitter Developers, s.f.).

6.4.2. Tweet JSON

Todas las API de Twitter que devuelven Tuits proporcionan esos datos codificados mediante la notación de objetos JavaScript (JSON). JSON se basa en pares clave-valor, con atributos con nombre y valores asociados. Estos atributos y su estado se utilizan para describir objetos.

En Twitter servimos muchos objetos como JSON, incluidos Tuits y Usuarios . Todos estos objetos encapsulan atributos centrales que describen el objeto. Cada Tuit tiene un autor, un mensaje, una identificación única, una marca de tiempo de cuándo se publicó y, a veces, metadatos geográficos compartidos por el usuario. Cada usuario tiene un nombre de

Twitter, una identificación, una cantidad de seguidores y, con mayor frecuencia, una biografía de la cuenta.

Con cada Tuit también se generan objetos de "entidad", que son matrices de contenidos comunes de Tuit, como hashtags, menciones, medios y enlaces. Si hay enlaces, la carga útil JSON también puede proporcionar metadatos como la URL completamente desenrollada y el título y la descripción de la página web.

Entonces, además del contenido de texto en sí, un Tuit puede tener más de 150 atributos asociados (Twitter Developers, s.f.).

6.4.2.1. Tuits

Al ingerir datos de Tuit, el objeto principal es el objeto Tuit, que es un objeto principal para varios objetos secundarios. Por ejemplo, todos los Tuits incluyen un objeto Usuario que describe quién fue el autor del Tuit. Si el Tuit está etiquetado geográficamente, se incluirá un objeto "lugar". Cada Tuit incluye un objeto de "entidades" que encapsula matrices de hashtags, menciones de usuarios, URL, etiquetas de efectivo y medios nativos. Si el Tuit tiene algún medio 'adjunto' o 'nativo' (fotos, video, GIF animado), habrá un objeto "extended_entities".

El JSON a continuación ilustra algunos fundamentos estructurales de los objetos Tuit. Aquí se muestran algunos de los atributos principales del Tuit en el 'nivel raíz' (o nivel superior), incluido cuándo se publicó el Tuit, su ID única y el mensaje del Tuit. También en este nivel de raíz hay otros objetos fundamentales (secundarios) como los objetos "usuario" y "entidades".

{

"created_at": "Thu May 10 15:24:15 +0000 2018",

"id_str": "850006245121695744",

"text": "Here is the Tweet message.",

"user": {

},

"place": {

```
 },
 "entities": {
 },
 "extended_entities": {
 }
}
```

6.4.2.2. Tuits Extendidos

El JSON que describe los tuits extendidos se introdujo cuando se lanzaron Tuits de 280 caracteres en noviembre de 2017. Tweet JSON se amplió para encapsular estos mensajes más largos, sin romper los miles de aplicaciones que analizan estos objetos fundamentales de Twitter. Para proporcionar compatibilidad total con versiones anteriores, se conservaron el campo de 'texto' original de 140 caracteres y los objetos de entidad analizados a partir de él. En el caso de Tuits de más de 140 caracteres, este campo de 'texto' de nivel raíz se truncará y, por lo tanto, quedará incompleto. Dado que los objetos de 'entidades' de nivel raíz contienen matrices de metadatos clave analizados a partir del mensaje de 'texto', como hashtags y enlaces incluidos, estas colecciones estarían incompletas. Por ejemplo, si un mensaje de Tuit tenía 200 caracteres de longitud, con un hashtag incluido al final, el nivel de raíz heredado 'entidades.hashtags'

Se introdujo un nuevo campo 'extended_tweet' para contener los mensajes de Tuit más largos y los metadatos completos de la entidad. El objeto "extended_tweet" proporciona el campo "full_text" que contiene el mensaje de Tuit completo y sin truncar cuando tiene más de 140 caracteres. El objeto "extended_tweet" también contiene un objeto "entidades" con matrices completas de hashtags, enlaces, menciones, etc.

Los Tuits extendidos se identifican con un booleano "truncado" a nivel de raíz. Cuando es verdadero ("truncado": verdadero), los campos "extended_tweet" deben analizarse en lugar de los campos de nivel raíz.

Hay que tener en cuenta que en el ejemplo JSON que se muestra a continuación, el campo "texto" de nivel raíz está truncado y la matriz "entidades.hashtags" de nivel raíz está vacía, aunque el mensaje del Tuit incluye tres hashtags. Dado que se trata de un Tuit extendido, el campo "truncado" se establece en verdadero, y el objeto "extended_tweet" proporciona metadatos completos de Tuit "full_text" y "entidades".

{

"created_at": "Thu May 10 17:41:57 +0000 2018",

"id_str": "994633657141813248",

"text": "Just another Extended Tweet with more than 140 characters, generated as a documentation example, showing that [\"tru… https://t.co/U7Se4NM7Eu",

"display_text_range": [0, 140],

"truncated": true,

"user": {

"id_str": "944480690",

"screen_name": "FloodSocial"

},

"extended_tweet": {

"full_text": "Just another Extended Tweet with more than 140 characters, generated as a documentation example, showing that [\"truncated\": true] and the presence of an \"extended_tweet\" object with complete text and \"entities\" #documentation #parsingJSON #GeoTagged https://t.co/e9yhQTJSIA",

"display_text_range": [0, 249],

"entities": {

"hashtags": [{

"text": "documentation",

"indices": [211, 225]

}, {

"text": "parsingJSON",

"indices": [226, 238]

}, {

```
                    "text": "GeoTagged",
                    "indices": [239, 249]
                }]
            }
        },
        "entities": {
            "hashtags": []
        }
}
```


6.4.2.3. Retuits y citas

Si se está trabajando con objetos Retuit o Cita, entonces esa carga útil JSON contendrá varios objetos Tuit, y cada objeto Tuit contendrá su propio objeto Usuario. El objeto de nivel raíz contendrá información sobre el tipo de acción realizada, es decir, si se trata de un retuit o un tuit de cotización, y contendrá un objeto que describa el tuit "original" que se está compartiendo.

6.4.2.4. Retuits

Los retuits siempre contienen dos objetos Tuit. El Tuit 'original' que se está retuiteando se proporciona en un objeto "retweeted_status". El objeto de nivel raíz encapsula el propio Retuit, incluido un objeto Usuario para la cuenta que realiza la acción Retuit y la hora del Retuit. Retuitear es una acción para compartir un Tuit con los seguidores y no se puede agregar ningún otro contenido nuevo. Además, no se puede proporcionar una (nueva) ubicación con un Retuit. Si bien el Tuit 'original' puede tener geoetiquetado, los objetos Retuit "geo" y "place" siempre serán nulos.

Incluso antes de la introducción de los Tuits extendidos, el objeto "entidades" de nivel raíz estaba en algunos casos truncado e incompleto debido a que la cadena "RT @nombredeusuario" se adjuntaba al mensaje del Tuit que se estaba retuiteando. Hay que tener en cuenta que, si un Retuit es Retuit, el "retweet_status" seguirá apuntando al Tuit original, lo que significa que el Retuit intermedio no está incluido. Se observa un

comportamiento similar cuando se usa twitter.com para "mostrar" un Retuit. Si se copia el ID de Tuit único asignado a la 'acción' Retuit, se muestra el Tuit original.

A continuación, se muestra una estructura de ejemplo para un Retuit. Nuevamente, al analizar los Retuits, es clave analizar el objeto "retweeted_status" para ver el mensaje de Tuit completo (original) y los metadatos de la entidad.


```
{
  "tweet": {
    "text": "RT @author original message"
    "user": {
         "screen_name": "Retweeter"
    },
    "retweeted_status": {
      "text": "original message".
        "user": {
           "screen_name": "OriginalTweeter"
        },
        "place": {
        },
        "entities": {
        },
        "extended_entities": {
        }
      },
    },
    "entities": {
```

```
    },
    "extended_entities": {
    }
  }
}
```

6.4.2.5. Citas

Los Tuits con citas son muy parecidos a los Retuits, excepto que incluyen un nuevo mensaje de Tuit. Estos nuevos mensajes pueden contener su propio conjunto de hashtags, enlaces y otros metadatos de "entidades". Los Tuits de Cita también pueden incluir información de ubicación compartida por el usuario que publica el Tuit de cotización, junto con medios como GIF, videos y fotos.

Tuits de tipo Cita contendrán al menos dos objetos Tuit y, en algunos casos, tres. El Tuit citado, que en sí mismo puede ser un Tuit citado, se proporciona en un objeto "quoted_status". El objeto de nivel raíz encapsula el Tuit de cotización en sí, incluido un objeto de usuario para la cuenta que realiza la acción de compartir y la hora del Tuit de cotización.

Hay que tener en cuenta que Quote Tweets ahora puede tener fotos, GIF o videos, agregados a ellos usando la interfaz de usuario 'publicar Tuit'. Cuando se incluyen enlaces a medios alojados externamente en el mensaje Quote Tweet, el nivel raíz "entity.urls" los describirá. Los medios adjuntos a los Tuits de cotización aparecerán en los metadatos de nivel raíz "extended_entities".

Cuando se lanzaron los Tuits de Cita por primera vez, se adjuntó un enlace abreviado (URL t.co) al mensaje del Tuit 'original' y se proporcionó en el campo de "texto" de nivel raíz. Además, los metadatos para esa URL de t.co se incluyeron en la matriz de nivel raíz 'entity.urls'. Sin embargo, a mediados de junio de 2020 se realizó una actualización mediante la cual estos detalles dejaron de estar disponibles en la respuesta JSON. En primer lugar, la URL t.co abreviada del Tuit citado no se incluirá en el campo "texto" de nivel raíz. En segundo lugar, los metadatos del Tuit citado no se incluirán en los metadatos "entity.urls". En cambio, los metadatos de URL para el Tuit citado estarán en un nuevo

objeto "quoted_status_permalink" en el nivel raíz (o nivel superior), por lo que en el mismo nivel del objeto "quoted_status".

A continuación, se muestra una estructura de ejemplo para un Tuit de Cita con el formato final. Hay que tener en cuenta que el atributo "texto" de nivel raíz se basa en el mensaje del Tuit citado, más una URL abreviada de Twitter al Tuit citado.


```
{
  "text": "My added comments to this Tweet",
  "user": {
    "screen_name": "TweetQuoter"
  },
  "quoted_status": {
    "text": "original message",
    "user": {
      "screen_name": "OriginalTweeter"
    },
    "place": {
    },
    "entities": {
    },
    "extended_entities": {
    }
  },
  "quoted_status_permalink": {
     "url": "https:\/\/t.co\/LinkToTweet",
     "expanded": "https:\/\/twitter.com\/OriginalTweeter\/status\/994281226797137920",
```

```
    "display": "twitter.com\/OriginalTweeter\/status\/994281226797137920"
  },
  "place": {
  },
  "entities": {
    "urls": [
    ]
  }
}
```


6.4.2.6. Buenas prácticas

- Twitter JSON está codificado con caracteres UTF-8.
- Los analizadores deben tolerar la variación en el orden de los campos con facilidad. Se debe suponer que Tweet JSON se sirve como un hash de datos desordenado.
- Los analizadores deben tolerar la adición de campos "nuevos". La plataforma de Twitter ha evolucionado continuamente desde 2006, por lo que existe una larga historia de nuevos metadatos agregados a los Tuits .
- Los analizadores JSON deben ser tolerantes con los campos "faltantes", ya que no todos los campos aparecen en todos los contextos.
- Por lo general, es seguro considerar un campo nulo, un conjunto vacío y la ausencia de un campo como lo mismo.

6.4.3. Objeto Tweet

Los tuits son el bloque de construcción atómico básico de todo lo relacionado con Twitter. Los tuits también se conocen como "actualizaciones de estado". El objeto Tuit tiene una larga lista de atributos " de nivel raíz, incluyendo atributos fundamentales tales como id, created_at, y text. Los objetos Tuit también son el objeto 'principal' de varios objetos secundarios. Tuit objetos hijo incluyen user, entities y extended_entities. Los tuits con etiquetas geográficas tendrán un objeto place secundario (Twitter Developers, s.f.).

6.4.3.1. Diccionario de datos de tuits

A continuación, se encuentra el diccionario de datos para estos atributos de 'nivel raíz', así como enlaces a diccionarios de datos de objetos secundarios.

Atributo	Tipo	Descripción
created_at	Cadena	Hora UTC cuando se creó este Tuit. Ejemplo: "created_at": "Mié 10 de octubre 20:19:24 +0000 2018"
id	Entero 64	La representación entera del identificador único de este Tuit. Este número es superior a 53 bits y algunos lenguajes de programación pueden tener dificultades / defectos silenciosos para interpretarlo. Usar un entero de 64 bits con signo para almacenar este identificador es seguro. Usar id_str para buscar el identificador para estar seguro. Ejemplo: "id": 1050118621198921728
id_str	Cadena	Representación de cadena del identificador único de este Tuit. Las implementaciones deberían usar esto en lugar del entero grande en id. Ejemplo: "id_str": "1050118621198921728"
text	Cadena	El texto UTF-8 real de la actualización de estado. Ejemplo: "text":"To make room for more expression, we will now count all emojis as equal—including those with gender and skin t… https://t.co/MkGjXf9aXm"

source	Cadena	Utilidad utilizada para publicar el Tuit, como una cadena con formato HTML. Los tuits del sitio web de Twitter tienen un valor fuente de web. Ejemplo: "source":"Twitter Web Client"
truncated	Booleano	Indica si el valor del parámetro text se truncó, por ejemplo, como resultado de un retuit que excede el límite de longitud del texto del Tuit original de 140 caracteres. El texto truncado terminará en puntos suspensivos, como este. ...Dado que Twitter ahora rechaza los Tuits largos en lugar de truncarlos, la gran mayoría de los Tuits tendrán este ajuste en false. Tenga en cuenta que si bien los retuits nativos pueden tener su propiedad text de nivel superior acortada, el texto original estará disponible debajo del objeto retweeted_status y el parámetro truncated se establecerá en el valor del estado original (en la mayoría de los casos, false). Ejemplo: "truncated": true
in_reply_to_status_id	Entero 64	Puede ser nulo. Si el Tuit representado es una respuesta, este campo contendrá la representación entera de la ID del Tuit original. Ejemplo: "in_reply_to_status_id": 1051222721923756032
in_reply_to_status_id_str	Cadena	Puede ser nulo. Si el Tuit representado es una respuesta, este campo contendrá la representación de cadena del ID del Tuit original. Ejemplo:

		"in_reply_to_status_id_str": "1051222721923756032"
in_reply_to_user_id	Entero 64	Puede ser nulo. Si el Tuit representado es una respuesta, este campo contendrá la representación entera del ID de autor del Tuit original. No siempre será necesariamente el usuario mencionado directamente en el Tuit. Ejemplo: "in_reply_to_user_id": 6253282
in_reply_to_user_id_str	Cadena	Puede ser nulo. Si el Tuit representado es una respuesta, este campo contendrá la representación de cadena del ID del autor del Tuit original. No siempre será necesariamente el usuario mencionado directamente en el Tuit. Ejemplo: "in_reply_to_user_id_str": "6253282"
in_reply_to_screen_name	Cadena	Puede ser nulo. Si el Tuit representado es una respuesta, este campo contendrá el nombre de pantalla del autor del Tuit original. Ejemplo: "in_reply_to_screen_name": "twitterapi"

user	Objeto de usuario	El usuario que publicó este Tuit. Ejemplo resaltando atributos seleccionados: { "user": { "id": 6253282, "id_str": "6253282", "name": "Twitter API", "screen_name": "TwitterAPI", "location": "San Francisco, CA", "url": "https://developer.twitter.com", "description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform. Don't get an answer? It's on my website.", "verified": true, "followers_count": 6129794, "friends_count": 12, "listed_count": 12899, "favourites_count": 31, "statuses_count": 3658, "created_at": "Wed May 23 06:01:13 +0000 2007", "utc_offset": null, "time_zone": null, "geo_enabled": false, "lang": "en",

```
    "contributors_enabled": false,
    "is_translator": false,
    "profile_background_color": "null",
    "profile_background_image_url": "null",
    "profile_background_image_url_https": "null",
    "profile_background_tile": null,
    "profile_link_color": "null",
    "profile_sidebar_border_color": "null",
    "profile_sidebar_fill_color": "null",
    "profile_text_color": "null",
    "profile_use_background_image": null,
    "profile_image_url": "null",
    "profile_image_url_https":
"https://pbs.twimg.com/profile_images/9428584795925
54497/BbazLO9L_normal.jpg",
    "profile_banner_url":
"https://pbs.twimg.com/profile_banners/6253282/14974
91515",
    "default_profile": false,
    "default_profile_image": false,
    "following": null,
    "follow_request_sent": null,
    "notifications": null
  }
```

```
}
```

coordinates	Coordenadas	Puede ser nulo. Representa la ubicación geográfica de este Tuit según lo informado por el usuario o la aplicación cliente. La matriz de coordenadas internas tiene el formato geoJSON (primero la longitud, luego la latitud). Ejemplo:
		"coordinates":
		{
		"coordinates":
		[
		-75.14310264,
		40.05701649
		],
		"type": "Point"
		}

place	Objeto Lugares	Puede ser nulo. Cuando está presente, indica que el tuit está asociado (pero no necesariamente se origina en) un lugar . Ejemplo: "place": { "attributes":{}, "bounding_box": { "coordinates": [[[-77.119759,38.791645],

[-76.909393,38.791645],

[-76.909393,38.995548],

[-77.119759,38.995548]

]],

"type":"Polygon"

},

"country":"United States",

"country_code":"US",

"full_name":"Washington, DC",

"id":"01fbe706f872cb32",

"name":"Washington",

"place_type":"city",

"url":"http://api.twitter.com/1/geo/id/0172cb32.json"

}

quoted_status_id	Entero 64	Este campo solo aparece cuando el Tuit es una cita. Este campo contiene el valor entero ID de Tuit del Tuit citado. Ejemplo: "quoted_status_id": 1050119905717055488
quoted_status_id_str	Cadena	Este campo solo aparece cuando el Tuit es una cita. Esta es la ID del Tuit de representación de cadena del Tuit citado. Ejemplo: "quoted_status_id_str": "1050119905717055488"

is_quote_status	Booleano	Indica si se trata de un Tuit citado. Ejemplo: "is_quote_status": false
quoted_status	Tuit	Este campo solo aparece cuando el Tuit es una cita. Este atributo contiene el objeto Tuit del Tuit original que se citó.
retweeted_status	Tuit	Los usuarios pueden amplificar la transmisión de Tuits creados por otros usuarios retuiteando . Los retuits se pueden distinguir de los tuits típicos por la existencia de un atributo retweeted_status. Este atributo contiene una representación del Tuit original que se retuiteó. Tenga en cuenta que los retuits de retuits no muestran representaciones del retuit intermediario, sino solo el Tuit original. (Los usuarios también pueden cancelar un retuit que crearon eliminando su retuit).
quote_count	Entero	Puede ser nulo. Indica aproximadamente cuántas veces este Tuit ha sido citado por usuarios de Twitter. Ejemplo: "quote_count": 33 Nota: este objeto solo está disponible con los productos de nivel Premium y Enterprise.
reply_count	Entero	Número de veces que se ha respondido a este Tuit. Ejemplo: "reply_count": 30 Nota: este objeto solo está disponible con los productos de nivel Premium y Enterprise.

retweet_count	Entero	Número de veces que se ha retuiteado este Tuit. Ejemplo: "retweet_count": 160
favorite_count	Entero	Anulable. Indica aproximadamente cuántas veces los usuarios de Twitter le han dado “me gusta” a este Tuit . Ejemplo: "favorite_count": 295

entities	Entidades	Entidades que se han analizado del texto del Tuit. Ejemplo:
		"entities":
		{
		"hashtags":[],
		"urls":[],
		"user_mentions":[],
		"media":[],
		"symbols":[]
		"polls":[]
		}

extended_entities	Entidades extendidas	Cuando hay entre una y cuatro fotos nativas o un video o un GIF animado en Tuit, contiene una matriz de metadatos de 'medios'. Esto también está disponible en Citas de Tuits. Ejemplo:
		"entities":
		{
		"media":[]
		}

favorited	Booleano	Puede ser nulo. Indica si este Tuit le ha gustado al usuario que se autentica. Ejemplo: "favorited":true
retweeted	Booleano	Indica si este Tuit ha sido retuiteado por el usuario que se autentica. Ejemplo: "retweeted":false
possibly_sensitive	Booleano	Puede ser nulo. Este campo solo aparece cuando un Tuit contiene un enlace. El significado del campo no se refiere al contenido del Tuit en sí, sino que es un indicador de que la URL contenida en el Tuit puede contener contenido o medios identificados como contenido sensible. Ejemplo: "possibly_sensitive": false
filter_level	Cadena	Indica el valor máximo del parámetro filter_level que se puede utilizar y seguir transmitiendo este Tuit. Por lo que un valor de medium será transmitido en none, low y medium arroyos. Ejemplo:

		"filter_level": "low"
lang	Cadena	Anulable. Cuando está presente, indica un identificador de idioma BCP 47 correspondiente al idioma detectado por la máquina del texto del Tuit, o und si no se pudo detectar ningún idioma. Ejemplo: "lang": "en"
matching_rules	Matriz de objetos de regla	Presente en productos filtrados como Twitter Search y PowerTrack. Proporciona la identificación y la etiqueta asociadas con la regla que coincidió con el Tuit. Con PowerTrack, más de una regla puede coincidir con un Tuit. Ejemplo: "matching_rules": " [{ "tag": "twitterapi emojis", "id": 1050118621198921728, "id_str": "1050118621198921728" }]"

6.4.3.2. Atributos adicionales de Tuit

Las API de Twitter que proporcionan Tuits (por ejemplo, los estados GET / método de búsqueda) pueden incluir estos atributos de Tuit adicionales:

Atributo	**Tipo**	**Descripción**
current_user_retweet	Objeto	Perspectival Only aparece en los métodos que admiten el parámetro include_my_retweet, cuando se establece en

		true. Detalla el ID de Tuit del propio retuit del usuario (si existe) de este Tuit. Ejemplo:
		"current_user_retweet": {
		"id": 6253282,
		"id_str": "6253282"
		}

scopes	Objeto	Un conjunto de pares clave-valor que indica la entrega contextual prevista del Tuit que lo contiene. Utilizado actualmente por los productos promocionados de Twitter. Ejemplo: "scopes":{"followers":false}
withheld_copyright	Booleano	Cuando está presente y se establece en "verdadero", indica que este contenido se ha retenido debido a una queja de la DMCA . Ejemplo: "withheld_copyright": true
withheld_in_countries	Matriz de cadena	Cuando está presente, indica una lista de códigos de país de dos letras en mayúsculas de los que se oculta este contenido. Twitter admite los siguientes valores no nacionales para este campo: "XX": el contenido se retiene en todos los países. "XY": el contenido se retiene debido a una solicitud de la DMCA. Ejemplo: "withheld_in_countries": ["GR", "HK", "MY"]
withheld_scope	Cadena	Cuando está presente, indica si el contenido que se retiene es el "estado" o un "usuario".

Ejemplo:

"withheld_scope": "status"

6.4.3.3. Atributos obsoletos

Campo	Tipo	Descripción
geo	Objeto	Obsoleto. Anulable. En su lugar, utilizar el coordinates campo. Este atributo obsoleto tiene sus coordenadas formateadas como [lat, long] , mientras que el resto de Tuit geo tiene el formato [long, lat] .

6.4.4. Objeto Usuario

El objeto Usuario contiene metadatos de la cuenta de usuario de Twitter que describen al usuario de Twitter al que se hace referencia. Los usuarios pueden crear tuits, retuitear, citar tuits de otros usuarios, responder tuits, seguir a usuarios, ser @mencionados en tuits y pueden agruparse en listas.

El objeto Tuit también contendrá objetos de Usuario de los Usuarios involucrados dentro de un Tuit. En el caso de Retuits y Tuits citados, el objeto user de nivel superior representa qué cuenta tomó esa acción, y la carga útil JSON incluirá un segundo user dentro retweeted_status de la cuenta que creó el Tuit original. Los objetos de usuario se pueden retirar utilizando ido screen_name.

En general, estos valores de metadatos user son relativamente constantes. Algunos campos nunca cambian, como el del usuario id(proporcionado como una cadena id_str) y cuándo se creó la cuenta. Otros metadatos en ocasiones pueden cambiar, como screen_name, description y location. Algunos metadatos cambian con frecuencia, como el número de Tuits que ha publicado la cuenta statuses_count y su número de seguidores followers_count (Twitter Developers, s.f.).

6.4.4.1. Diccionario de datos de usuario

Atributo	Tipo	Descripción
id	Entero 64	La representación entera del identificador único para este usuario. Este número es superior a 53 bits y algunos lenguajes de programación pueden tener dificultades / defectos silenciosos para interpretarlo. Usar un entero de 64 bits con signo para almacenar este identificador es seguro. Usar id_str para buscar el identificador para estar seguro. Ejemplo: "id": 6253282
id_str	Cadena	Representación de cadena del identificador único de este usuario. Las implementaciones deberían usar esto en lugar del entero grande, posiblemente no consumible en id. Ejemplo: "id_str": "6253282"
name	Cadena	El nombre del usuario, tal como lo han definido. No necesariamente el nombre de una persona. Por lo general, tiene un límite de 50 caracteres, pero está sujeto a cambios. Ejemplo: "name": "Twitter API"
screen_name	Cadena	El nombre de pantalla, identificador o alias con el que este usuario se identifica. Los nombres de pantalla son únicos, pero están sujetos a cambios. Usar id_str como identificador de usuario siempre que sea posible. Por lo general, un máximo de 15 caracteres, pero algunos relatos históricos pueden existir con nombres más largos. Ejemplo:

		"screen_name": "twitterapi"
location	Cadena	Puede ser nulo . La ubicación definida por el usuario para el perfil de esta cuenta. No es necesariamente una ubicación, ni se puede analizar por máquina. En ocasiones, el servicio de búsqueda interpretará este campo de forma imprecisa. Ejemplo: "location": "San Francisco, CA"
derived	Matrices de objetos enriquecidos	Solo API empresariales Recopilación de metadatos de enriquecimiento derivados para el usuario. Proporciona los metadatos de Enriquecimiento geográfico del perfil . Ejemplo: "derived":{"locations": [{"country":"United States","country_code":"US","locality":"Denver"}]}
url	Cadena	Puede ser nulo. Una URL proporcionada por el usuario en asociación con su perfil. Ejemplo: "url": "https://developer.twitter.com"
description	Cadena	Puede ser nulo. La cadena UTF-8 definida por el usuario que describe su cuenta. Ejemplo: "description": "The Real Twitter API."
protected	Cadena	Cuando es verdadero, indica que este usuario ha elegido proteger sus Tuits. Ejemplo: "protected": true

verified	Booleano	Cuando es verdadero, indica que el usuario tiene una cuenta verificada. Ejemplo: "verified": false
followers_count	Entero	El número de seguidores que tiene esta cuenta actualmente. Bajo ciertas condiciones de coacción, este campo indicará temporalmente "0". Ejemplo: "followers_count": 21
friends_count	Entero	El número de usuarios que sigue esta cuenta (también conocido como sus "seguidores"). Bajo ciertas condiciones de coacción, este campo indicará temporalmente "0". Ejemplo: "friends_count": 32
listed_count	Entero	El número de listas públicas de las que este usuario es miembro. Ejemplo: "listed_count": 9274
favourites_count	Entero	La cantidad de Tuits que le han gustado a este usuario durante la vida de la cuenta. Ortografía británica utilizada en el nombre del campo por razones históricas. Ejemplo: "favourites_count": 13
statuses_count	Entero	El número de Tuits (incluidos los retuits) emitidos por el usuario. Ejemplo: "statuses_count": 42
created_at en	Cadena	La fecha y hora UTC en que se creó la cuenta de usuario en Twitter. Ejemplo:

		"created_at": "Mon Nov 29 21:18:15 +0000 2010"
profile_banner_url	Cadena	La URL basada en HTTPS que apunta a la representación web estándar del banner de perfil subido por el usuario. Al agregar un elemento de ruta final de la URL, es posible obtener diferentes tamaños de imagen optimizados para pantallas específicas. Ejemplo: "profile_banner_url": "https://si0.twimg.com/profile_banners/819797/1348102824"
profile_image_url_https	Cadena	Una URL basada en HTTPS que apunta a la imagen de perfil del usuario. Ejemplo: "profile_image_url_https": "https://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png"
default_profile	Booleano	Cuando es verdadero, indica que el usuario no ha modificado el tema o el fondo de su perfil de usuario. Ejemplo: "default_profile": false
default_profile_image	Booleano	Cuando es verdadero, indica que el usuario no ha subido su propia imagen de perfil y se usa una imagen predeterminada en su lugar. Ejemplo: "default_profile_image": false

withheld_in_countries	Matriz de cadena	Cuando está presente, indica una lista de códigos de país de dos letras en mayúsculas de los que se oculta este contenido. Twitter admite los siguientes valores no nacionales para este campo: "XX": el contenido se retiene en todos los países. "XY": el contenido se retiene debido a una solicitud de la DMCA. Ejemplo: "withheld_in_countries": ["GR", "HK", "MY"]
withheld_scope	Cadena	Cuando está presente, indica que el contenido que se retiene es un "usuario". Ejemplo: "withheld_scope": "user"

6.4.4.2. Atributos ya no admitidos (obsoletos)

Campo	**Tipo**	**Descripción**
utc_offset	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET account/settings
time_zone	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET account/settings
lang	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET account/settings como idioma

geo_enabled	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET account/settings. Este campo debe ser verdadero para que el usuario actual adjunte datos geográficos al usar estados / actualizaciones de POST
following	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET friendships/lookup
follow_request_sent	nulo	El valor se establecerá en nulo. Todavía disponible a través de GET friendships/lookup
has_extended_profile	nulo	En desuso. El valor se establecerá en nulo.
notifications	nulo	En desuso. El valor se establecerá en nulo.
profile_location	nulo	En desuso. El valor se establecerá en nulo.
contributors_enabled	nulo	En desuso. El valor se establecerá en nulo.
profile_image_url	nulo	En desuso. El valor se establecerá en nulo. NOTA: Las imágenes de perfil solo están disponibles mediante el campo profile_image_url_https.

profile_background_color	nulo	En desuso. El valor se establecerá en nulo.
profile_background_image_url	nulo	En desuso. El valor se establecerá en nulo.
profile_background_image_url_https	nulo	En desuso. El valor se establecerá en nulo.
profile_background_tile	nulo	En desuso. El valor se establecerá en nulo.
profile_link_color	nulo	En desuso . El valor se establecerá en nulo.
profile_sidebar_border_color	nulo	En desuso. El valor se establecerá en nulo.
profile_sidebar_fill_color	nulo	En desuso. El valor se establecerá en nulo.
profile_text_color	nulo	En desuso. El valor se establecerá en nulo.
profile_use_background_image	nulo	En desuso. El valor se establecerá en nulo.
is_translator	nulo	En desuso. El valor se establecerá en nulo.
is_translation_enabled	nulo	En desuso. El valor se establecerá en nulo.

translator_type	nulo	En desuso. El valor se establecerá en nulo.

6.4.4.3. Objeto de usuario de ejemplo


```
{
	"id": 6253282,
	"id_str": "6253282",
	"name": "Twitter API",
	"screen_name": "TwitterAPI",
	"location": "San Francisco, CA",
	"profile_location": null,
	"description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform. Don't get an answer? It's on my website.",
	"url": "https:\/\/t.co\/8IkCzCDr19",
	"entities": {
		"url": {
			"urls": [{
				"url": "https:\/\/t.co\/8IkCzCDr19",
				"expanded_url": "https:\/\/developer.twitter.com",
				"display_url": "developer.twitter.com",
				"índices": [
					0,
					23
				]
			}]
```

```
            },
            "description": {
                "urls": []
            }
        },
        "protected": false,
        "followers_count": 6133636,
        "friends_count": 12,
        "listed_count": 12936,
        "created_at": "Wed May 23 06:01:13 +0000 2007",
        "favourites_count": 31,
        "utc_offset": null,
        "time_zone": null,
        "geo_enabled": null,
        "verified": true,
        "statuses_count": 3656,
        "lang": null,
        "contributors_enabled": null,
        "is_translator": null,
        "is_translation_enabled": null,
        "profile_background_color": null,
        "profile_background_image_url": null,
        "profile_background_image_url_https": null,
        "profile_background_tile": null,
```

"profile_image_url": null,

"profile_image_url_https":
"https:\/\/pbs.twimg.com\/profile_images\/942858479592554497\/BbazLO9L_normal.jpg",

"profile_banner_url": null,

"profile_link_color": null,

"profile_sidebar_border_color": null,

"profile_sidebar_fill_color": null,

"profile_text_color": null,

"profile_use_background_image": null,

"has_extended_profile": null,

"default_profile": false,

"default_profile_image": false,

"following": null,

"follow_request_sent": null,

"notifications": null,

"translator_type": null

}

6.4.5. Objeto Entidades

Las entidades proporcionan metadatos e información contextual adicional sobre el contenido publicado en Twitter. La sección entities proporciona conjuntos de cosas comunes incluidas en los Tuits: hashtags, menciones de usuarios, enlaces, tickers de acciones (símbolos), encuestas de Twitter y medios adjuntos. Estas matrices son convenientes para los desarrolladores al consumir Tuits, ya que Twitter esencialmente ha procesado previamente o analizado previamente el cuerpo del texto. En lugar de tener que buscar y encontrar explícitamente estas entidades en el cuerpo del Tuit, su analizador puede ir directamente a esta sección JSON y ahí están.

Más allá de proporcionar conveniencias de análisis, la sección entities también proporciona metadatos útiles de 'valor agregado'. Por ejemplo, si está utilizando el enriquecimiento de URL mejoradas, los metadatos de URL incluyen URL completamente expandidas, así como títulos y descripciones de sitios web asociados. Otro ejemplo es cuando hay menciones de usuarios, los metadatos de las entidades incluyen el ID de usuario numérico, que son útiles al realizar solicitudes a muchas API de Twitter.

Cada carga útil Tweet JSON incluye una sección entities, con el conjunto mínimo de hashtags, urls, user_mentions, y symbolsatributos, aunque ninguna de esas entidades son parte del mensaje Tuit. Por ejemplo, si examina el JSON en busca de un Tuit con un cuerpo de "¡Hola mundo!" y sin medios adjuntos, Tweet JSON incluirá el siguiente contenido con matrices de entidades que contienen cero elementos:

```
"entities": {
    "hashtags": [
    ],
    "urls": [
    ],
    "user_mentions": [
    ],
    "symbols": [
    ]
  }
```

Las entidades de medios y encuestas solo aparecerán cuando ese tipo de contenido sea parte del Tuit. Y si se trabaja con medios nativos (fotos, videos o GIF), el objeto Entidades extendidas es el que las incluirá.

Las secciones entities y extended_entities están formadas por matrices de objetos de entidad. A continuación, se añaden descripciones para cada uno de estos objetos de entidad, incluidos los diccionarios de datos que describen los nombres de los atributos del objeto, los tipos y una breve descripción. También se indicará qué operadores de

PowerTrack coinciden con estos atributos e incluiremos algunas cargas útiles JSON de muestra.

Una colección de entidades comunes que se encuentran en los Tuits, incluidos hashtags, enlaces y menciones de usuarios. Este objeto entities incluye un atributo media, pero su implementación en la sección entities solo es completamente precisa para Tuits con una sola foto. Para todos los Tuits con más de una foto, un video o un GIF animado, el lector es dirigido a la sección extended_entities (Twitter Developers, s.f.).

6.4.5.1. Diccionario de datos de entidades

El objeto de las entidades es un contenedor de matrices de otros subobjetos de entidad. Después de ilustrar la estructura entities, se proporcionarán diccionarios de datos para estos subobjetos y los operadores que los coinciden.

Campo	Tipo	Descripción
hashtags	Matriz de objetos Hashtag	Representa hashtags que se han analizado del texto del Tuit. Ejemplo:

```
{
  "hashtags": [
    {
      "indices": [
        32,
        38
      ],
      "text": "nodejs"
    }
  ]
```

media	Matriz de objetos multimedia	Representa elementos multimedia subidos con el Tuit. Ejemplo:
		{
		"media": [
		{
		"display_url": "pic.twitter.com/5J1WJSRCy9",
		"expanded_url": "https://twitter.com/nolan_test/status/930077847535812610/photo/1",
		"id": 9.300778475358126e17,
		"id_str": "930077847535812610",
		"indices": [
		13,
		36
		],
		"media_url": "http://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg",
		"media_url_https": "https://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg"
		"sizes": {
		"thumb": {
		"h": 150,
		"resize": "crop",
		"w": 150
		},

```
            "large": {
                "h": 1366,
                "resize": "fit",
                "w": 2048
            },
            "medium": {
                "h": 800,
                "resize": "fit",
                "w": 1200
            },
            "small": {
                "h": 454,
                "resize": "fit",
                "w": 680
            }
        },
        "type": "photo",
        "url": "https://t.co/5J1WJSRCy9",
      }
    ]
  }
```

urls	Matriz de objetos URL	Representa las URL incluidas en el texto de un Tuit.
		Ejemplo (sin el enriquecimiento de URL mejorado habilitado):
		{
		"urls": [
		{
		"indices": [
		32,
		52
		],
		"url": "http://t.co/IOwBrTZR",
		"display_url": "youtube.com/watch?v=oHg5SJ…",
		"expanded_url": "http://www.youtube.com/watch?v=oHg5SJYRHA0"
		}
		]
		}
		Ejemplo (con el enriquecimiento de URL mejorado habilitado):
		{"urls": [
		{
		"url": "https://t.co/D0n7a53c2l",
		"expanded_url": "http://bit.ly/18gECvy",
		"display_url": "bit.ly/18gECvy",
		"unwound": {

```
        "url":
"https://www.youtube.com/watch?v=oHg5SJYRHA0",
        "status": 200,
        "title": "RickRoll'D",
        "description":  "http://www.facebook.com/rickroll548  As
long as trolls are still trolling, the Rick will never stop rolling."
      },
      "indices": [
        62,
        85
      ]
    }
  ]
}
```

user_mentions	Matriz de objetos de mención del usuario	Representa a otros usuarios de Twitter mencionados en el texto del Tuit. Ejemplo:
		{
		"user_mentions": [
		{
		"name": "Twitter API",
		"indices": [
		4,
		15
		],
		"screen_name": "twitterapi",
		"id": 6253282,
		"id_str": "6253282"
		}
		]
		}

symbols	Matriz de objetos de símbolo	Representa símbolos, es decir, $ hashtags, incluidos en el texto del Tuit. Ejemplo: { "symbols": [{ "indices": [12,

```
        17
      ],
      "text": "twtr"
    }
  ]
}
```


polls	Matriz de objetos de encuesta	Representa las encuestas de Twitter incluidas en el tuit. Ejemplo:
		{"polls": [
		{
		"options": [
		{
		"position": 1,
		"text": "I read documentation once."
		},
		{
		"position": 2,
		"text": "I read documentation twice."
		},
		{
		"position": 3,
		"text": "I read documentation over and over again."
		}

```
        ],
        "end_datetime": "Thu May 25 22:20:27 +0000 2017",
        "duration_minutes": 60
      }
    ]
  }
```


6.4.5.2. Objeto hashtag

La sección entities contendrá una matriz hashtags que contiene un objeto para cada hashtag incluido en el cuerpo del Tuit e incluirá una matriz vacía si no hay hashtags presentes.

El operador # PowerTrack se utiliza para hacer coincidir el atributo text. El operador has:hashtags coincidirá si hay al menos un elemento en la matriz.

Campo	Tipo	Descripción
indices	Matriz de Enteros	Una matriz de números enteros que indican las compensaciones dentro del texto del Tuit donde comienza y termina el hashtag. El primer número entero representa la ubicación del carácter # en la cadena de texto del Tuit. El segundo número entero representa la ubicación del primer carácter después del hashtag. Por lo tanto, la diferencia entre los dos números será la longitud del nombre del hashtag más uno (para el carácter '#'). Ejemplo: "indices": [32,38]

text	Cadena	Nombre del hashtag, menos el carácter "#" inicial. Ejemplo: "text": "nodejs"

6.4.5.3. Objeto multimedia

La sección entities contendrá una matriz media que contiene un solo objeto multimedia si algún objeto multimedia ha sido 'adjunto' al Tuit. Si no se ha adjuntado ningún medio nativo, no habrá ningún array media en el entities. Por las siguientes razones, la sección extended_entities debe usarse para procesar los medios nativos de Tuit:

- Los medios type siempre indicarán ‘photo’ incluso en los casos en que se adjunte un video y un GIF al Tuit.
- Aunque se pueden adjuntar hasta cuatro fotos, solo la primera aparecerá en la sección entities.

El operador has:media coincidirá si se completa esta matriz.

Campo	Tipo	Descripción
display_url	Cadena	URL del medio para mostrar a los clientes. Ejemplo: "display_url": "pic.twitter.com/rJC5Pxsu"
expand_url	Cadena	Una versión ampliada de display_url. Enlaces a la página de visualización de medios. Ejemplo: "extended_url": "http://twitter.com/yunorno/status/114080493036773378/photo/1"
id	Entero 64	ID del medio expresado como un número entero de 64 bits. Ejemplo: "id": 114080493040967680

id_str	Cadena	ID del medio expresado como una cadena. Ejemplo: "id_str": "114080493040967680"
indices	Matriz de Enteros	Una matriz de números enteros que indica las compensaciones dentro del texto del Tuit donde comienza y termina la URL. El primer número entero representa la ubicación del primer carácter de la URL en el texto del Tuit. El segundo entero representa la ubicación del primer carácter no URL que aparece después de la URL (o el final de la cadena si la URL es la última parte del texto del Tuit). Ejemplo: "indices": [15,35]
media_url	Cadena	Una URL http:// que apunta directamente al archivo multimedia cargado. Ejemplo: "media_url": "http://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg" Para los medios en mensajes directos, media_url es la misma URL https que media_url_https y se debe acceder firmando una solicitud con el token de acceso del usuario mediante OAuth 1.0A. No es posible acceder a las imágenes a través de una sesión autenticada de twitter.com. No se puede incrustar estas imágenes directamente en una página web.

media_url_https	Cadena	Una URL https:// que apunta directamente al archivo multimedia cargado, para incrustar en páginas https. Ejemplo: "media_url_https": "https://p.twimg.com/AZVLmp-CIAAbkyy.jpg" Para medios en mensajes directos, media_url_https debe accederse firmando una solicitud con el token de acceso del usuario mediante OAuth 1.0A. No es posible acceder a las imágenes a través de una sesión autenticada de twitter.com. No se puede incrustar estas imágenes directamente en una página web.

sizes	Objeto de tamaño	Un objeto que muestra los tamaños disponibles para el archivo multimedia. Ejemplo:
		{
		"sizes": {
		"thumb": {
		"h": 150,
		"resize": "crop",
		"w": 150
		},
		"large": {
		"h": 1366,
		"resize": "fit",
		"w": 2048
		},
		"medium": {
		"h": 800,
		"resize": "fit",
		"w": 1200
		},
		"small": {
		"h": 454,
		"resize": "fit",
		"w": 680

```
    }
  }
}
```

source_status_id	Entero 64	Anulable. En el caso de los Tuits que contienen medios que se asociaron originalmente con un tuit diferente, este ID apunta al Tuit original. Ejemplo: "source_status_id": 205282515685081088

source_status_id_str	Entero 64	Anulable. Para los Tuits que contienen medios que se asociaron originalmente con un tuit diferente, esta ID basada en cadenas apunta al Tuit original. Ejemplo: "source_status_id_str": "205282515685081088"
type	Cadena	Tipo de medios cargados. Los tipos posibles incluyen foto, video y animated_gif. Ejemplo: "type":"photo"
url	Cadena	URL envuelta para el enlace de medios. Esto se corresponde con la URL incrustada directamente en el texto del Tuit sin procesar y los valores del parámetro indices. Ejemplo: "url": "http://t.co/rJC5Pxsu"

6.4.5.4. Objetos de tamaño de medio

Todos los Tuits con medios nativos (fotos, videos y GIF) incluirán un conjunto de tamaños de 'pulgar', 'pequeño', 'mediano' y 'grande' con tamaños de píxeles de alto y ancho. Para fotos y URL de medios de imagen de vista previa, el formato de URL de Photo Media especifica cómo construir diferentes URL para cargar medios fotográficos de diferentes tamaños.

Campo	Tipo	Descripción
thumb	Objeto de tamaño	Información para una versión en miniatura de los medios. Ejemplo: "thumb":{"h":150, "resize":"crop", "w":150} Los medios fotográficos del tamaño de una miniatura se limitarán a llenar un límite de 150x150 y se recortarán.

large	Objeto de tamaño	Información para una versión de gran tamaño de los medios. Ejemplo: "large":{"h":454, "resize":"fit", "w":680} Los soportes fotográficos de tamaño pequeño se limitarán a ajustarse a un límite de 680x680.
medium	Objeto de tamaño	Información para una versión mediana del medio. Ejemplo: "medium": {"h": 800, "resize": "fit", "w": 1200} Los soportes fotográficos de tamaño mediano estarán limitados para ajustarse a un límite de 1200x1200.
small	Objeto de tamaño	Información para una versión de tamaño reducido de los medios. Ejemplo: "small": {"h": 1366, "resize": "fit", "w": 2048} Los soportes fotográficos de gran tamaño se limitarán a ajustarse a un límite de 2048x2048.

Campo	Tipo	Descripción
w	Entero	Ancho en píxeles de este tamaño. Ejemplo: "w": 150
h	Entero	Altura en píxeles de este tamaño. Ejemplo: "h": 150

redimensión	Cadena	Método de cambio de tamaño utilizado para obtener este tamaño. Un valor de fit significa que los medios se redimensionaron para adaptarse a una dimensión, manteniendo su relación de aspecto original. Un valor de crop significa que el medio se recortó para ajustarse a una resolución específica. Ejemplo: "resize": "crop"

6.4.5.5. Formato de URL de medios fotográficos

Los medios fotográficos en Twitter se pueden cargar en diferentes tamaños. Es mejor cargar la imagen de menor tamaño que sea lo suficientemente grande para caber en una ventana de visualización de imagen en particular. Para cargar diferentes tamaños, el objeto de tamaño y media_url (o media_url_https) deben combinarse en un formato particular. Se usará el objeto de ejemplo de entidad de medios ya proporcionado para nuestro ejemplo al construir una URL de medios fotográficos.

Se pueden cargar media_url o media_url_https por sí solos, lo que hará que la variante media se cargue de forma predeterminada. Sin embargo, es preferible proporcionar una URL de medios fotográficos con formato completo cuando sea posible.

Hay tres partes de la URL de un medio fotográfico:

URL base	La URL base es la URL multimedia sin la extensión del archivo. Por ejemplo: "media_url_https": "https://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg", La URL base es entonces: https://pbs.twimg.com/media/DOhM30VVwAEpIHq

Formato	El formato es el tipo de foto con la que se formatea la imagen. Los formatos posibles son jpg o png, que se proporcionan como la extensión de la URL multimedia. Por ejemplo: "media_url_https": "https://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg", El formato es entonces: jpg

Nombre	El nombre es el nombre del campo del tamaño a cargar.
	
	Por ejemplo:
	
	{
	
	"sizes": {
	
	"thumb": {
	
	"h": 150,
	
	"resize": "crop",
	
	"w": 150
	
	},
	
	"large": {
	
	"h": 1366,
	
	"resize": "fit",
	
	"w": 2048
	
	},
	
	"medium": {
	
	"h": 800,
	
	"resize": "fit",
	
	"w": 1200
	
	},
	
	"small": {
	
	"h": 454,
	
	"resize": "fit",
	
	"w": 680

```
  }
 }
}
```

El nombre al cargar la foto de gran tamaño sería: grande

Se tomará estas tres partes (URL base, formato y nombre) y se combinarán en la URL del medio fotográfico para cargar. Hay 2 formatos para cargar imágenes de esta manera, *heredado* y *moderno* . Todas las cargas de imágenes deben dejar de usar el formato *heredado* y usar el formato *moderno* . El uso del formato *moderno* dará como resultado una mejor tasa de aciertos de CDN para la persona que llama, mejorando así las latencias

de carga al ser menos probable que tenga que generar y cargar los medios desde el centro de datos.

Formato heredado	El formato heredado está en desuso. Todas las cargas de medios fotográficos deben pasar al formato moderno. <url_base>. <formato>: <nombre> Por ejemplo: https://pbs.twimg.com/media/DOhM30VVwAEpIHq.jpg:large
Formato moderno	El formato moderno para cargar fotos se estableció en Twitter en 2015 y ha sido de facto desde 2017. Todas las cargas de medios fotográficos deben pasar a este formato. <base_url>? format = <format> & name = <name> Por ejemplo: https://pbs.twimg.com/media/DOhM30VVwAEpIHq?format=jpg&name=large Nota: los elementos de la cadena de consulta de la URL del medio fotográfico están en orden alfabético. Si la carga de medios añadiera elementos de consulta adicionales, el orden alfabético seguiría siendo necesario. Por ejemplo, si hubiera un nuevo elemento de consulta hipotético llamado formato_preferido , iría después de formato y nombre en la cadena de consulta.

6.4.5.6. Objeto URL

La sección entities contendrá una matriz urls que contiene un objeto para cada enlace incluido en el cuerpo del Tuit e incluirá una matriz vacía si no hay enlaces presentes.

El operador has:links coincidirá si hay al menos un elemento en la matriz. El operador url: se utiliza para hacer coincidir el atributo expanded_url. Si está utilizando el enriquecimiento de URL expandido , el operador url: se utiliza para hacer coincidir el atributo unwound.url (URL completamente desenrollado). Si está utilizando el enriquecimiento de URL mejorado , los operadores url_title: y url_decription: se utilizan para hacer coincidir los atributos unwound.title y unwound.description.

Campo	Tipo	Descripción
display_url	Cadena	URL pegada / escrita en Tuit. Ejemplo: "display_url": "bit.ly/2so49n2"
expand_url	Cadena	Versión ampliada de '' display_url ''. Ejemplo: "extended_url": "http://bit.ly/2so49n2"
indices	Matriz de Enteros	Matriz de números enteros que representan compensaciones dentro del texto del Tuit donde comienza y termina la URL. El primer número entero representa la ubicación del primer carácter de la URL en el texto del Tuit. El segundo entero representa la ubicación del primer carácter que no es de URL después del final de la URL. Ejemplo: "indices": [30,53]
url	Cadena	URL envuelta, correspondiente al valor incrustado directamente en el texto del Tuit sin procesar y los valores del parámetro de índices. Ejemplo: "url": "https://t.co/yzocNFvJuL"

Si está utilizando los enriquecimientos de URL expandidos y / o mejorados, los siguientes metadatos están disponibles bajo el atributo unwound:

Campo	Tipo	Descripción

url	Cadena	La versión completamente desenrollada del enlace que está incluido en el Tuit. Ejemplo: "url": "https://blog.twitter.com/en_us/topics/insights/2016/using-twitter-as-a-go-to-communication-channel-during-severe-weather-events.html"
status	Entero	Estado HTTP final del proceso de desenrollado, un '200' que indica éxito. Ejemplo: 200
title	Cadena	Título HTML del enlace. Ejemplo: "title":"Using Twitter as a ‘go-to’ communication channel during severe weather"
description	Cadena	Descripción HTML del enlace. Ejemplo: "description":"Using Twitter as a ‘go-to’ communication channel during severe weather"

6.4.5.7. Objeto de mención de usuario

La sección entities contendrá una matriz user_mentions que contiene un objeto por cada mención de usuario incluida en el cuerpo del Tuit, e incluirá una matriz vacía si no hay mención de usuario presente.

El operador @ PowerTrack se utiliza para hacer coincidir el atributo screen_name. El operador has:mentions coincidirá si hay al menos un elemento en la matriz.

Campo	Tipo	Descripción
id	Entero 64	ID del usuario mencionado, como un número entero. Ejemplo:

		"id": 6253282
id_str	Cadena	Si es del usuario mencionado, como una cadena. Ejemplo: "id_str": "6253282"
indices	Matriz de Enteros	Matriz de números enteros que representan las compensaciones dentro del texto del Tuit donde comienza y termina la referencia del usuario. El primer número entero representa la ubicación del carácter '@' de la mención del usuario. El segundo número entero representa la ubicación del primer carácter que no es del nombre de pantalla después de la mención del usuario. Ejemplo: "índices": [4,15]
name	Cadena	Nombre para mostrar del usuario al que se hace referencia. Ejemplo: "name":"Twitter API"
screen_name	Cadena	Nombre de pantalla del usuario referenciado. Ejemplo: "screen_name": "twitterapi"

6.4.5.8. Objeto símbolo

La sección entities contendrá una matriz symbols que contiene un objeto por cada $ hashtag incluido en el cuerpo del Tuit, e incluirá una matriz vacía si no hay ningún símbolo presente.

El operador $ PowerTrack se utiliza para hacer coincidir el atributo text. El operador has:symbols coincidirá si hay al menos un elemento en la matriz.

Campo	Tipo	Descripción
indices	Matriz de Enteros	Una matriz de números enteros que indican las compensaciones dentro del texto del Tuit donde el símbolo / etiqueta de efectivo comienza y termina. El primer número entero representa la ubicación del carácter $ en la cadena de texto del Tuit. El segundo número entero representa la ubicación del primer carácter después de la etiqueta de efectivo. Por lo tanto, la diferencia entre los dos números será la longitud del nombre del hashtag más uno (para el carácter '$'). Ejemplo: "indices": [12,17]
text	Cadena	Nombre de la etiqueta de efectivo, menos el carácter "$" inicial. Ejemplo: "text": "twtr"

6.4.5.9. Objeto de encuesta

La sección entities contendrá una matriz polls que contiene un solo objeto poll si el Tuit contiene una encuesta. Si no se incluye una encuesta, no habrá ningún arreglo polls en la sección entities.

Campo	Tipo	Descripción

options	Matriz de objeto de opción	Una serie de opciones, cada una con una posición de encuesta y el texto para esa posición. Ejemplo:
		
		{"options": [
		
		{
		
		"position": 1,
		
		"text": "I read documentation once."
		
		}

		]
		}
end_datetime	Cadena	Marca de tiempo (UTC) de cuándo finaliza la encuesta. Ejemplo: "end_datetime": "Thu May 25 22:20:27 +0000 2017"
duration_minutes	Cadena	Duración de la encuesta en minutos. Ejemplo: "duration_minutes": 60

6.4.5.10. Retuits y Citas de Tuits

Desde la perspectiva de la API de Twitter, el Retuit y la Cita son tipos especiales de Tuits que contienen el Tuit original como un objeto incrustado. Por lo tanto, los objetos Retuits y Citas son padres de un Twuit "original" hijo (y, por lo tanto, duplican el tamaño). Los retuits tienen un objeto "retweeted_status" de nivel superior y los tuits citados tienen un objeto "quoted_status". Para mantener la coherencia, estos objetos Retuit y Cita de nivel superior también tienen una propiedad de texto y entidades asociadas. Sin embargo, las entidades del nivel superior pueden diferir de las entidades proporcionadas por las entidades "originales" integradas. En el caso de los Retuits, el texto nuevo se antepone al cuerpo del Tuit original. En el caso de los tuits itados, se agrega texto nuevo al cuerpo del tuit.

En general, la mejor práctica es recuperar el texto, las entidades, el autor original y la fecha del Tuit original en retweeted_status siempre que exista. Una excepción es la obtención de entidades de Twitter que formen parte de la cotización aditiva.

6.4.5.11. Retuits

Un detalle importante con Retuits es que no se pueden agregar *entidades de* Twitter adicionales al Tuit. Los usuarios no pueden agregar hashtags, URL u otros detalles cuando retuit. Sin embargo, el atributo de texto Retuit (nivel superior) está compuesto por el texto original del Tuit con "RT @nombre de usuario:" antepuesto.

En algunos casos, especialmente con cuentas con nombres de usuario largos, la combinación de estos nuevos caracteres y el cuerpo del Tuit original puede exceder fácilmente el límite de longitud del texto del Tuit original de 140 caracteres. Para preservar la compatibilidad con la visualización y el almacenamiento de 140 caracteres, el cuerpo de nivel superior trunca el final del cuerpo del Tuit y agrega puntos suspensivos (“…”). En consecuencia, algunas entidades de nivel superior ubicadas al final del Tuit original pueden ser incorrectas o faltar, por ejemplo, en el caso de un hashtag truncado o una entrada de URL.

Este Tuit, https://twitter.com/FloodSocial/status/907974220298125312, tiene el siguiente texto del Tuit:

Just another test Tweet that needs to be exactly 140 characters with trailing URL and hashtag …… http://wapo.st/2w8iwPQ #Testing

En el ejemplo anterior, tanto la URL como el hashtag se vieron afectados. Dado que el hashtag se truncó por completo y la URL se truncó parcialmente, faltan en las entidades de nivel superior. También notará que la entidad de nivel superior user_mentions adicional proviene del prefijo “RT @floodsocial:” en el campo de texto.

Sin embargo, el texto y las entidades del Tuit en retweeted_status reflejan perfectamente el Tuit original sin truncamiento o entidades incorrectas, de ahí la recomendación de confiar en el objeto *retweeted_status* anidado para Retuits.

6.4.5.12. Citas

Los Tuits con citas se introdujeron en 2016 y se diferencian de los Retuits en que cuando "citas" un Tuit, estás agregando contenido nuevo "encima" de un Tuit compartido. Este nuevo contenido puede incluir casi cualquier cosa que pueda tener un Tuit original, incluidos texto nuevo, hashtags, menciones y URL.

Los Tuits de Citas pueden contener medios nativos (fotos, videos y GIF) y aparecerán debajo del objeto de entidades.

Dado que se pueden agregar entidades de Twitter, es probable que las entidades de Cotización sean diferentes de las entidades originales.

En este ejemplo, se colocaron una nueva URL y un hashtag al final del Tuit de cotización.

Este Tuit, https://twitter.com/FloodSocial/status/907983973225160704, tiene el siguiente texto del Tuit:

strange and equally tragic when islands flood... trans-atlantic testing of quote tweets | @thisuser @thatuser http://bit.ly/2vMMDuu #testing

En este caso, las entidades de nivel superior no reflejan los detalles de la cotización.

Sin embargo, el texto y las entidades del Tuit en extended_tweet reflejan perfectamente el Tuit de Cotización sin truncamiento o entidades incorrectas, de ahí la recomendación de confiar en el objeto *extended_tweet* anidado para los Tuits de Cotización.

6.4.5.13. Entidades para objeto de usuario

Las entidades para objetos de usuario describen las URL que aparecen en los campos de descripción y URL del perfil definido por el usuario. No describen hashtags ni menciones de usuario. A diferencia de las entidades de Tuit, las entidades de usuario pueden aplicar a múltiples campos dentro de su objeto principal; para eliminar la ambigüedad, encontrará un nodo principal llamado URL y una descripción que indica qué campo contiene la URL autorizada.

En este ejemplo, el campo de la URL del usuario contiene un enlace t.co que está completamente expandido dentro del nodo Entidades / URL / URL [0] de la respuesta. El usuario no tiene una URL envuelta en su descripción.

Un ejemplo JSON:


```
{
  "id": 6253282,
  "id_str": "6253282",
  "name": "Twitter API",
  "screen_name": "twitterapi",
  "location": "San Francisco, CA",
```

```
  "description": "The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
  "url": "http:\/\/t.co\/78pYTvWfJd",
  "entities": {
    "url": {
      "urls": [
        {
          "url": "http:\/\/t.co\/78pYTvWfJd",
          "expanded_url": "http:\/\/dev.twitter.com",
          "display_url": "dev.twitter.com",
          "indices": [
            0,
            22
          ]
        }
      ]
    },
    "description": {
      "urls": [

      ]
    }
  }
}
```

6.4.6. GET statuses/user_timeline

Devuelve una colección de los Tuits más recientes publicados por el usuario indicado por los parámetros screen_name o user_id.

Los cronogramas de usuario que pertenecen a usuarios protegidos solo se pueden solicitar cuando el usuario autenticado "posee" el cronograma o es un seguidor aprobado del propietario.

La línea de tiempo devuelta es equivalente a la que se ve como el perfil de un usuario en Twitter.

Este método solo puede devolver hasta 3200 de los Tuits más recientes de un usuario. Los retuits nativos de otros estados por parte del usuario se incluyen en este total, independientemente de si include_rts se establece en false al solicitar este recurso (Twitter Developers, s.f.).

6.4.6.1. URL del recurso

https://api.twitter.com/1.1/statuses/user_timeline.json

6.4.6.2. Información de recursos

Formatos de respuesta	JSON
¿Requiere autenticación?	si
¿Tarifa limitada?	si
Solicitudes / ventana de 15 minutos (autenticación de usuario)	900
Solicitudes / ventana de 15 minutos (autenticación de la aplicación)	1500
Solicitudes / ventana de 24 horas	100.000

El límite de solicitudes de 24 horas se basa en un reloj continuo, que comienza en el momento de la primera solicitud y se monitorea durante las próximas 24 horas.

6.4.6.3. Parámetros

Nombre	Necesario	Descripción	Valor por defecto	Ejemplo
user_id	Opcional	El ID del usuario para quien devolver los resultados.		*12345*
screen_name	Opcional	El nombre de pantalla del usuario para el que se mostrarán los resultados.		*noradio*
since_id	Opcional	Devuelve resultados con un ID mayor que (es decir, más reciente que) el ID especificado. Hay límites en la cantidad de Tuits a los que se puede acceder a través de la API. Si el límite de Tuits ha ocurrido desde since_id, since_id se forzará al ID más antiguo disponible.		*12345*

count	Opcional	Especifica la cantidad de Tuits que se deben intentar recuperar, hasta un máximo de 200 por solicitud distinta. El valor del *recuento* se considera mejor como un límite al número de Tuits que se devolverán porque el contenido suspendido o eliminado se elimina después de que se ha aplicado el recuento. Incluimos retuits en el recuento, incluso si no se proporciona *include_rts* . Se recomienda que siempre envíe *include_rts = 1* cuando utilice este método de API.	
max_id	Opcional	Devuelve resultados con un ID menor que (es decir, mayor que) o igual al ID especificado.	*54321*
trim_user	Opcional	Cuando se establece en cualquiera de *true* , *t* o *1* , cada Tuit devuelto en una línea de tiempo incluirá un objeto de usuario incluyendo solo los autores en estado numérico ID. Omita este parámetro para recibir el objeto de usuario completo.	*true*
exclude_replies	Opcional	Este parámetro evitará que aparezcan respuestas en la línea de tiempo devuelta. El uso de *exclude_replies* con el parámetro de *recuento* significará que recibirá tuits hasta contar; esto se debe a que el parámetro de *recuento* recupera esa	*true*

		cantidad de tuits antes de filtrar los retuits y las respuestas.	
include_rts	Opcional	Cuando se establece en *false*, la línea de tiempo eliminará los retuits nativos (aunque seguirán contando tanto para la duración máxima de la línea de tiempo como para el segmento seleccionado por el parámetro de recuento). Nota: Si está utilizando el parámetro trim_user junto con include_rts, los retuits aún contendrán un objeto de usuario completo.	*false*

6.4.6.4. Paginación

La API de Twitter tiene varios métodos, como GET status / user_timeline y GET status / home_timeline , que devuelven una línea de tiempo de los datos del Tuit. Tales líneas de tiempo pueden crecer mucho, por lo que existen límites en cuanto a la cantidad de una línea de tiempo que una aplicación cliente puede obtener en una sola solicitud. Por lo tanto, las aplicaciones deben recorrer los resultados de la línea de tiempo para crear una lista más completa.

Debido a la naturaleza en tiempo real de Twitter y al volumen de datos que se agregan constantemente a las líneas de tiempo, los enfoques de paginación estándar no siempre son efectivos. Es por ello por lo que pueden sucederse algunos problemas a los que los desarrolladores de Twitter se pueden enfrentar al paginar los conjuntos de resultados y brindar las mejores prácticas para procesar una línea de tiempo.

En un mundo ideal, la paginación sería muy fácil de implementar. Si se considera el caso en el que una línea de tiempo tiene 10 Tuits ordenados cronológicamente al revés. Una

aplicación puede intentar leer la línea de tiempo completa en dos solicitudes estableciendo un tamaño de página de 5 elementos y solicitando la primera página, luego la segunda página.

El problema con este método es que en las líneas de tiempo de Twitter se agregan constantemente nuevos Tuits al frente. Si se agregan dos nuevos Tuis a la línea de tiempo entre la primera y la segunda llamada, la segunda búsqueda recupera dos Tuits que se devolvieron en la llamada anterior. De hecho, si se agregan 5 o más Tuits entre llamadas, las llamadas posteriores eventualmente recuperarían todos los Tuits devueltos desde la primera solicitud, lo que hace que una solicitud API completa sea completamente redundante.

La solución al problema descrito anteriormente es utilizar una técnica para trabajar con flujos de datos llamada cursor. En lugar de leer una línea de tiempo relativa a la parte superior de la lista (que cambia con frecuencia), una aplicación debe leer la línea de tiempo relativa a los ID de los Tuits que ya ha procesado. Esto se logra mediante el uso del parámetro de solicitud max_id.

Para usar max_id correctamente, la primera solicitud de una aplicación a un método de la línea de tiempo solo debe especificar un recuento. Al procesar esta respuesta y las siguientes, lleve un registro de la identificación más baja recibida. Este ID debe pasarse como el valor del parámetro max_id para la próxima solicitud, que solo devolverá Tuits con ID menores o iguales al valor del parámetro max_id. Hay que tener en cuenta que dado que el parámetro max_id es inclusivo, el Tuit con el ID coincidente en realidad se devolverá nuevamente.

Si bien un Tuit redundante no es terriblemente ineficiente, aún es posible optimizar las solicitudes max_id para abordar este problema si su plataforma es capaz de trabajar con enteros de 64 bits. Los entornos en los que un ID de Tuit no se puede representar como un número entero con 64 bits de precisión (como JavaScript) deben omitir este paso. Reste 1 del ID de Tuit más bajo devuelto de la solicitud anterior y utilícelo para el valor de max_id. No importa si este max_id ajustado es un ID de Tuit válido o si se corresponde con un Tuit publicado por un usuario diferente; el valor solo se utiliza para decidir qué Tuits filtrar. Cuando se ajusta de esta manera, es posible recorrer una línea de tiempo sin recibir Tuits redundantes.

Las aplicaciones que procesan una línea de tiempo, esperan una cierta cantidad de tiempo y luego necesitan procesar nuevos Tuits que se han agregado desde la última vez que se

procesó la línea de tiempo, pueden hacer una optimización más usando el parámetro since_id.

Si se considera el ejemplo anterior donde se procesaron los Tuits del 1 al 10 y que los Tuits 11 a 18 se agregaron a la línea de tiempo desde que comenzó el procesamiento en el ejemplo anterior. Un enfoque ineficiente para procesar los nuevos Tuits sería iterar desde el principio de la lista hasta que apareciera el Tuit 10.

Este problema se evita configurando el parámetro since_id en el ID más grande de todos los Tuits que su aplicación ya ha procesado. A diferencia de max_id, el parámetro since_id no es inclusivo, por lo que no es necesario ajustar el ID de ninguna manera. Twitter solo devolverá Tuits con ID superiores al valor pasado para since_id.

Las aplicaciones que utilizan los parámetros max_id y since_id minimizan correctamente la cantidad de datos redundantes que obtienen y procesan, al tiempo que conservan la capacidad de iterar sobre todo el contenido disponible de una línea de tiempo (Twitter Developers, s.f.).

6.4.7. API de búsqueda estándar

La API de búsqueda estándar de Twitter (search/tweets) permite consultas simples contra los índices de tuits recientes o populares y se comporta de manera similar, pero no exactamente, a la función de interfaz de usuario de búsqueda disponible en los clientes móviles o web de Twitter. La API de búsqueda de Twitter busca en una muestra de tuits recientes publicados en los últimos 7 días. Es por ello por lo que es importante saber que la API de búsqueda estándar se centra en la relevancia y no en la integridad. Esto significa que algunos Tuits y usuarios pueden faltar en los resultados de búsqueda (Twitter Developers, s.f.).

La API de búsqueda estándar, por tanto, devuelve una colección de Tuits relevantes que coinciden con una consulta específica (Twitter Developers, s.f.).

6.4.7.1. URL del recurso

https://api.twitter.com/1.1/search/tweets.json

6.4.7.2. Información de recursos

Formatos de respuesta	JSON

¿Requiere autenticación?	si
¿Tarifa limitada?	si
Solicitudes / ventana de 15 minutos (autenticación de usuario)	180
Solicitudes / ventana de 15 minutos (autenticación de la aplicación)	450

6.4.7.3. Parámetros

Nombre	Necesario	Descripción	Valor por defecto	Ejemplo
q	necesario	Una consulta de búsqueda UTF-8 codificada en URL de 500 caracteres como máximo, incluidos los operadores. Además, las consultas pueden estar limitadas por la complejidad.		@noradio

geocode	Opcional	Devuelve tuits de usuarios ubicados dentro de un radio determinado de la latitud / longitud determinada. La ubicación se toma preferentemente de la API de geoetiquetado, pero recurrirá a su perfil de Twitter. El valor del parámetro se especifica mediante " latitude,longitude,radius", donde las unidades de radio deben especificarse como " mi" (millas) o " km" (kilómetros). Tenga en cuenta que no puede utilizar el operador cercano a través de la API para geocodificar ubicaciones arbitrarias; sin embargo, puede utilizar este parámetro geocode para buscar códigos geográficos cercanos directamente. Se considerará un máximo de 1,000 "subregiones" distintas cuando se use el modificador de radio.	37.781157 - 122.398720 1mi

lang	Opcional	Restringe los tuits al idioma dado, dado por un código ISO 639-1 . La detección del idioma es el mejor esfuerzo.	eu
locale	Opcional	Especifique el idioma de la consulta que está enviando (solo está vigente actualmente). Está destinado a consumidores de idiomas específicos y el valor predeterminado debería funcionar en la mayoría de los casos.	ja
result_type	Opcional	Opcional. Especifica qué tipo de resultados de búsqueda prefiere recibir. El valor predeterminado actual es "mixto". Los valores válidos incluyen: * mixed: Incluya resultados populares y en tiempo real en la respuesta. * recent: devuelve solo los resultados más recientes en la respuesta * popular: devuelve solo los resultados más populares en la respuesta.	mixed recent popular
count	Opcional	El número de tuits que se devolverán por página, hasta un máximo de 100. El valor predeterminado es 15. Este era	100

		antes el parámetro "rpp" en la antigua API de búsqueda.	
until	Opcional	Devuelve tuits creados antes de la fecha indicada. La fecha debe tener el formato AAAA-MM-DD. Tenga en cuenta que el índice de búsqueda tiene un límite de 7 días. En otras palabras, no se encontrarán tuits para una fecha anterior a una semana.	2015-07-19
since_id	Opcional	Devuelve resultados con un ID mayor que (es decir, más reciente que) el ID especificado. Hay límites en la cantidad de Tuits a los que se puede acceder a través de la API. Si el límite de Tuits ha ocurrido desde since_id, since_id se forzará al ID más antiguo disponible.	12345
max_id	Opcional	Devuelve resultados con un ID menor que (es decir, mayor que) o igual al ID especificado.	54321
include_entities	Opcional	El nodo entities no se incluirá cuando se establezca en false.	false

6.4.7.4. Operadores

Con la función de búsqueda reciente de Labs, se envía una sola consulta (también conocida como 'regla' o 'filtro') con una solicitud GET y se devuelven los Tuits coincidentes. Las consultas se componen de operadores que se utilizan para hacer coincidir una variedad de atributos de Tuit.

Para crear una consulta, se puede especificar un solo operador "independiente" o combinar varios operadores.

Los operadores especificados juntos se combinarán en una sola cláusula AND. Por ejemplo, “snow day #NoSchool” coincidirá con los Tuits que contengan las palabras clave snow y day y el hashtag #NoSchool.

También se puede hacer coincidir Tuits en función de la presencia de cualquiera de los operadores utilizando la conjunción OR. Por ejemplo, especificar “gato OR gruñón OR #meme” coincidirá con cualquier Tuit que contenga al menos uno de gato, gruñón o el hashtag #meme .

Se puede utilizar paréntesis para agrupar operadores. Por ejemplo, (gato gruñón) OR (#meme has:images) devolverá Tuits que contengan los términos gato y gruñón, o Tuits con imágenes que contengan el hashtag #meme .

Se puede anteponer un guion a una palabra clave (o cualquier operador) para negarla (NOT). Por ejemplo, “gato #meme -gruñón” coincidirá con los Tuits que contengan el hashtag #meme y el término gato, pero solo si no contienen el término gruñón. También se puede negar los operadores agrupados mediante paréntesis. Una cláusula de consulta común es -is:retweet , que no coincidirá en Retuits, por lo que solo coincidirá en Tuits originales.

Algunos detalles importantes acerca de la construcción de consultas son:

- Las consultas pueden tener 512 caracteres.
- Todos los operadores se pueden negar. Los operadores negados no se pueden utilizar solos.
- No negar un conjunto de operadores agrupados en un conjunto de paréntesis. En su lugar, negar a cada operador individual. Por ejemplo, en lugar de usar -(gato OR gruñón OR meme), sería más conveniente usar -gato -gruñón -meme (Twitter Developers, s.f.).

6.4.7.5. Operadores independientes

Estos operadores se pueden usar solos o junto con cualquier otro operador (incluidos los de la lista de operadores de Entidad).

Operador	Descripción
keyword	Coincide con una palabra clave dentro del cuerpo de un Tuit. Esta es una coincidencia simbólica, lo que significa que su cadena de palabras clave se comparará con el texto simbólico del cuerpo del Tuit. La tokenización divide las palabras según la puntuación, los símbolos y los caracteres separadores de plano básico Unicode. Por ejemplo, un Tuit con el texto "Me gusta la coca-cola" se dividiría en las siguientes fichas: yo, me gusta, coca, cola. Estos tokens luego se compararían con la cadena de palabras clave utilizada en su regla. Para hacer coincidir cadenas que contengan signos de puntuación (por ejemplo, coca-cola), símbolos o caracteres separadores, debe envolver la palabra clave entre comillas dobles.
emoji	Coincide con un emoji dentro del cuerpo de un Tuit. Al igual que una palabra clave, los emojis son una coincidencia simbólica, lo que significa que su emoji se comparará con el texto simbólico del cuerpo del Tuit. Tenga en cuenta que si un emoji tiene una variante, debe envolverlo entre comillas dobles para agregar a una regla.
"concordancia de frase exacta"	Coincide con la frase exacta dentro del cuerpo de un Tuit.
#	Coincide con cualquier Tuit que contenga un hashtag reconocido, si el hashtag es una entidad reconocida en un Tuit. Este operador realiza una coincidencia exacta, NO una coincidencia tokenizada, lo que significa que la regla #thanku hará coincidir las publicaciones con el hashtag exacto #thanku, pero no con las que tengan el hashtag #thankunext.

@	Coincide con cualquier Tuit que mencione el nombre de usuario dado, si el nombre de usuario es una entidad reconocida (incluido el carácter @).
from:	Coincide con cualquier Tuit de un usuario específico. El valor puede ser el nombre de usuario (excluyendo el carácter @) o el ID de usuario numérico del usuario.
to:	Coincide con cualquier Tuit que responda a un usuario en particular. El valor puede ser el nombre de usuario (excluyendo el carácter @) o el ID de usuario numérico del usuario.

url:	Realiza una coincidencia tokenizada en cualquier URL con formato válido de un Tuit. Este operador puede coincidir con el contenido tanto de la URL como de la URL expandida. Por ejemplo, un Tuit que contenga "Debería consultar Twitter Developer Labs: https://t.co/c0A36SWil4" (con la URL corta que redirecciona a https://developer.twitter.com) coincidirá con las dos reglas siguientes: • from:TwitterDev url:"https://developer.twitter.com"(porque coincidirá con el contenido de entities.urls.expanded_url) • from:TwitterDev url:"https://t.co"(porque coincidirá con el contenido de entities.urls.url) Los símbolos y frases que contienen signos de puntuación o caracteres especiales deben estar entre comillas dobles (por ejemplo: url:"/developer"). De manera similar, para hacer coincidir un protocolo específico, escríbalo entre comillas dobles (por ejemplo: url:"https://developer.twitter.com").
retweets_of:	Coincide con los Tuits que son Retuits del usuario especificado. El valor puede ser el nombre de usuario (excluyendo el carácter @) o el ID de usuario numérico del usuario.
context:	Coincide con los tuits con un ID de dominio específico y / o un ID de dominio, par de ID de entidad donde * representa un comodín. Para obtener más información sobre este operador, visite nuestra página sobre Anotaciones . context: domain_id.entity_id Ejemplo:

	• context: 10.799022225751871488 (domain_id.entity_id devuelve Tuits que coinciden con ese par específico de dominio-entidad)
entity:	Coincide con los tuits con un valor de cadena de entidad específico. Para obtener más información sobre este operador, visite nuestra página sobre Anotaciones . entity: "declaración de cadena de entidad / lugar" Ejemplos: • entity: "Michael Jordan" • entity: "Barcelona"

6.4.7.6. Operadores de entidad

Los siguientes operadores no se pueden utilizar como cláusulas independientes en una consulta; solo se pueden usar junto con al menos un operador de la lista de operadores independientes.

Por ejemplo, la siguiente consulta no es compatible ya que contiene solo operadores de entidad, sería demasiado general y probablemente coincidiría con un volumen extremo de tuits.

has:mentions (has:media OR has:links)

Si agregamos un operador independiente, como la palabra clave 'nieve', se admite la siguiente consulta.

nieve has:mentions (has:media OR has:links)

Operador	**Descripción**

is:retweet	Coincidencias en Retuits que coinciden con el resto de la regla especificada. Este operador busca solo Retuits verdaderos (por ejemplo, los generados mediante el botón Retuit). Este operador no hará coincidir los tuits con comentarios (también conocidos como tweets con citas) y tuits modificados. Este operador se puede negar para que coincida solo con los Tuits originales, por ejemplo, -is:retweet.
is:verified	Envíe solo Tuits cuyos autores estén verificados por Twitter.
has:hashtag	Coincide con Tuits que contienen al menos un hashtag.
has:links	Este operador coincide con los Tuits que contienen enlaces en el cuerpo del Tuit.
has:mentions	Coincide con los Tuits que mencionan a otro usuario de Twitter.
has:media	Coincide con los Tuits que contienen una URL multimedia reconocida por Twitter.
has:images	Hace coincidir los tuits que contienen una URL reconocida con una imagen.
has:videos	Coincide con los tuits que contienen videos nativos de Twitter, subidos directamente a Twitter. Esto no coincidirá con los videos creados con Periscope o los Tuits con enlaces a otros sitios de alojamiento de videos.

lang:	Coincide con los tuits que han sido clasificados por Twitter como de un idioma en particular (si, y solo si, el tuit ha sido clasificado). Es importante tener en cuenta que cada Tuit actualmente solo se clasifica como de un idioma, por lo que si se combinan varios idiomas con Y, no se obtendrán resultados. Nota: si no se puede realizar una clasificación de idioma, el resultado proporcionado es 'und' (por indefinido). La siguiente lista representa los idiomas admitidos actualmente y su identificador de idioma BCP 47 correspondiente:

am Amárico	kn Canarés	sv sueco	el griego
hu húngaro	si Sinhala	et Estonio	or Oriya
pt portugués	zh chino	ml Malayalam	tr turco
ar Arábica	km Jemer	tl Tagalo	gu Gujarati
is islandés	sk eslovaco	fi finlandés	pa Panjabi
ro rumano	cs checo	dv Maldivas	uk ucranio
hy armenio	ko coreano	ta Tamil	ht haitiano
in indonesio	sl esloveno	fr francés	ps Pashto
ru ruso	da danés	mr Marathi	ur Urdu
	lo Lao	te Telugu	iw hebreo

bn bengalí	ckb Kurdo sorani	ka georgiano	fa persa
it italiano	nl holandés	ne Nepalí	ug Uigur
sr serbio	lv letón	th tailandés	hi hindi
bg búlgaro	es codicioso	de alemán	pl polaco
ja japonés	en Inglés	no noruego	vi vietnamita
sd Sindhi	lt lituano	bo Tibetano	cy galés
my birmano			

6.4.7.7. Paginación

Las consultas de búsqueda suelen coincidir en más Tuits de los que se pueden devolver en una única respuesta de API. Cuando eso sucede, los datos se devuelven en una serie de 'páginas'. La paginación se refiere a métodos para solicitar todas las páginas a fin de recuperar el conjunto de datos completo.

Los detalles fundamentales de la paginación de búsqueda reciente son los siguientes:

- El método de búsqueda reciente de Labs responderá a una consulta con al menos una página y proporcionará un next_token en su respuesta JSON si hay páginas adicionales disponibles. Para recibir Tuits coincidentes, este proceso se puede repetir hasta que no se incluya ningún token en la respuesta.
- Los tuits se envían en orden cronológico inverso, en la zona horaria UTC. Esto es cierto dentro de las páginas individuales, así como en varias páginas:
    - El primer Tuit de la primera respuesta será el más reciente que coincida con su consulta.
    - El último Tuit de la última respuesta será el más antiguo que coincida con su consulta.

- El parámetro de solicitud max_results permite configurar la cantidad de Tuits devueltos por respuesta. Este valor predeterminado es 10 tuits y tiene un máximo de 100.

- Cada implementación de paginación implica analizar next_tokens de la carga útil de respuesta e incluirlos en la solicitud de búsqueda de la 'página siguiente'.

La función de búsqueda reciente de Labs se diseñó para admitir dos patrones de uso fundamentales:

- Obtener un histórico: solicitar tuits coincidentes de un período de interés. Por lo general, se trata de solicitudes únicas en apoyo de la investigación histórica. Las solicitudes de búsqueda se pueden basar en los parámetros de solicitud start_time y end_time. La función de búsqueda reciente de Labs responde con Tuits entregados en orden cronológico inverso, comenzando con el Tuit coincidente más reciente.

- Encuestas: solicitando Tuits coincidentes que se han publicado desde el último Tuit recibido. Estos casos de uso suelen tener un enfoque casi en tiempo real y se caracterizan por solicitudes frecuentes, "escuchando" nuevos Tuits de interés. La función de búsqueda reciente de Labs proporciona el parámetro de solicitud since_id en apoyo del patrón de "sondeo". Para ayudar con la navegación por ID de Tuit, el parámetro de solicitud until_id también está disponible.

El modo histórico es el modo predeterminado de la búsqueda reciente de Labs e ilustra los fundamentos de la paginación. Para recuperar Tuits de un período de interés dentro de los últimos siete días se utilizan los parámetros de solicitud start_time y end_time. Las solicitudes de historial suelen ser solicitudes únicas en apoyo de la investigación y el análisis.

Realizar solicitudes para un período de datos es el modo predeterminado del método de búsqueda reciente de Labs. Si una solicitud de búsqueda no especifica un parámetro de solicitud start_time, end_time o since_id, end_time se establecerá de forma predeterminada en "ahora" (en realidad, 30 segundos antes de la hora de la consulta) y start_time se establecerá de forma predeterminada en siete días.

El método responderá con la primera 'página' de Tuits en orden cronológico inverso, comenzando con el Tuit más reciente. La carga útil JSON de respuesta también incluirá un next_token si hay páginas adicionales de datos. Para recopilar todo el conjunto de Tuits

coincidentes, independientemente del número de páginas, se realizan solicitudes hasta que no se proporciona next_token.

Por ejemplo, aquí hay una solicitud inicial de Tuits con la palabra clave “snow” de la última semana:

/tweets/search?query=snow

La respuesta incluye los 10 tuits más recientes, junto con estos atributos "meta" en la respuesta JSON:


```
 "meta": {
      "newest_id": "1204860593741553664",
      "oldest_id": "1204860580630278147",
      "next_token": "b26v89c19zqg8o3fobd8v73egzbdt3qao235oql",
      "result_count": 10
  }
```


Para recuperar los siguientes 10 tuits, next_token se agrega a la solicitud original. La solicitud sería:

/tweets/search?query=snow&next_token=b26v89c19zqg8o3fobd8v73egzbdt3qao235oql

El proceso de buscar un next_token e incluirlo en una solicitud posterior se puede repetir hasta que se recopilen todos (o una cantidad de) Tuits, o hasta que se haya realizado una cantidad específica de solicitudes. Si la fidelidad de los datos (recopilar todas las coincidencias de su consulta) es clave para el caso de uso, será suficiente con un "repeat until request.next_token is null".

Esta "paginación hacia atrás a través de la historia" es la forma más simple de paginación (Twitter Developers, s.f.).

6.4.8. GET followers/ids

Esta función devuelve una colección con cursor de los IDs de usuario de todos los usuarios que siguen al usuario especificado.

Los resultados se ordenan con el seguimiento más reciente primero; sin embargo, este orden está sujeto a cambios no anunciados y posibles problemas de coherencia. Los resultados se dan en grupos de 5000 ID de usuario y se puede navegar por varias "páginas" de resultados utilizando el valor next_cursor en solicitudes posteriores.

Este método es especialmente poderoso cuando se usa junto con GET users / lookup , un método que permite convertir los IDs de usuario en objetos de usuario completos de forma masiva (Twitter Developers, s.f.).

6.4.8.1. URL del recurso

https://api.twitter.com/1.1/followers/ids.json

6.4.8.2. Información de recursos

Formatos de respuesta	JSON
¿Requiere autenticación?	si
¿Tarifa limitada?	si
Solicitudes / ventana de 15 minutos (autenticación de usuario)	15
Solicitudes / ventana de 15 minutos (autenticación de la aplicación)	15

6.4.8.3. Parámetros

Nombre	Necesario	Descripción	Valor por defecto	Ejemplo

user_id	Opcional	El ID del usuario para quien devolver los resultados.		12345
screen_name	Opcional	El nombre de pantalla del usuario para el que se mostrarán los resultados.		noradio
cursor	semi-opcional	Hace que la lista de conexiones se divida en páginas de no más de 5000 ID a la vez. No se garantiza que el número de ID devueltos sea 5000, ya que los usuarios suspendidos se filtran después de consultar las conexiones. Si no se proporciona ningún cursor, se asumirá un valor de -1, que es la primera "página". La respuesta de la API incluirá un previous_cursor y next_cursor para permitir la paginación hacia adelante y hacia atrás.	-1	1289376 4510938

stringify_ids	Opcional	Algunos entornos de programación no consumirán ID de Twitter debido a su tamaño. Proporcione esta opción para que los ID se devuelvan como cadenas.	false	true
count	Opcional	Especifica el número de ID que intenta recuperar, hasta un máximo de 5000 por solicitud distinta. El valor de count es mejor como un límite para el número de resultados a devolver. Al utilizar el parámetro de recuento con este método, es aconsejable utilizar un valor de recuento coherente en todas las solicitudes a la colección del mismo usuario. Se recomienda el uso de este parámetro en entornos donde los 5000 ID constituyen una respuesta demasiado grande.		2048

6.4.9. Códigos de respuesta

La API de Twitter Estándar devuelve los códigos de estado HTTP además de códigos y mensajes de error basados en JSON, intentando devolver los más apropiados para cada solicitud (Twitter Developers, s.f.).

6.4.9.1. Códigos de estado HTTP

La API de Twitter intenta devolver los códigos de estado HTTP apropiados para cada solicitud.

Código	Texto	Descripción
200	OK	¡Éxito!
304	No modificado	No hubo nuevos datos para devolver.
400	Solicitud incorrecta	La solicitud no es válida o no se puede atender de otro modo. Un mensaje de error adjunto le explicará más. Las solicitudes sin autenticación se consideran inválidas y generarán esta respuesta.
401	No autorizado	Credenciales de autenticación faltantes o incorrectas. Esto también puede volver en otras circunstancias no definidas.
403	Prohibido	Se entiende la solicitud, pero se ha rechazado o no se permite el acceso. Un mensaje de error adjunto explicará el motivo. Este código se utiliza cuando se rechazan solicitudes debido a límites de actualización . Otras razones por las que se devuelve este estado se enumeran junto con los códigos de error en la tabla siguiente.

404	Extraviado	El URI solicitado no es válido o el recurso solicitado, como un usuario, no existe.
406	Inaceptable	Se devuelve cuando se especifica un formato no válido en la solicitud.
410	Desactivado	Este recurso se ha desactivado. Se utiliza para indicar que un método de la API se ha desactivado y ya no está disponible.
420	Mejora tu calma	Se devuelve cuando una aplicación tiene una tarifa limitada por realizar demasiadas solicitudes.
422	Entidad no procesable	Se devuelve cuando los datos no se pueden procesar (por ejemplo, si una imagen cargada en POST account / update_profile_banner no es válida, o el cuerpo JSON de una solicitud tiene un formato incorrecto).
429	Demasiadas solicitudes	Se devuelve cuando no se puede atender una solicitud debido a que el límite de frecuencia de la aplicación se ha agotado para el recurso.
500	Error de servidor interno	Algo está roto. Suele ser un error temporal, por ejemplo, en una situación de carga elevada o si un método tiene problemas temporalmente.
502	Puerta de enlace incorrecta	Twitter está caído o se está actualizando.

503	Servicio no disponible	Los servidores de Twitter están activos, pero sobrecargados de solicitudes. Inténtelo de nuevo más tarde.
504	Tiempo de espera de puerta de enlace	Los servidores de Twitter están activos, pero la solicitud no se pudo atender debido a alguna falla dentro de la pila interna. Inténtelo de nuevo más tarde.

6.4.9.2. Mensajes de error

Los mensajes de error de la API de Twitter se devuelven en formato JSON. Por ejemplo, un error podría verse así:

{ "errors" : [{ "message" : "Lo sentimos, esa página no existe" , "code" : 34 }]}

6.4.9.3. Códigos de error

Además del texto de error descriptivo, los mensajes de error contienen códigos que se pueden analizar por máquina. Si bien el texto de un mensaje de error puede cambiar, los códigos seguirán siendo los mismos.

La siguiente tabla describe los códigos que pueden aparecer al trabajar con la API estándar (tenga en cuenta que la API de anuncios y algunas otras familias de recursos pueden presentar códigos de error adicionales). Si no se incluye una respuesta de error en la tabla, vuelva a examinar los códigos de estado HTTP anteriores para determinar la mejor manera de abordar el problema.

Código	**Texto**	**Descripción**
3	Coordenadas no válidas.	Corresponde con HTTP 400. Las coordenadas proporcionadas como parámetros no eran válidas para la solicitud.

13	No hay ubicación asociada con la dirección IP especificada.	Corresponde con HTTP 404. No fue posible derivar una ubicación para la dirección IP proporcionada como parámetro en la solicitud de búsqueda geográfica.
17	Ningún usuario coincide con términos específicos.	Corresponde con HTTP 404. No fue posible encontrar un perfil de usuario que coincida con los parámetros especificados.
32	No se pudo autenticar	Corresponde con HTTP 401. Hubo un problema con los datos de autenticación para la solicitud.
34	lo siento, esa pagina no existe	Corresponde con HTTP 404. No se encontró el recurso especificado.
36	No puede reportarse a sí mismo por spam.	Corresponde con HTTP 403. No puede utilizar su propio ID de usuario en una llamada de notificación de spam.
38	Falta el parámetro <name>.	Corresponde con HTTP 403. A la solicitud le falta el parámetro <name> (como medios, texto, etc.) en la solicitud.
44	El parámetro attach_url no es válido	Corresponde con HTTP 400. El valor de URL proporcionado no es una URL que pueda adjuntarse a este Tuit.
50	Usuario no encontrado.	Corresponde con HTTP 404. No se encuentra el usuario.
63	El usuario ha sido suspendido.	Corresponde con HTTP 403 La cuenta de usuario se ha suspendido y no se puede recuperar la información.

64	Su cuenta está suspendida y no se le permite acceder a esta función.	Corresponde con HTTP 403. El token de acceso que se utiliza pertenece a un usuario suspendido.
68	La API REST de Twitter v1 ya no está activa. Migra a la API v1.1.	Corresponde con HTTP 410. La solicitud se realizó a una URL retirada de la era v1.
87	El cliente no puede realizar esta acción.	Corresponde con HTTP 403. El método llamado no es una URL permitida.
88	Excede el límite de velocidad	Corresponde con HTTP 429. Se alcanzó el límite de solicitud para este recurso para la ventana de límite de velocidad actual.
89	Token no válido o caducado	Corresponde con HTTP 403. El token de acceso utilizado en la solicitud es incorrecto o ha caducado.
92	Se requiere SSL	Corresponde con HTTP 403. Solo se permiten conexiones TLS v1.2 en la API. Actualice la solicitud a una conexión segura.
93	Esta aplicación no puede acceder ni eliminar sus mensajes directos.	Corresponde con HTTP 403. El token OAuth no proporciona acceso a Mensajes Directos.

99	No se pueden verificar sus credenciales.	Corresponde con HTTP 403. Las credenciales de OAuth no se pueden validar. Compruebe que el token sigue siendo válido.
120	Error al actualizar la cuenta: el *valor* es demasiado largo (el máximo son *nn* caracteres)	Corresponde con HTTP 403. Se lanza cuando uno de los *valores* pasados al extremo update_profile.json excede el valor máximo permitido actualmente para ese campo. El mensaje de error especificará el número máximo permitido de *nn* caracteres.
130	Sobre capacidad	Corresponde con HTTP 503. Twitter está temporalmente sobre capacidad.
131	Error interno	Corresponde con HTTP 500. Se produjo un error interno desconocido.
135	No se pudo autenticar	Corresponde con HTTP 401. Marca de tiempo fuera de los límites (a menudo causado por una desviación del reloj al autenticarse; verifique el reloj del sistema)
139	Ya ha marcado como favorito este estado.	Corresponde con HTTP 403. No se puede marcar como favorito (“me gusta”) un Tuit más de una vez.
144	No se encontró ningún estado con esa identificación.	Corresponde con HTTP 404. No se encuentra el ID de tuit solicitado (si existía, probablemente se eliminó)
150	No puedes enviar mensajes a	Corresponde con HTTP 403. Error al enviar un mensaje directo.

	usuarios que no te siguen.	
151	Hubo un error al enviar su mensaje: *motivo*	Corresponde con HTTP 403. Error al enviar un mensaje directo. El valor del *motivo* proporcionará más información.
160	Ya solicitó seguir al *usuario* .	Corresponde con HTTP 403. Esta fue una solicitud de seguimiento duplicada y una solicitud anterior aún no fue reconocida.
161	No puedes seguir a más personas en este momento.	Corresponde con HTTP 403. Se lanza cuando un usuario no puede seguir a otro usuario debido a que alcanzó el límite. Este límite se aplica a cada usuario de forma individual, independientemente de las aplicaciones que utilice para acceder a la plataforma Twitter.
179	Lo sentimos, no está autorizado para ver este estado.	Corresponde con HTTP 403. Se lanza cuando el usuario que se autentica no puede ver un Tuit, generalmente debido a que el autor del Tuit ha protegido sus Tuits.
185	El usuario supera el límite de actualización de estado diario	Corresponde con HTTP 403. Se lanza cuando un Tuit no se puede publicar debido a que el usuario no tiene permiso para publicar. A pesar del texto en el mensaje de error que indica que este error solo se genera cuando se alcanza un límite diario, este error se generará siempre que se alcance un límite de publicación. Los derechos de emisión tienen períodos de tiempo de itinerancia de duración no especificada.

186	El tuit debe ser un poco más corto.	Corresponde con HTTP 403. El texto de estado es demasiado largo.
187	El estado es un duplicado	Corresponde con HTTP 403. El texto de estado ya ha sido tuiteado por la cuenta autenticada.
195	Parámetro de URL no válido o faltante	Corresponde con HTTP 403. La solicitud debe tener un parámetro de URL válido.
205	Has superado el límite de informes de spam.	Corresponde con HTTP 403. Se alcanzó el límite de cuenta para informar spam. Inténtelo de nuevo más tarde.
214	El propietario debe permitir dms de cualquier persona.	Corresponde con HTTP 403. El usuario no está configurado para tener Mensajes Directos abiertos cuando intenta configurar un mensaje de bienvenida.
215	Datos de autenticación incorrectos	Corresponde con HTTP 400. El método requiere autenticación, pero no se presentó o fue totalmente inválido.
220	Sus credenciales no permiten el acceso a este recurso.	Corresponde con HTTP 403. El token de autenticación en uso está restringido y no puede acceder al recurso solicitado.

226	Parece que esta solicitud podría estar automatizada. Para proteger a nuestros usuarios del spam y otras actividades maliciosas, no podemos completar esta acción en este momento.	Corresponde con HTTP 403. Controlamos y ajustamos constantemente nuestros filtros para bloquear el spam y la actividad maliciosa en la plataforma de Twitter. Estos sistemas se sintonizan en tiempo real. Si recibe esta respuesta, nuestros sistemas han marcado el Tuit o el Mensaje directo como posiblemente adecuado para este perfil.
251	Este método se ha retirado y no debe utilizarse.	Corresponde con HTTP 410. La aplicación realizó una solicitud a una URL retirada.
261	La aplicación no puede realizar acciones de escritura.	Corresponde con HTTP 403. Se debe a que la aplicación está restringida para las acciones POST, PUT o DELETE. Se recomienda verificar la información en el panel de la aplicación. También se puede presentar un ticket en https://help.twitter.com/forms/platform.
271	No puedes silenciarte.	Corresponde con HTTP 403. La cuenta de usuario autenticado no se puede silenciar.
272	No estás silenciando al usuario especificado.	Corresponde con HTTP 403. La cuenta de usuario autenticado no silencia la cuenta que una llamada intenta dejar de silenciar.

323	No se permiten GIF animados al cargar varias imágenes.	Corresponde con HTTP 400. Solo se puede adjuntar un GIF animado a un solo Tuit.
324	Falló la validación de los identificadores de medios.	Corresponde con HTTP 400. Hubo un problema con el ID de medio enviado con el Tuit.
325	No se encontró una identificación de medios.	Corresponde con HTTP 400. No se encontró el ID de medio adjunto al Tuit.
326	Para proteger a nuestros usuarios del spam y otras actividades maliciosas, esta cuenta está bloqueada temporalmente.	Corresponde con HTTP 403. El usuario debe iniciar sesión en https://twitter.com para desbloquear su cuenta antes de que se pueda usar el token de usuario.
327	Ya has retuiteado este Tuit.	Corresponde con HTTP 403. El usuario no puede retuitear el mismo Tuit más de una vez.
349	No puede enviar mensajes a este usuario.	Corresponde con HTTP 403. El remitente no tiene privilegios para enviar mensajes directos al destinatario.

354	El texto de su mensaje directo supera el límite máximo de caracteres.	Corresponde con HTTP 403. El tamaño del mensaje excede la cantidad de caracteres permitidos en un mensaje directo.
355	La suscripción ya existe.	Corresponde con HTTP 409 Conflict. Relacionado con la solicitud de la API de actividad de la cuenta para agregar una nueva suscripción para un usuario autenticado.
385	Intentó responder a un Tuit que se eliminó o que no está visible para usted.	Corresponde con HTTP 403. Solo se puede enviar una respuesta con referencia a un Tuit público existente.
386	El Tuit supera el número de tipos de archivos adjuntos permitidos.	Corresponde con HTTP 403. Un Tuit está limitado a un único recurso adjunto (medios, Cotización de Tuit, etc.)
407	La URL proporcionada no es válida.	Corresponde con HTTP 400. No se pudo manejar una URL incluida en el Tuit. Esto puede deberse a que no se pudo convertir una URL que no es ASCII, o por otras razones.

415	URL de devolución de llamada no aprobada para esta aplicación cliente. Las URL de devolución de llamada aprobadas se pueden ajustar en la configuración de su aplicación	Corresponde con HTTP 403. Las URL de devolución de llamada de la aplicación deben incluirse en la lista blanca a través de la página de detalles de la aplicación en el portal para desarrolladores . La aplicación de Twitter solo puede utilizar URL de devolución de llamada aprobadas.
416	Solicitud no válida / suspendida	Corresponde con HTTP 401. La aplicación se ha suspendido y no se puede utilizar para Iniciar sesión con Twitter .
417	Las aplicaciones de escritorio solo admiten el valor oauth_callback 'oob'	Corresponde con HTTP 401. La aplicación está intentando utilizar OAuth basado en PIN fuera de banda , pero se ha especificado una URL de devolución de llamada en la configuración de la aplicación.
421	Este Tuit ya no está disponible	Corresponde con HTTP 404. No se puede recuperar el Tuit. Esto puede deberse a varias razones.
422	Este Tuit ya no está disponible porque violó las Reglas de Twitter.	Corresponde con HTTP 404. El Tuit no está disponible en la API.

433	El autor original del Tuit restringió quién puede responder a este Tuit.	Corresponde con HTTP 403. Se lanza al responder a un Tuit, y el autor de ese Tuit original limita quién puede responder. En este caso, solo se puede enviar una respuesta si el autor sigue o ha sido mencionado por el autor del Tuit original.

6.4.10. Límites y cuotas en las solicitudes a la API

Los siguientes límites solo se aplican a los métodos de la API Estándar, y no a la API Premium (Twitter Developers, s.f.).

6.4.10.1. Por usuario o por aplicación de desarrollador

La limitación de la tasa de la API estándar se realiza principalmente por usuario, o se describe con más precisión, por token de acceso de usuario. Si un método permite 15 solicitudes por ventana de límite de velocidad, entonces permite 15 solicitudes por ventana por token de acceso.

Cuando se usa el OAuth 2.0 Bearer Token, los límites de tasa se determinan globalmente para toda la aplicación para desarrolladores. Si un método permite 180 solicitudes por ventana de límite de velocidad, entonces le permite realizar 15 solicitudes por ventana, en nombre de su aplicación. Este límite se considera completamente separado de los límites por usuario.

6.4.10.2. Ventanas de 15 minutos

Los límites de frecuencia se dividen en intervalos de 15 minutos. Todos los terminales requieren autenticación, por lo que no existe el concepto de llamadas no autenticadas y límites de velocidad.

Hay dos depósitos iniciales disponibles para solicitudes GET: 15 llamadas cada 15 minutos y 180 llamadas cada 15 minutos.

6.4.10.3. Encabezados HTTP y códigos de respuesta

Se utilizan los encabezados HTTP para comprender dónde se encuentra la aplicación para un límite de velocidad determinado, en el método que se acaba de utilizar.

Hay que tener en cuenta que los encabezados HTTP son contextuales. Cuando se usa el OAuth 2.0 Bearer Token, indican el límite de velocidad para el contexto de la aplicación. Cuando se usa el OAuth 1.0a User Context, indican el límite de frecuencia para ese contexto de aplicación de usuario.

- x-rate-limit-limit: el límite máximo de tasa para ese método dado
- x-rate-limit-remaining: la cantidad de solicitudes que quedan para la ventana de 15 minutos
- x-rate-limit-reset:la ventana restante antes de que se restablezca el límite de velocidad, en segundos de época UTC

Cuando una aplicación excede el límite de frecuencia para un método de API estándar determinado, la API devolverá un código de respuesta HTTP 429 “Too Many Requests” y se mostrará el siguiente error en el cuerpo de la respuesta:

{ "errors": [{ "code": 88, "message": "Rate limit exceeded" }] }

Para predecir mejor los límites de tarifas disponibles, es necesario considerar el uso periódico de GET application / rate_limit_status. Al igual que los encabezados HTTP que limitan la velocidad, la respuesta de este recurso indicará el estado del límite de velocidad para el contexto de la llamada; cuando se usa el OAuth 2.0 Bearer Token, los límites pertenecerán a ese contexto de autenticación. Cuando se utiliza el OAuth 1.0a User Context, los límites pertenecerán al contexto de usuario de aplicación.

6.4.10.4. Límites de las solicitudes GET y POST

Los límites de tasa de lectura del sistema (GET) se definen por usuario y por aplicación, mientras que los límites de tasa de escritura en el sistema (POST) se definen únicamente a nivel de cuenta de usuario. En otras palabras, para los límites de velocidad de lectura, se considera el siguiente escenario:

- Si el usuario A inicia la aplicación Z, y la aplicación Z hace 10 llamadas a la línea de tiempo de mención del usuario A en una ventana de 15 minutos, entonces la aplicación Z tiene 5 llamadas pendientes para esa ventana.
- Luego, el usuario A inicia la aplicación X, y la aplicación X llama a la línea de tiempo de mención del usuario A 3 veces, luego a la aplicación X le quedan 12 llamadas para esa ventana

- El valor restante de las llamadas en la aplicación X está aislado de la aplicación Z, a pesar de que el mismo usuario A

Esto contrasta con las asignaciones de escritura, que se definen por cada usuario. Por lo tanto, si el usuario A termina publicando 5 Tuits con la aplicación Z, entonces durante ese mismo período, independientemente de cualquier otra aplicación que abra el usuario A, esos 5 POST contarán contra cualquier otra aplicación que actúe en nombre del usuario A durante esa misma ventana de hora.

Por último, puede haber ocasiones en las que los valores de límite de tasa que se devuelven sean inconsistentes, o casos en los que no se devuelva ningún encabezado. Quizás la caché ha sido reiniciada o una caché estaba ocupada, por lo que el sistema se comunicó con una instancia diferente: los valores pueden ser inconsistentes de vez en cuando. Hay un mejor esfuerzo para mantener la coherencia, con una tendencia a dar llamadas adicionales a una aplicación si hay una inconsistencia.

6.4.10.5. Consejos para evitar tener una tarifa limitada

Los siguientes consejos están ahí para ayudarlo a codificar de manera defensiva y reducir la posibilidad de tener una tasa limitada. Algunas características de la aplicación que quizás desee proporcionar son simplemente imposibles a la luz de la limitación de la velocidad, especialmente en lo que respecta a la frescura de los resultados.

- Almacenamiento en caché: Es recomendable almacenar las respuestas de la API en la aplicación o en el sitio si espera que se utilicen mucho. Por ejemplo, no es recomendable llamar a la API de Twitter en cada carga de página de la página de destino de su sitio web. En su lugar, es mejor llamar a la API con poca frecuencia y cargar la respuesta en un caché local. Cuando los usuarios acceden al sitio web, cargan la versión en caché de los resultados.

- Dar prioridad a los usuarios activos: Si el sitio realiza un seguimiento de muchos usuarios de Twitter (por ejemplo, obteniendo su estado actual o estadísticas sobre su uso de Twitter), es recomendable solicitar solo datos para los usuarios que hayan iniciado sesión recientemente en el sitio.

- Adaptarse a los resultados de la búsqueda: Si la aplicación monitorea un gran volumen de términos de búsqueda, es recomendable consultar con menos frecuencia las búsquedas que no tienen resultados que las que sí lo tienen. Al usar un retroceso, se puede mantener actualizado sobre las consultas que son populares, pero no se desperdicia ciclos que solicitan consultas que rara vez

cambian. Alternativamente, es conveniente usar las API de transmisión y filtrar los términos de búsqueda.

- Utilizar el OAuth 2.0 Bearer Token como "reserva": Las solicitudes que utilizan el OAuth 2.0 Bearer Token se evalúan en un contexto separado de los límites de tasa por usuario de una aplicación. Para muchos escenarios, es posible que sea mejor utilizar este grupo de límite de tasa adicional como una "reserva" para las operaciones típicas basadas en el usuario.
- Lista de denegación: Si una aplicación abusa de los límites de tarifas, será denegada en la lista. Las aplicaciones de la lista denegadas no pueden obtener una respuesta de la API de Twitter. Si el usuario o la aplicación han sido denegados en la lista y se cree que ha habido un error, se puede utilizar los formularios de soporte de plataforma para solicitar asistencia. Para ello, es necesario incluir la siguiente información:
    - Si se está utilizando la API REST estándar, realizar una llamada a GET application / rate_limit_status desde la cuenta o computadora que se cree que está en la lista de denegaciones.
    - Explicar por qué se cree que la solicitud fue denegada.
    - Describir en detalle cómo se ha solucionado el problema que se cree que provocó que se le denegara la lista.

6.4.10.6. Streaming API

La Streaming API tiene límites de velocidad y niveles de acceso que son apropiados para conexiones de larga duración. Aprovechar la Streaming API es una excelente manera de liberar los límites de velocidad para usos más ingeniosos de la API de Twitter.

6.4.10.7. Patrón de retroceso exponencial para streaming

Si el intento de reconexión inicial no tiene éxito, el cliente debe continuar intentando reconectarse usando un patrón de retroceso exponencial hasta que se vuelva a conectar con éxito.

Independientemente de cómo se desconecte el cliente, se debe configurar la aplicación para que se vuelva a conectar inmediatamente. Si el primer intento de reconexión no tiene éxito, es recomendable que la aplicación implemente un patrón de retroceso exponencial en los intentos de reconexión posteriores (por ejemplo, esperar 1 segundo, luego 2

segundos, luego 4, 8, 16, etc.), con un límite superior razonable. Si se alcanza este límite superior, se debe configurar el cliente para que notifique al equipo para que se pueda investigar más.

6.4.10.8. Límites por ventana por recurso

La duración de la ventana de límite de tasa de API es de 15 minutos. Hay que tener en cuenta que los puntos finales / recursos que no figuran en el gráfico anterior tienen por defecto 15 solicitudes por usuario asignado.