TESIS DE MÁSTER

Análisis y evolución del contenido de los mensajes a través de la red social de Twitter

Autor: Andriy Yatsyk
Directores:
Dr. Miguel Rebollo Pedruelo
Dra. Elena del Val Noguera y
Dra. Ángeles Calduch Losa

Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València,
Camino de Vera, s/n
46022 Valencia, Spain
Curso 2018/2019,

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a la doctora Ángeles Calduch Losa por su orientación, dedicación y paciencia a lo largo de la realización de este trabajo. Agradecer también enormemente a la doctora Elena del Val Noguera su gran colaboración, por haberme demostrado que con la dedicación y el esfuerzo uno es capaz de conseguir lo que se propone. Gracias también al doctor Miguel Rebollo Pedruelo por demostrarme que por muy absurdo que suene un comentario, bien planteado se puede convertir en una buena idea. A mis familiares y amigos y todos los que me han animado y visto el esfuerzo dedicado en esta nueva etapa de mi vida.

Índice general

Índice de figuras

Índice de cuadros

CAPÍTULO

1

Introducción

1.1. Motivación

Las redes sociales generan grandes cantidades de datos cada día, debido a que las personas tratan de descubrir por una parte lo que está sucediendo en el mundo y por otra parte quieren compartir información al instante acerca de lo que opinan sobre algún tema o bien para conectarse con amigos o familiares.

Toda esta información es aprovechable para entender las necesidades y opiniones de los usuarios acerca de diversos temas. Día a día se usa para realizar investigaciones y tomar decisiones [18, 37, 41].

Tanto para el entorno político como para los usuarios sería interesante saber de qué temas hablan los cargos gubernamentales o cómo reaccionan estos en las redes sociales a nivel profesional debido a los acontecimientos, normas y leyes que se llevan a cabo [5, 20, 40].

Para todo esto se utilizan técnicas estadísticas que permitan recabar información para responder a las cuestiones que se ha planteado [32].

1.2. Objetivos

El principal objetivo de esta tesis de máster es proponer un método de análisis de información procedente de la red social Twitter. Este análisis se realizará teniendo en cuenta el contenido de los mensajes producidos y la estructura comunicativa que generan las interacciones de los usuarios en la red social. Para ello, se utilizarán técnicas de clasificación, análisis de sentimientos y análisis estructural de redes sociales. Para alcanzar este objetivo se plantean los siguientes subobjetivos:

- Realizar un preprocesado automático de la información extraída de la red social Twitter para su posterior análisis.
- Aplicar técnicas de aprendizaje automático y análisis de sentimientos para el análisis del contenido de los mensajes extraídos de Twitter.
- Realizar un análisis de las propiedades estructurales de redes a nivel global e individual para extraer elementos significativos que nos permitan entender el comportamiento de los usuarios en determinados eventos.
- Integrar las técnicas para el análisis del contenido y las técnicas de análisis de la estructura de redes sociales para ofrecer nuevas medidas híbridas.
- Validar el análisis propuesto mediante el caso de estudio de los dos partidos políticos más relevantes en Estados Unidos.

1.3. Organización del documento

Considerando la motivación y los objetivos de la tesis, el resto del documento se organiza de la siguiente manera:

- **Capítulo 2:**
  En el segundo capítulo se ha realizado una revisión de trabajos previos relacionados con el análisis del contenido de mensajes en redes sociales y de las interacciones de los usuarios para extraer información acerca del comportamiento de los usuarios, además de plantear qué aporte hará este trabajo con respecto a los trabajos existentes.
- **Capítulo 3:**
  En el capítulo 3 se presenta una visión general de la propuesta de análisis de información

procedente de las redes sociales. Posteriormente se detallan las distintas etapas llevadas a cabo en el análisis, empezando con el preprocesado de los datos. A continuación se explica el análisis del contenido de los mensajes mediante la aplicación los algoritmos de aprendizaje automático (LDA) para la extracción de temas. También se describe la aplicación del algoritmo Hierarchically Supervised Latent Dirichlet Allocation (HSLDA), y el algoritmo de análisis de sentimientos. Además se describen los análisis de la estructura de la red llevados a cabo a nivel global y a nivel de usuario.

- **Capítulo 4:**
  En el capítulo 4 se presenta un caso de estudio donde se aplica la propuesta de análisis de información descrita en el capítulo 3. Para ello se consideran los mensajes extraídos entre los días 12 y 13 de diciembre del 2018 que mencionan a alguno de los partidos políticos más representativos de Estados Unidos. En el caso de estudio se interpretarán los resultados ofrecidos por los distintos tipos de análisis.

- **Capítulo 5:**
  En este capítulo se comentarán las principales conclusiones de la tesis y posibles trabajos futuros.

CAPÍTULO

2

Estado del Arte

2.1. Redes sociales

Gran parte de la información que se genera en hoy en día proviene de redes sociales como LinkedIn, Facebook, Twitter o YouTube entre otras, ya que son usadas por millones de personas. Por ejemplo, en la región de América del Norte, según el artículo de Gordon [14], en 2017 el 60 % de la población tiene al menos una cuenta en una red social.

El tipo de usuario que está presente en las redes sociales no tiene un perfil en concreto, ya que puede ser cualquier ciudadano. Hay cuentas con muchos seguidores como son las celebridades, representantes de empresas, políticos e incluso altos cargos como presidentes de países. Gracias a estos diferentes perfiles es posible recopilar contenido de texto de diferentes usuarios o grupos sociales.

Todas estas condiciones crean un nicho potencial de fuente de datos que pueden ser utilizados para obtener opciones de usuarios, análisis de sentimientos de mensajes [15], control de actividades [2], análisis de grupos políticos [40], etc.

La red social Twitter trata un gran número de textos diariamente. Con toda esa información se abre un gran abanico de oportunidades de investigación, y con las herramientas adecuadas se puede analizar desde cómo actúa un partido político en la red social hasta comparaciones con

otros partidos. Herramientas de software como R Studio o lenguajes de programación como Python permiten trabajar con una gran cantidad de datos mediante diferentes librerías con el fin de analizar y tratar la información.

Este trabajo investiga los temas de los que hablan diferentes partidos políticos, tanto por parte del gobierno como de la oposición. Esto se ha realizado mediante el análisis del contenido de los tuits y de las interacciones generadas por cuentas asociadas a partidos políticos. Concretamente, se ha analizado los sentimientos asociados a temas latentes de los que se habla en determinadas comunidades de usuarios, así como la estructura comunicativa que emerge de las interacciones entre usuarios.

2.2. Twitter

El planteamiento presentado en este trabajo está relacionado con el tratamiento de información en redes sociales, en este caso centrado en Twitter por su flexibilidad de permisos. Twitter es una plataforma social que combina aspectos de redes sociales y mensajería instantánea. En sí, es un modo de comunicación rápido y sencillo. Los usuarios registrados en esta plataforma pueden publicar mensajes cortos de estado, noticias, enlaces, fotos, vídeos, mencionar a otros usuarios, etc. [33].

Según afirma Wang [42], Twitter es la plataforma tipo microblog abierta más popular, tiene aproximadamente 320 millones de usuarios activos [19]. Mediante esta nueva forma de socialización, los usuarios tienen la posibilidad de publicar tuits sobre distintos aspectos de su vida cotidiana, desde el desarrollo profesional hasta actualizaciones personales y familiares.

Twitter fue originalmente fue diseñado para ser utilizado con servicios de mensajería de texto de teléfonos móviles [24]. La brevedad de formato y la restricción a 140 caracteres por cada tuit crean un canal de comunicación en forma de reto para los usuarios que pueden expresarse de forma informal y rápida.

La información personal que se divulga a través de esta plataforma, en la sección de perfil de usuario, es reducida, opcional y breve. Por lo general solo se coloca el nombre, la ubicación, una breve biografía de 160 caracteres y una dirección web (en caso de que se disponga de alguna). Desde su creación en 2007, Twitter se ha considerado más que una aplicación de redes sociales una plataforma de noticias, comentarios, opiniones, marketing, activismo político, fotos compartidas, documentación de eventos y conversaciones entre otros aspectos. El acceso a los pensamientos, intenciones, cooperaciones y actividades de millones de usuarios en tiempo

real ha creado un potente canal para entender lo que está pasando en el momento en cualquier lugar del mundo.

De acuerdo con estudios realizados por Ma et al [24], este servicio de mensajería cubre un gran número de cuestiones, tal como se ha mencionado, incluyendo comentarios sobre diversos temas de actividades personales, política, noticias y muchos más. Debido a su corta extensión, los usuarios recurren a introducir en el texto etiquetas llamadas hashtags es decir, palabras clave prefijadas con el símbolo (#), lo cual los hace más llamativos.

Los hashtags han demostrado ser muy efectivos para organizar la información en Twitter, hasta tal punto que es posible ver en la plataforma el hashtag que predomina en una región del mundo. También mejoran la información y la búsqueda de los tuits, facilitando la interacción social. De acuerdo con una investigación que realizaron en Ma et al [24], el 58 % de los usuarios de Twitter utiliza un hashtag en sus comentarios, por lo que éste se ha convertido en una característica clave en muchas redes sociales como Telegram, FriendFeed, Facebook, Instagram, entre otras redes.

Si los tuits a analizar pertenecen a usuarios con cuentas verificadas, es decir, son certificadas por entidades terceras, se demuestra que la cuenta es realmente del usuario o la persona que la creó. En este trabajo, las cuentas a analizar están relacionadas con partidos políticos o representan a un partido y utilizan hashtags concretos y mencionan a otros usuarios que son o suelen ser relevantes en el tema.

En los últimos años se han realizado diversas investigaciones relacionadas con los comentarios que exponen las personas en las redes sociales. En las siguientes secciones se abordarán algunos artículos en los que se han analizado el contenido de tuits y/o la estructura comunicativa que emerge en base a los tuits.

2.3. Análisis del contenido de los mensajes

La extracción de los topicos cada vez es más utilizada en diferentes ámbitos de minería de textos. Uno de los algoritmos más utilizados es el LDA (Linear discriminant analysis) [23], que se ha extendido de diferentes formas y particularmente para las redes sociales, disponiendo de varias extensiones para ser utilizado. Un buen ejemplo es el presentado por Chang et al. [10], que propusieron un nuevo modelo probabilístico de temas para inferir descripciones y las relaciones entre las entidades que se estudian.

La idea de unir un algoritmo como LDA junto con un análisis de sentimientos es empleada

en muchos propósitos. A veces, el objetivo es identificar los sentimientos de los usuarios en la red social a través de sus conversaciones sobre un tema como en el artículo de Naskar et al. [28], donde analizan si los usuarios tienden a reunirse de acuerdo con la semejanza de sus sentimientos.

Destacar que el artículo de Wang [42] inspirado por el principio de homofilia (en el ámbito de las redes sociales se utiliza para referise a la asortatividad), sugiere que las opiniones están influenciadas por las conexiones de los usuarios.

En el artículo escrito por Torres Nabel [39] se presenta una red de "influencers" para explicar los fenómenos virales que suceden en las redes sociales.

Se desarrolla un análisis a partir del primer estudio basado en la metodología de big data en México, donde el Instituto Nacional de Estadística y Geografía expuso el estado de ánimo de los tuiteros en México a partir de 63 millones de tuits desde el 1 febrero de 2014 hasta el 15 de mayo de 2015. El artículo abarca el algoritmo de las tendencias en las redes sociales, estudio de las tendencias emocionales en las redes sociales y consecuencias del estado de ánimo.

El artículo presentado por Artcila Calderón et al. [3] describe y evalúa la técnica de análisis supervisado de sentimientos en comunicación política. Se describen las técnicas utilizadas para el análisis de sentimientos en la comunicación política y tratamiento distribuido para grandes cantidades de datos.

También es relevante el articulo de Concha et al. [11] que se centran en la influencia política antes de las elecciones.

2.4. Análisis de las estructuras comunicativas

En el contexto del análisis de redes sociales hay un interés desde el punto de vista empresarial. Actualmente, existen empresas que ofrecen servicios para el análisis de redes sociales como Audiense[1] o TweetBinder [2]. Estas herramientas ofrecen un análisis estático de la red en un determinado punto de tiempo. Normalmente utilizan métricas que se centran principalmente en datos estadísticos sobre los mensajes transmitidos en la red y dejan en un segundo plano medidas estructurales que están presentes en el área de Redes Complejas.

El análisis de la estructura de redes sociales tiene una aplicación directa en diversas áreas. Concretamente, el análisis de redes sociales ha sido utilizado para campañas de marketing on-line

[1] `https://audiense.com/`
[2] `https://www.tweetbinder.com/`

y estrategias de persuasión [21], diseño de interfaces de usuario, sistemas de recomendación, para determinar consumidores potenciales [1] o para determinar la personalidad de los usuarios a través de sus interacciones [13]. El análisis de las interacciones en las redes sociales facilita la comprensión de los flujos de información explicando como un mensaje puede llegar a un usuario que no esta relacionado con este y la localización de usuarios tienen una posición influyente en la red "influencers" [29, 34]. Otros ámbitos donde se ha aplicado el análisis de la redes sociales son la personalización de los resultados de búsquedas basándose en los intereses de nodos vecinos en la red [9], así como para el terrorismo [30] o el cyberbullying [6].

Los análisis de redes se centran en un análisis estático de las propiedades de la topología de la red, como por ejemplo la distribución del grado de conexión, la longitud de caminos, el coeficiente de clustering, o propiedades de centralidad en un instante de tiempo determinado [17, 26, 38].

Otros trabajos que analizan las redes sociales de eventos políticos o conferencias profesionales se centran principalmente en la evolución de las estadísticas de la información generada por los usuarios de la red a lo largo de un evento [7, 12, 22, 27] y en cómo se difunde esa información dependiendo de la temática [35], dejando a un lado el análisis de las propiedades estructurales de la red.

En el contexto del análisis de redes sociales hay un interés desde el punto de vista empresarial. Actualmente, existen empresas que ofrecen servicios para el análisis de redes sociales como Audiense[3] o TweetBinder [4]. Estas herramientas ofrecen un análisis estático de la red en un determinado punto de tiempo. Normalmente utilizan métricas que se centran principalmente en datos estadísticos sobre los mensajes transmitidos en la red y dejan en un segundo plano medidas estructurales que están presentes en el área de Redes Complejas.

Destacar la hipótesis que se propone en la investigación de [25] en redes sociales. Este estudio trata de analizar cómo las redes públicas dan forma a una red global y facilitan la comunicación entre comunidades con diferentes orientaciones políticas. Se analiza la posibilidad de que los perfiles políticos generen interacciones al introducir contenido partidista en los flujos de información, cuya audiencia principal consiste en usuarios opuestos ideológicamente.

[3] `https://audiense.com/`
[4] `https://www.tweetbinder.com/`

CAPÍTULO

3

Propuesta

3.1. Introducción

El análisis de redes sociales mediante diferentes algoritmos permiten identificar fenómenos que muchas veces no se ven, o no se manifiestan a simple vista. En este capítulo se describe la propuesta planteada junto a su diseño y estructura. La propuesta consiste en un proceso de análisis de datos extraídos de una red social que comienza con la limpieza de los tuits para posteriormente aplicar algoritmos y otras técnicas de análisis que nos permitan extraer nueva información. Entre los análisis que se van a tener en cuenta están: un análisis de contenido de los mensajes, pudiendo obtener de este modo los temas más relevantes de los que se habla y el sentimiento de los tuits en el ámbito de los partidos políticos, un análisis de la estructura comunicativa, y un análisis que combina características del análisis del contenido y de la estructura.

En este capítulo se plantea el reto de poder extraer información concreta sobre si hay usuarios hablando de un tema o temas concretos, si los usuarios tienen una percepción positiva o negativa sobre éstos, o qué estructura comunicativa está presente en sus interacciones.

En el diagrama de la Figura 3.1 se puede ver la estructura del proceso que se ha realizado. El

proceso de análisis esta distribuido en las siguientes etapas:

1. Etapa: Extracción de los tuits.
2. Etapa: Tratamiento de datos.
3. Etapa: Análisis del contenido de los mensajes.
4. Etapa: Análisis sobre la estructura comunicativa.
5. Etapa: Análisis híbrido.

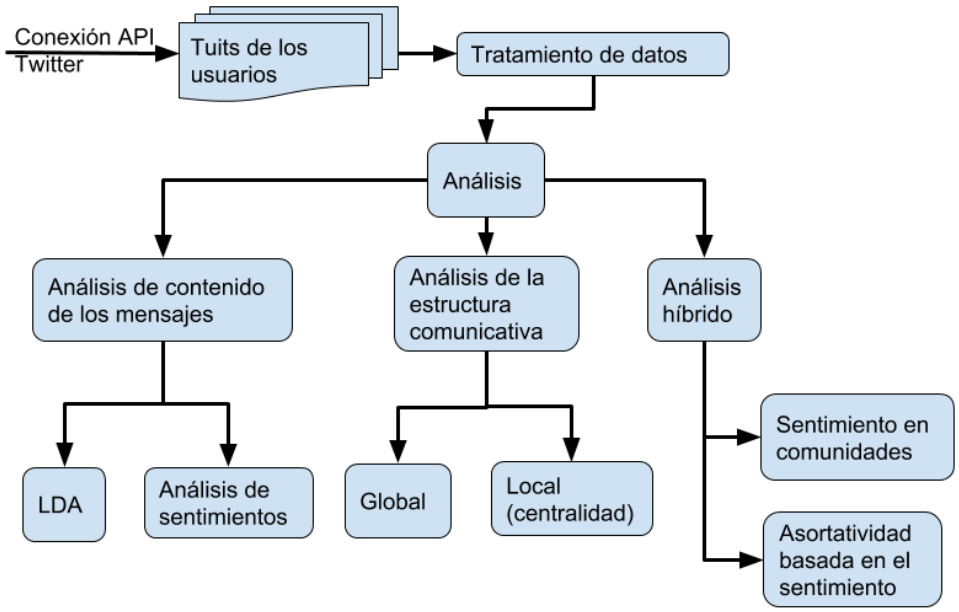

Figura 3.1: Estructura del proceso de análisis

3.2. Extracción de los tuits y tratamiento de datos

El primer reto que se ha planteado es la gran abundancia de información en Twitter que hay que extraer, filtrar y posteriormente analizar. Para ello se hizo uso de la API de Twitter para poder extraer todos los tuits relacionados con un tema específico. Para utilizar el API es necesario disponer de unas credenciales. Concretamente api_key, api_secret, access_token, y access_token_secret. Los datos recogidos se almacenan en una base de datos MongoDB [1] para poder recuperarlos posteriormente.

[1] `https://www.mongodb.com/es`

Una vez que tenemos la información extraída de la red social, pasaremos al filtrado y limpieza de los datos. De todos los campos que tiene un tuit, nos quedaremos sólo con aquellos que nos interesan para los análisis posteriores (ver Tabla 3.1).

Campo	**Descripción**
id	ID del tuit.
text	Texto del tuit.
user.screen_name	Nombre de usuario propietario del tuit o retuit.
entities.user_mentions.screen_name	Nombre de los usuarios que se ha mencionado en el tuit.
created_at	Fecha de creación del tuit.

Cuadro 3.1: Campos extraídos del tuit

Con el uso de la librería de NLTK **(Natural Language Toolkit)** y de las expresiones regulares, que facilitan el trabajo de modificaciones masivas, se realiza un proceso de limpieza de contenido que no es útil para algunos de los análisis posteriores. Concretamente, se aplican las siete reglas siguientes:

1. Regla: Eliminar los caracteres especiales como "$, &" seguido de texto.

2. Regla: Quitar los hipervínculos.

3. Regla: Eliminar el símbolo de la almohadilla.

4. Regla: Eliminar palabras que no aportan nada relevante, es decir, formadas por uno o dos caracteres, dejando solo las más relevantes.

5. Regla: Eliminación de espacios en blanco seguidos.

6. Regla: Normalización de las palabras, esto se debe a que hay palabras escritas de forma incorrecta o algunas de las letras son repetidas, sea por equivocación de autores u otro efecto.

7. Regla: Eliminación de las palabras vacías, también más conocidas como "stop words".

Una vez realizados los apartados anteriores, se procede a llevar a cabo diferentes análisis del contenido de los mensajes.

3.3. Análisis del contenido de los mensajes

En esta sección se plantea un análisis del contenido de los mensajes. Primero se aplicará un algoritmo no supervisado (LDA) para la extracción de los temas latentes dentro de los tuits. Posteriormente aplicaremos algoritmos de análisis de sentimientos para poder observar la actitud de los usuarios con respecto a determinados temas, en base a sus publicaciones en la red social.

3.3.1. Análisis de LDA

Las siglas de LDA son de Latent Dirichlet Allocation, que corresponde a una técnica de aprendizaje no supervisado, es decir, en un principio no se conoce el objetivo buscado (no hay pautas predefinidas y puede que no se conozca el número de grupos de resultado). El principal objetivo es identificar los temas más relevantes en la información existente.

Al utilizar LDA como una orientación de "contenedor de palabras", significa que las palabras del documento son totalmente intercambiables, por lo que su orden no es importante. Cada documento tiene una distribución de probabilidad sobre algunos temas, mientras que cada tema se representa como una distribución de probabilidad sobre un número de palabras, como se explica en el artículo de Hong and Davison [16].

Por ejemplo, dentro del contenido de documentos sobre la Universidad Politécnica de Valencia, habría ponencias individuales que forman parte del Departamento de Informática de Sistemas y Computadores. Hay una probabilidad de que haya algunas palabras que se utilicen con más frecuencia cuando se habla del Departamento de Informática de Sistemas y Computadores que de otros departamentos en la universidad, tales como: computadoras, algoritmos, gráficos, cadenas, datos, modelado de software y redes. Otros departamentos, como Ingeniería de Alimentos, pueden tener temas como cadenas, alimentación, simulación de procesos, microestructura de alimentos.

LDA ve toda la información de forma global y a partir de ese punto elige los temas. Si se han analizado los documentos de forma individual, ciertos temas podrían no estar recogidos, y sólo cuando se ve todo el cuerpo se empiezan a notar ciertos tópicos. En este ejemplo, palabras como "cadenas" pueden aparecer varias veces en los documentos independientemente del departamento. De esta manera se crea un modelo más realista del cuerpo, y por lo tanto, de los documentos individuales.

También las palabras que aparecen con menos frecuencia en los documentos únicos, pero son

comunes en muchos de ellos, probablemente estén indicando que hay un tema en común entre los documentos. La capacidad de recoger los matices de los tópicos permite destacar la información más relevante tanto incluida con menos posibilidades de repetición como con más, y dar así un mejor resultado.

Para entender con más claridad el proceso anteriormente explicado, a continuación se presenta un ejemplo sencillo del funcionamiento del LDA.

Supongamos que tenemos las siguientes oraciones:

- Me gusta programar en Python y en java.
- Esta mañana programé una aplicación web en PHP y Javascrip.
- Los gatos y los perros son mascotas.
- Mi primo adoptó un gato ayer.
- Mira esta web programada en Python para un refugio de animales.

Dadas las oraciones anteriores, los elementos con los que trabaja el LDA son:

- Documento: cada una de las oraciones contenidas en el ejemplo, en este caso hay cinco documentos.
- Corpus: formado por el conjunto de los cinco documentos.
- Palabras: Cada uno de los ítem que conforman los documentos, sin tomar en cuenta las palabras vacías, es decir, las que por sí solas no tienen ningún significado.

En este caso, lo que hace el LDA es descubrir automáticamente los temas que contienen el grupo de oraciones. Por ejemplo, si se le pidieran dos temas al algoritmo, realizaría un proceso como el siguiente:

- Oraciones 1 y 2: 100 % del Tema A
- Oraciones 3 y 4: 100 % del Tema B
- Oración 5: 60 % del Tema A y 40 % del Tema B

Los temas quedarían distribuidos de la siguiente manera:

- Tema A: 30 % Python, 15 % java, 10 % programar, 10 % aplicación... (en este caso se puede interpretar que este tema está relacionado con la categoría de programación).
- Tema B: 20 % gatos, 20 % perros, 20 % mascotas, 15 % animales... (en este caso se interpretaría como una categoría de animales).

3.3.2. Análisis de sentimientos

La clasificación de las emociones, el medio por el cual una persona puede distinguir una emoción de otra, es un tema complejo en la investigación. Se ha abordado por parte de los investigadores la clasificación de las emociones desde dos puntos de vista fundamentales:

1. Las emociones son construcciones discretas y fundamentalmente diferentes. En esta teoría, se piensa que todos los humanos tienen un conjunto innato de emociones básicas que son reconocibles a través de la cultura. Estas emociones básicas se describen como "discretas" porque se cree que son distinguibles por la expresión facial y los procesos biológicos de un individuo.
2. Las emociones pueden caracterizarse sobre una base dimensional en agrupaciones. Eugene Bann [4] propuso una teoría según la cual las personas transmiten su comprensión de las emociones a través del lenguaje que usan, que rodea a las palabras clave de emoción mencionadas. Él afirma que cuanto más distinto es el lenguaje que se usa para expresar cierta emoción, entonces más clara es la percepción de esa emoción, y por lo tanto más básica. Esto nos permite seleccionar las dimensiones que mejor representan el espectro completo de la emoción.

Para el análisis de sentimientos tendremos en cuenta modelos que siguen el punto de vista propuesto por Eugene Bann.

Modelo del circunflejo

Este modelo fue desarrollado por James Russell [36]. Este autor sugiere que las emociones se distribuyen en un espacio circular bidimensional, que contiene diferentes dimensiones de representación y valencia. El nivel de atención representa el eje vertical y la valencia representa el horizontal, mientras que el centro del círculo representa una valencia neutral y un nivel medio de atención (ver Figura 3.2). El modelo puede representarse en cualquier nivel tanto horizontal como vertical. Los modelos circunflejos se usan en ámbitos de mayor frecuencia para probar los estímulos de las palabras de emoción.

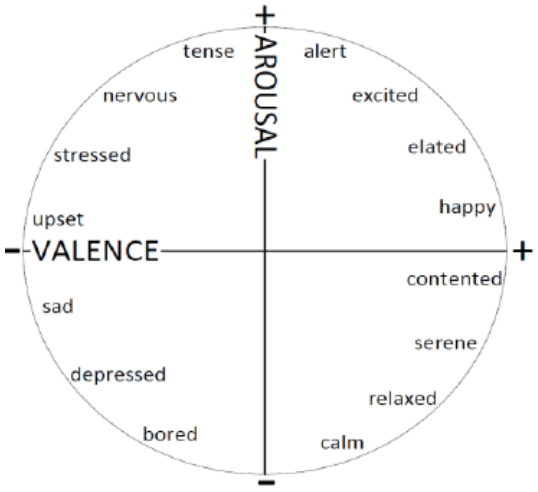

Figura 3.2: Modelo circunflejo

Modelo del vector

En el caso de "Modelo del vector" apareció por primera vez en 1992 [8]. Este modelo bi-dimensional consta de vectores que apuntan en dos direcciones, representando una forma de "boomerang" (ver Figura 3.3). Dicho modelo asume que siempre hay una dimensión en la que se encuentra una emoción particular.

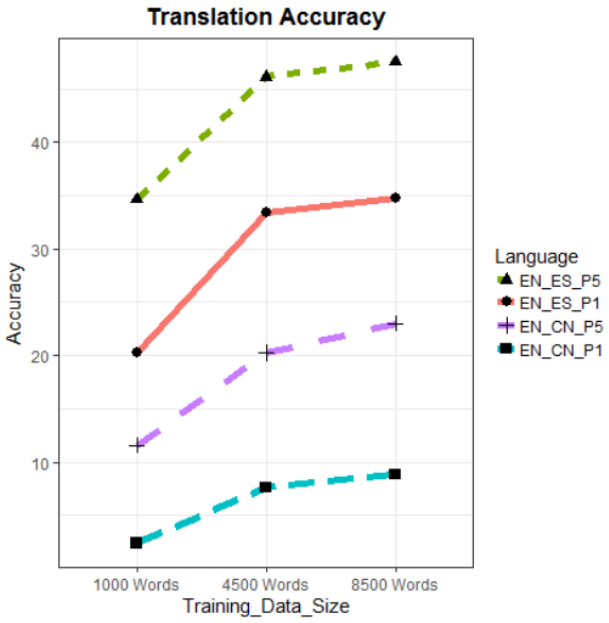

Figura 3.3: Modelo del vector

Modelo del clasificador Bayesiano ingenuo

Este modelo está basado en el teorema de Bayes con algunas simplificaciones que se resumen en la la hipótesis de independencia entre las variables que se predicen. Este modelo asume que la presencia o ausencia de una característica particular no está o no se relaciona con la presencia de cualquier otra. Por ejemplo, en una red social consideraremos que un perfil es seguidor de la política si hace comentarios diarios, si su profesión esta orientada a ese ámbito y si siempre habla de noticias con perfiles similares al suyo.

$$p(C \mid F_1 ... F_n) = \frac{p(C)p(F_1, ..., F_n \mid C)}{p(F_1, ..., F_n)}$$

Modelo de Plutchik's

El modelo de Robert Plutchik [31] plantea un modelo tridimensional que organiza las emociones en círculos concéntricos, donde los círculos internos son más básicos y los círculos externos más complejos. Los círculos externos también se forman mezclando las emociones del círculo interno (ver Figura 3.4). El modelo de Plutchik, como el de Russell, emana de una representación circunfleja, donde las palabras emocionales se representan en función de la similitud.

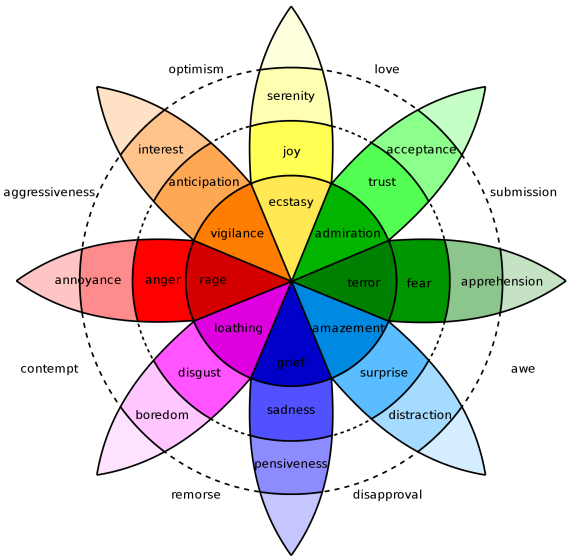

Figura 3.4: Modelo de Plutchik's

Modelo de PANA

La activación positiva - activación negativa "PANA", originalmente creado por Watson y Tellegen

[43], comenta que el afecto positivo y el negativo son dos sistemas separados. Similar al modelo vectorial, los estados de mayor excitación tienden a definirse por su valencia, y los estados de menor excitación tienden a ser más neutrales en términos de valencia (ver Figura 3.5).

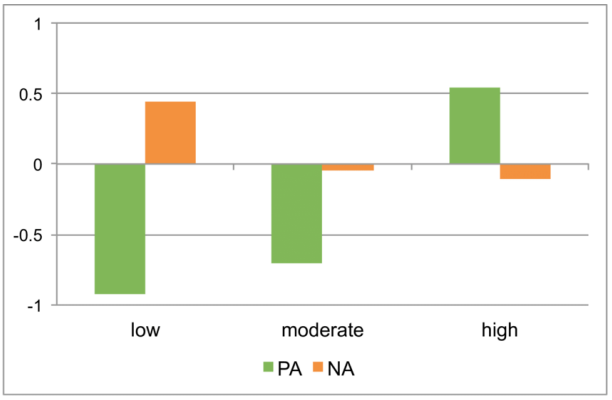

Figura 3.5: Modelo de PANA

Dado que los tuits están relacionados con la política, el contenido de estos son muchas veces textos formales con una calidad de expresión neutra del sentimiento, es decir, no se clasifican como positivos y tampoco como negativos.

En el lenguaje de Python existen diferentes librerías que ayudan a implementar los modelos citados en este documento. En este estudio se hizo uso de la librería "TextBlob" (procesamiento de texto simplificado), que por debajo realiza un análisis de sentimientos mediante el modelo del clasificador Bayesiano ingenuo.

Esta librería, a partir del contenido de un tuit, nos proporciona como resultado dos variables:

Polaridad: Esto nos indica si el sentimiento del contenido del tuit es positivo o de lo contrario negativo. Representado por un número donde el valor cero es un tuit neutral, valores menores que cero se consideran negativos y valores mayores que cero son los positivos.

Subjetividad: El rango de los valores que representan la subjetividad son entre cero y el uno, donde cero son los tuits clasificados como muy objetivos y uno como muy subjetivos.

Para que los resultados no fueran representados de forma complicada y nos proporcionen la polaridad de "TextBlob" se ha utilizado el modelo PANA (Positive - Negative activation). De esta manera el análisis y tratamiento los resultados obtenidos es más sencillo, clasificando el contenido de un tuit como a favor o en contra.

3.4. Análisis de la estructura comunicativa

En esta sección se presentan un conjunto de medidas utilizadas dentro del área de análisis de redes sociales. Estas medidas proporcionan información sobre el tipo de estructura de red y también sobre la relevancia de determinados nodos dentro de ellas. Las medidas se pueden agrupar en dos niveles teniendo en cuenta el nivel de abstracción que analizan: global y local.

Dentro del contexto de este trabajo, consideraremos que una red está compuesta por nodos y enlaces. Los nodos representarán a los usuarios que han generado mensajes sobre el tema a analizar y los enlaces representarán las interacciones entre los usuarios de la red. Estas interacciones son referencias que unos usuarios han hecho a otros en sus mensajes (i.e., retuit, mención, respuesta). Formalmente, la estructura de la red se podría representar con la notación que describimos a continuación: $\mathcal{G}$ representa una red que está formada por un conjunto de nodos N, donde cada nodo $i \in \{1, ..., n\}$ representa a una cuenta de usuario. Los nodos establecen relaciones que se representan como enlaces $E \in N \times N$. Para representar la red de nodos y sus relaciones se puede utilizar también la matriz de adyacencia $\mathbf{A}$. Teniendo en cuenta esta matriz, un enlace (relación) entre un nodo i y un nodo j se representa como $A_{ij} = 1$. Si no existe un enlace entre i y j se representa como $A_{ij} = 0$. Teniendo en cuenta qué representa una red en este trabajo, vamos a definir propiedades a nivel global y a nivel local que son interesantes para analizar mensajes generados en Twitter.

3.4.1. Medidas a nivel global

Para poder realizar un análisis de la estructura global de la red asociada a las interacciones entre usuarios como un todo, vamos a empezar describiendo las medidas a nivel global. Estas medidas determinan propiedades estructurales o topológicas que proporcionan información sobre cómo se relacionan unos usuarios con otros. Estas medidas nos ayudan a caracterizar las redes y poder determinar si siguen un determinado patrón ya definido dentro del área del análisis de redes sociales (i.e., small world, scale-free, o random).

Entre las medidas que existen a nivel global, en este trabajo nos vamos a centrar en: el número de nodos, el número de enlaces, la longitud media de los caminos, el diámetro de la red, la densidad, la modularidad, la asortatividad, la componente gigante y la componente k-core. A continuación se describen cada una de ellas y se explica su significado dentro del contexto del trabajo.

Nodos. Representa el número de cuentas de usuario dentro de la red social que han participado mediante un mensaje, ya sea generándolo o siendo mencionadas en él. Es un indicador de la participación alrededor de un tema.

Enlaces. Los enlaces representan las relaciones de interacción entre cuentas de usuarios a través de mensajes que pueden ser menciones, retuits y mensajes de respuesta. Los enlaces nos permiten analizar cómo se ha interactuado e intercambiado información entre las cuentas de usuarios. También pueden ser utilizados como un indicador de actividad de interacción.

Densidad. La densidad es una propiedad que nos permite determinar el grado de cohesión de la red. Esta medida consiste en dividir el número de enlaces existentes entre el número total de enlaces posibles. Es interesante para determinar si las interacciones entre cuentas de usuarios se están produciendo de manera global o sólo en determinadas partes de la red. En el caso de que se consideren redes dirigidas, la densidad también nos permite detectar si existe reciprocidad en las interacciones entre cuentas de usuario. Por ejemplo, en algunos eventos hay cuentas de usuarios que escriben mensajes a otras cuentas de entidades famosas pero no suelen recibir respuesta. En estos casos, si consideramos la red dirigida, la densidad tendrá un valor bajo.

Longitud media de los caminos y diámetro de la red. Mide la distancia media entre todos los pares de nodos. En nuestro caso, representa la distancia media entre cuentas de usuario que han formado parte de una publicación, bien porque la han generado o porque han sido mencionados dentro de ella. Formalmente, en una red $\mathcal{G}$ con un conjunto de nodos N, siendo $d(i, j)$ la distancia más corta entre i y j y asumiendo que $d = (i, j) = 0$, si i no puede ser alcanzado por j, se puede definir la longitud media del camino más corto l_g en la red como:

$$l_g = \frac{1}{|N| \cdot (|N| - 1)} \cdot \sum_{i \neq j} d(i, j)$$

donde $|N|$ es el número de nodos de la red $\mathcal{G}$.

El diámetro es la distancia máxima de entre todas las distancias cortas entre dos nodos (en nuestro caso cuentas de usuario) de la red. Estas medidas nos sirven de indicadores de la actividad de la red que estamos analizando. En redes donde la actividad interaccionando (i.e., generando mensajes que involucren a otras cuentas) sea alta, el diámetro y el camino medio será menor que en redes donde la actividad sea simplemente la generación de mensajes que no impliquen interacción. También son medidas que nos pueden dar una aproximación de la

capacidad de difusión de un mensaje dentro de red.

Modularidad. La modularidad es la fracción de los enlaces que caen dentro de los grupos establecidos menos la fracción esperada si los enlaces se distribuyeran al azar. Para una agrupación determinada de los vértices de la red, la modularidad refleja la concentración de los enlaces dentro de los módulos en comparación con la distribución aleatoria de enlaces entre todos los nodos independientemente de los módulos. La modularidad nos permite detectar si hay cuentas de usuarios que son más propensas a interactuar con un grupo de usuarios que con otros. Esta propiedad nos permite detectar comunidades en base a las interacciones (enlaces).

Asortatividad. Esta propiedad indica la preferencia de los nodos de una red por unirse a otros que le son similares en base a algunos de sus atributos. Normalmente, el atributo que se suele utilizar para analizar la asortatividad de una red es el grado de los nodos. El coeficiente de asortatividad r se trata del coeficiente de correlación de Pearson de los grados entre dos pares de nodos conectados. Valores $r > 0$ indican que existe una correlación entre nodos con grado similar. Un valor negativo de r indica correlaciones entre nodos de diferente grado. En general, r toma un valor comprendido entre -1 y 1. Cuando r = 1, se dice que la red es totalmente asortativa, cuando r = 0 la red es no asortativa y cuando r = -1 la red es disortativa. En el caso de las redes sociales, los nodos con alta conectividad tienden a conectarse a otros que también tienen un alto grado de conectividad.

Componente gigante. Si una red está formada por varias componentes, la componente gigante representa al conjunto de nodos enlazados entre si y que agrupan a la mayoría de los nodos de la red.

Componente K-Core. Especifica un grupo de nodos, los cuales están conectadas a al menos un número k nodos dentro del grafo, es decir, nodos que al menos tienen grado k.

3.4.2. Medidas a nivel de nodo

Las medidas a nivel de nodo se centran en determinar la relevancia de los nodos (usuarios) dentro de la red teniendo en cuenta distintos criterios. Existen muchas aproximaciones para determinar la relevancia de un nodo. Dependiendo de lo que nos interese analizar, nos centraremos en unas medidas o en otras. En el proyecto hemos considerado las siguientes: centralidad

basada en el grado, centralidad basada en el vector de valores propios, centralidad basada en cercanía y centralidad basada en intermediación.

Centralidad basada en grado. La relevancia de un nodo se puede determinar en base a su grado dentro de la red, es decir, el número de enlaces conectados al nodo. En el caso de una red social, de manera intuitiva, podemos deducir que los nodos que tengan un número de enlaces elevado puedan tener más acceso a más información o más relevancia que aquellos nodos que tienen menos.

Formalmente, la centralidad basada en grado de un nodo i se puede calcular de la siguiente manera:

$$C_{D_i} = \sum_j A_{ij}$$

Centralidad basada en el vector de valores propios. Esta medida es similar a la anterior, pero en este caso todos los vecinos (enlaces) de un nodo no tienen la misma importancia. Si un nodo está conectado a nodos que a su vez están bien conectados, se puede decir que este nodo es relevante.

Formalmente, se puede definir la centralidad basada en vectores propios de un nodo v como:

$$C_{V_i} = \frac{1}{\lambda} \sum_{j \in k_i} C_{v_j} = \frac{1}{\lambda} \sum_{j \in \mathcal{G}} A_{ij} C_{v_j}$$

donde k_i es el conjunto de vecinos de i y λ es una constante.

En el contexto del trabajo, una cuenta de usuario que tenga valores propios grandes implica que es una autoridad con la que se interactúa por los contenidos/mensajes que genera dentro o fuera de la red y que a su vez esta cuenta interactúa con otras cuentas. Una medida muy relacionada con la centralidad basada en valores propios es el PageRank.

Centralidad basada en cercanía. Esta medida de centralidad tiene en cuenta la distancia media de un nodo al resto de nodos. Formalmente, la cercanía de un nodo i se define como:

$$C_{CLO_i} = \frac{|N|}{\sum_j S_{ij}}$$

donde S es la matriz cuyos elementos (i,j) corresponden a la distancia más corta desde el nodo i hasta el nodo j.

En nuestro caso, la centralidad basada en cercanía puede utilizarse como un indicador de la rapidez de la difusión de la información desde un usuario concreto de la red a todos los demás. Si un usuario tiene un grado alto de centralidad basada en cercanía, será interesante utilizarlo como difusor de información. También puede verse como un indicador de accesibilidad. Si un nodo tiene un valor alto de este tipo de centralidad, será un indicador de que es cercano al resto de usuarios de la red.

Centralidad basada en intermediación. Esta medida de centralidad está basada en la intermediación. Esta medida tiene en cuenta para un nodo a cuántos caminos entre otros dos nodos de la red (i, j) pasan a través de él mismo (a). Si un nodo de la red tiene un valor de intermediación alto, puede ser considerado como relevante en una red ya que "hace de puente"de la información entre grupos de nodos (comunidades). En el proyecto, una cuenta de usuario que tenga un alto grado de intermediación puede considerarse como influyente dentro del proceso de comunicación, pudiendo controlar su flujo.

Formalmente, la intermediación C_{BET_i} de un nodo i en una red se define como:

$$C_{BET_i} = \sum_{j,k} \frac{b_{jik}}{b_{jk}}$$

donde b_{jk} es el número de caminos más cortos desde el nodo j hasta el nodo k, y b_{jik} el número de caminos más cortos desde j hasta k que pasan a través del nodo i.

3.5. Análisis híbrido

El análisis híbrido que se plantea en el trabajo integra el análisis del contenido de los mensajes con el análisis de la estructura comunicativa que generan los mismos. El objetivo es tener una visión más completa de qué es lo que está pasando en la red social. En este análisis vamos a tener en cuenta el sentimiento de las comunidades de usuarios, el sentimiento de las interacciones entre usuarios, los temas que hablan en las distintas comunidades de usuarios y la agrupación de usuarios en base al sentimiento.

Para analizar el *sentimiento por comunidades* de usuarios se ha calculado la modularidad de la red de interacciones. En este caso, se ha considerado que la red de interacciones es dirigida.

Esto nos permitirá diferenciar entre enlaces de entrada (in-degree) y de salida (out-degree), es decir, mensajes que hacen referencia al usuario y mensajes que genera el usuario. Para establecer los grupos ('clusters') de usuarios, hemos utilizado la propiedad estructural de la red de modularidad. Por cada cluster, se han analizado los tuits de los usuarios miembros, estableciendo su sentimiento. El sentimiento de un usuario dentro del cluster se ha considerado que será el sentimiento promedio de los tuits que ha generado, es decir, de los enlaces de salida (out-degree). Con este análisis podremos determinar si el hecho de pertenecer a una comunidad desde el punto de vista estructural de la red implica compartir un sentimiento. También nos permite detectar si el sentimiento es común en todas las comunidades o si por el contrario se mantiene local a la comunidad.

Para analizar el *sentimiento de los enlaces* hemos considerado que la red era no dirigida. Considerar la red no dirigida nos permite evaluar el total de interacciones entre dos nodos sin diferenciar quién generó el mensaje. En este caso, nos interesa evaluar el sentimiento de la relación de comunicación en global. Para ello, para cada enlace hemos localizado los mensajes asociados, hemos calculado el sentimiento de cada mensaje y finalmente hemos agregado todos los valores del sentimiento. El valor del sentimiento del enlace será el agregado de todos los valores del sentimiento de los mensajes entre el total de mensajes intercambiados.

El análisis de *tópicos por comunidades* nos permite detectar si hay determinados temas que están presentes en todas las conversaciones de la red de interacciones o sólo en algunos grupos concretos de usuarios. Dependiendo de la comunidad, puede ser que surjan unos temas u otros, ya que la información no tiene por qué fluir de manera homogénea en la red. Este análisis nos puede ayudar a detectar por dónde fluye la información y qué nodos (usuarios) pueden ser claves a la hora de su propagación.

Analizar la relación de *asortatividad en base al sentimiento* calcula la relación que tienen los nodos de mayor orden junto con los de menor. Si aplicamos dicho algoritmo por un atributo como es el sentimiento, podemos observar si existe relación entre los tuits positivos y los negativos.

CAPÍTULO

4

Caso de estudio

4.1. Introducción

El tema a analizar en este trabajo es la política de Estados Unidos de América, por la gran actividad y presencia en las redes sociales. Para poder dar un uso de los datos, se tiene que acotar su alcance. Los tuits analizados son tanto del partido político que actualmente está en el poder (partido republicano) como el de la oposición (partido demócrata).

Para la implementación de la propuesta de análisis de información procedente de redes sociales se ha utilizado el lenguaje de programación Python. Python y R son los lenguajes pioneros para la ciencia de los datos, ya que gracias a las diferentes librerías disponibles facilitan la no implementación desde cero de los algoritmos y utilidades. La elección entre el lenguaje Python frente el lenguaje de R, fue que Python dispone de una fácil comunicación con elementos externos, como es la API de **Twitter** o la creación de aplicaciones web cuya implementación es mucho más sencilla.

Para este estudio se han extraído diez mil tuits tanto del partido político republicano como del demócrata, sumando veinte mil en total.

En la tabla 4.1 se indican las cuentas de usuario que se utilizaron para realizar la consulta a través del API de Twitter que nos permitió recoger la información de los tuits.

Demócratas	Republicanos
@TheDemocrats	@GOP
@HouseDemocrats	@POTUS
@DNC	@housegop
@HillaryClinton	@SenateGOP
@BarackObama	

Cuadro 4.1: Cuentas de usuario (asociadas a los partidos políticos) que se utilizaron para recuperar la información de Twitter.

Las fechas en las que se han recogido los tuits han sido entre el 12 y el 13 de diciembre del 2018. En esos días hubo una polémica en una de las entrevistas realizada por "Fox News" a Donald Trump, tal y como se puede ver en la Figura 4.1.

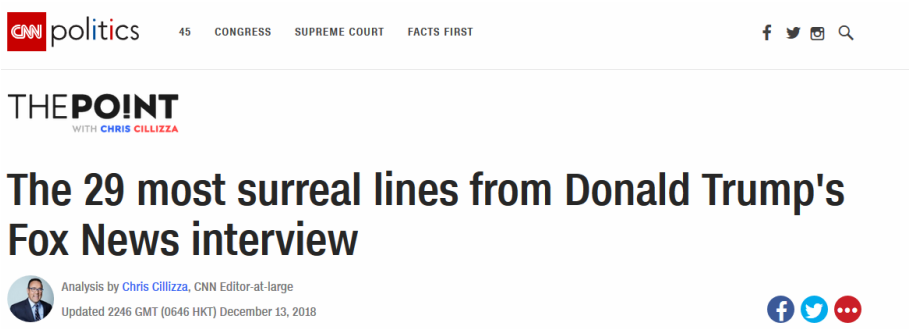

Figura 4.1: Noticias de CNN Politics (13/12/2018)

4.2. Análisis del contenido de los mensajes

En el caso de estudio, concretamente en la sección de análisis de contenido, se muestra la aplicación del algoritmo LDA y con el resultado de éste se ha realizado una búsqueda de noticias, comprobando que el resultado se corresponde con algún hecho relevante.

El primer paso ha sido aplicar el algoritmo LDA sobre los tuits de cada uno de los partidos políticos. En la Figura 4.2 se pueden observar el número de topicos y la relación entre ellos para el partido demócrata.

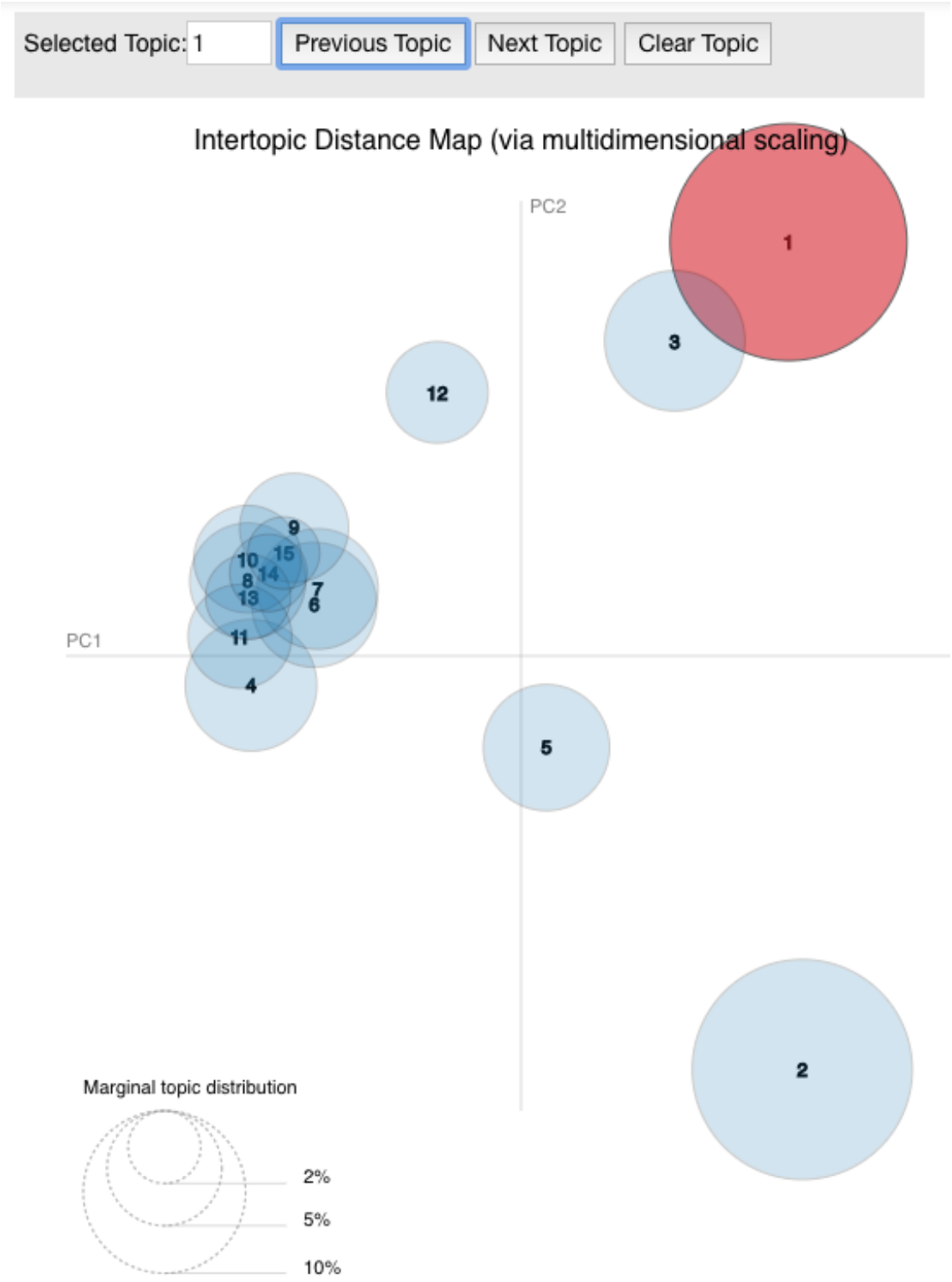

Figura 4.2: Número de topicos y su relación en el partido demócrata

Las palabras más destacadas en el partido político demócrata nos aparecen ordenadas desde las más repetidas a las que menos en la Figura 4.3.

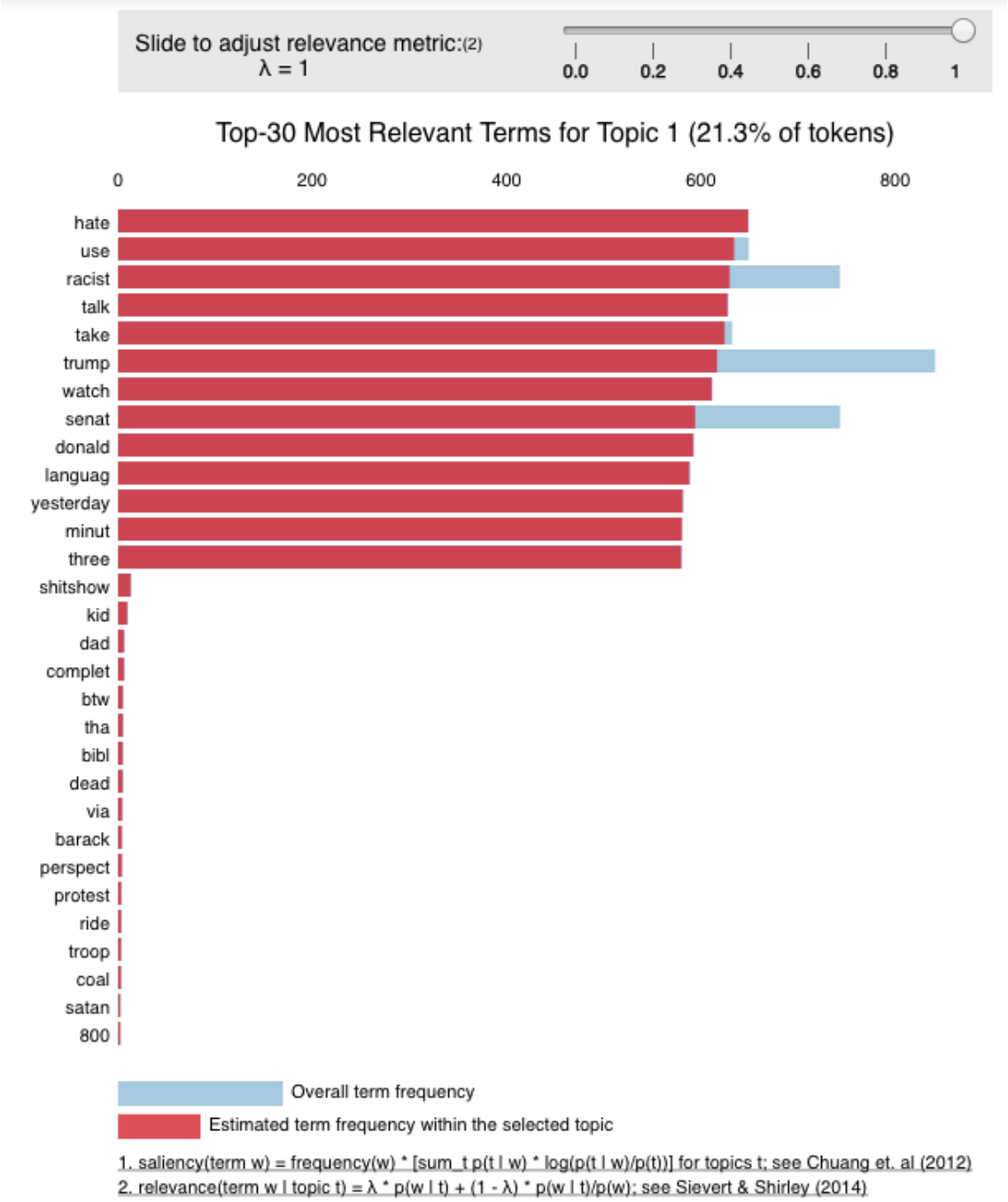

Figura 4.3: Palabras más destacadas en el partido demócrata

En la Figura 4.4 se pueden observar los 15 topicos generados y sus solapamientos en una escala multidimensional, también se puede observar la relación entre ellos.

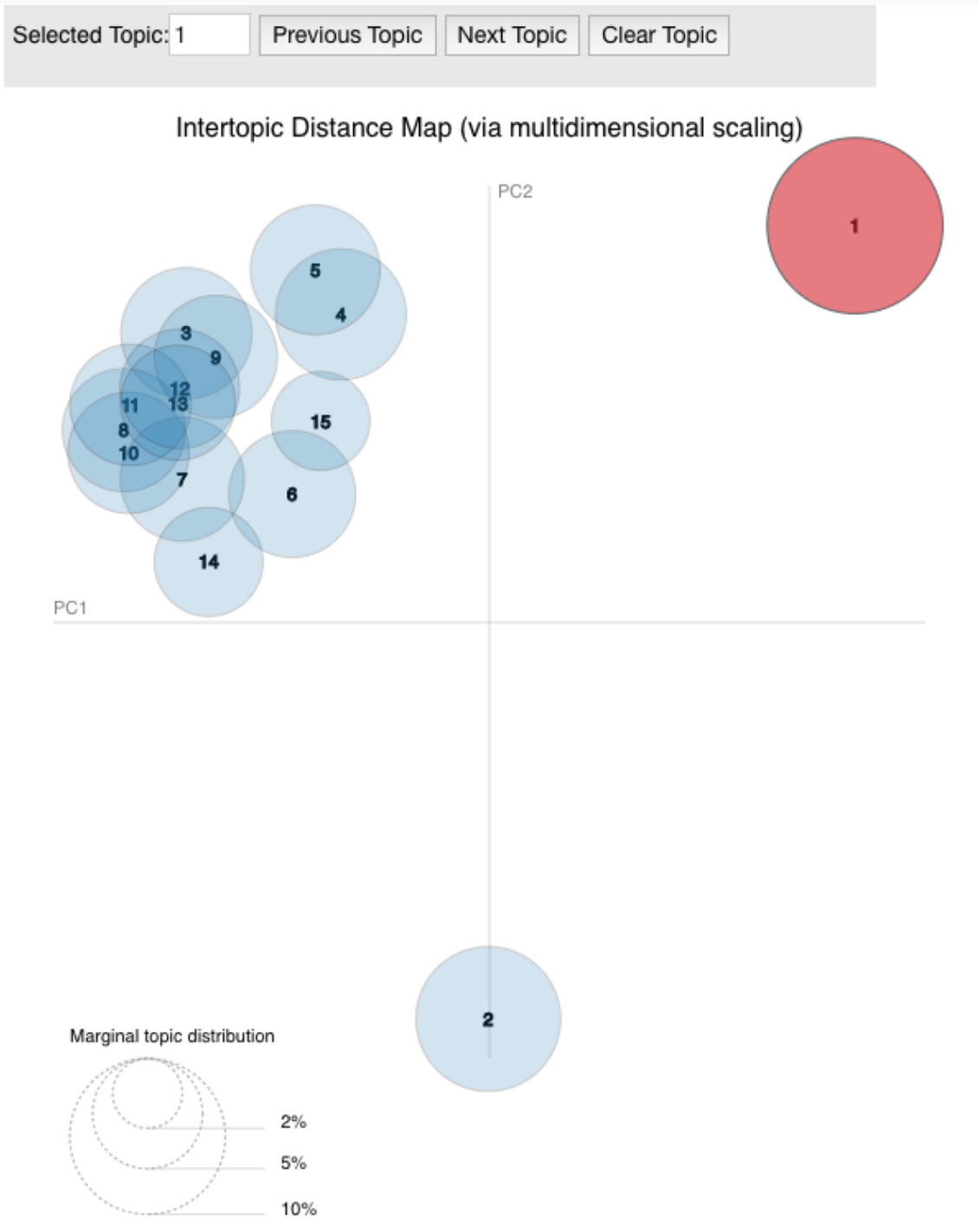

Figura 4.4: Número de topicos y su relación en el partido republicano

La Figura 4.5 se pueden ver las 30 palabras más relevantes por su frecuencia de aparición, también comentar que disponemos de 30 palabras diferentes para cada uno de los 15 tópicos.

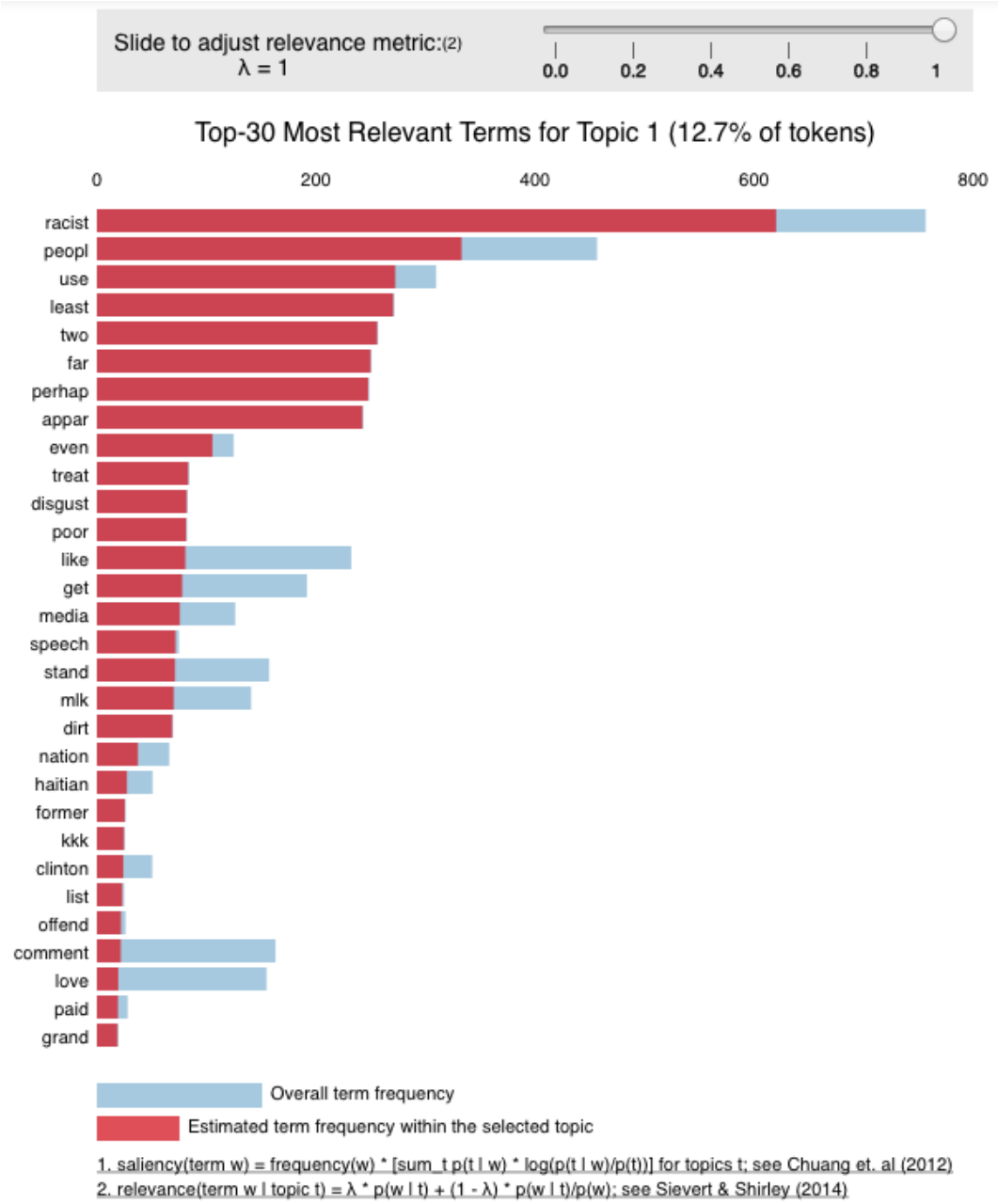

Figura 4.5: Palabras más destacadas en el partido republicano

Hemos escogido el topico número 1 clasificado por LDA y se ha realizado la búsqueda por la

palabra russianïncluyendo el nombre del partido político y en la Figura 4.6 se pueden observar que realmente ha habido noticias relacionadas con dicha palabra.

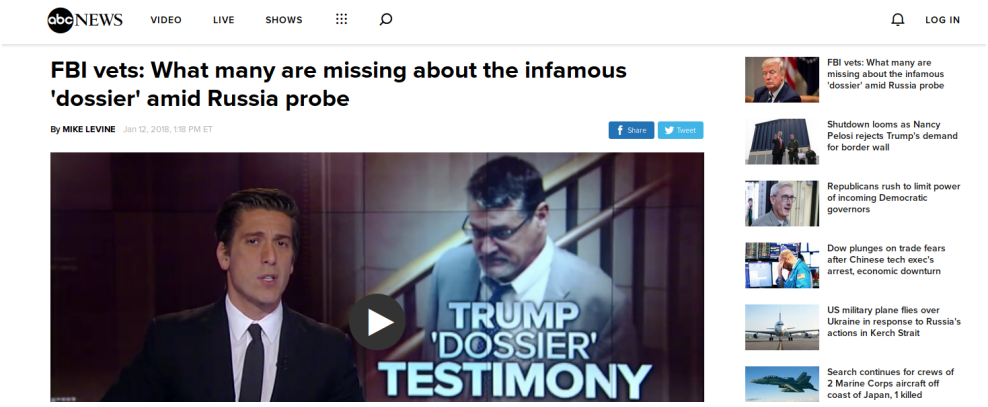

Figura 4.6: Noticia relacionada con Rusia (ABC News)

Por otra parte en el topico 2 se ha buscado noticias con la palabra ”haiti”, para ver si realmente ha habido polémica con dicha palabra o alguna noticia sobre esta.

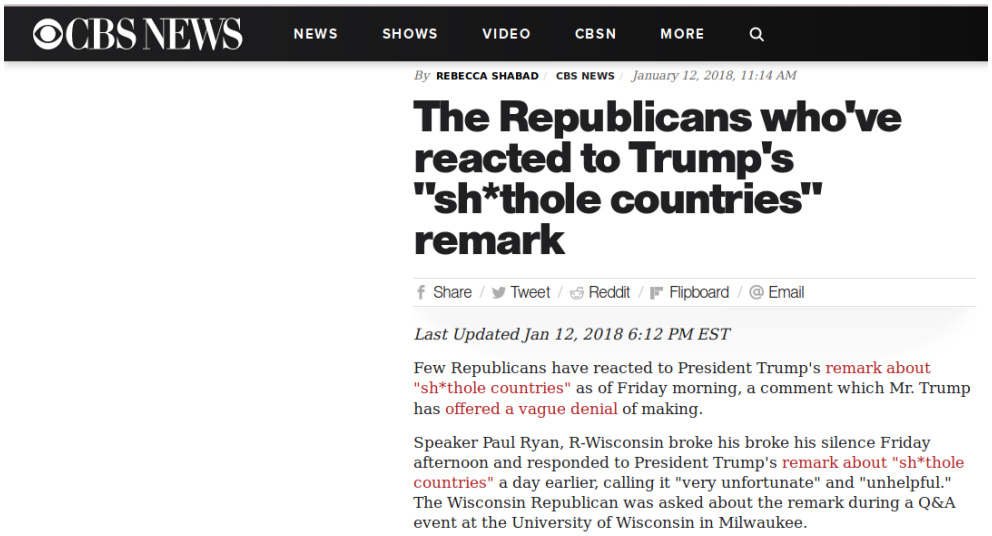

Figura 4.7: Trump declara que no insultó a Haití (CBS News)

De la misma manera se han seleccionado un par de topicos y sus palabras asociadas a partir de de los tuits analizados del partido demócrata. Una de las palabras más frecuentes fue ”Durbin”, uno de los senadores del partido demócrata.

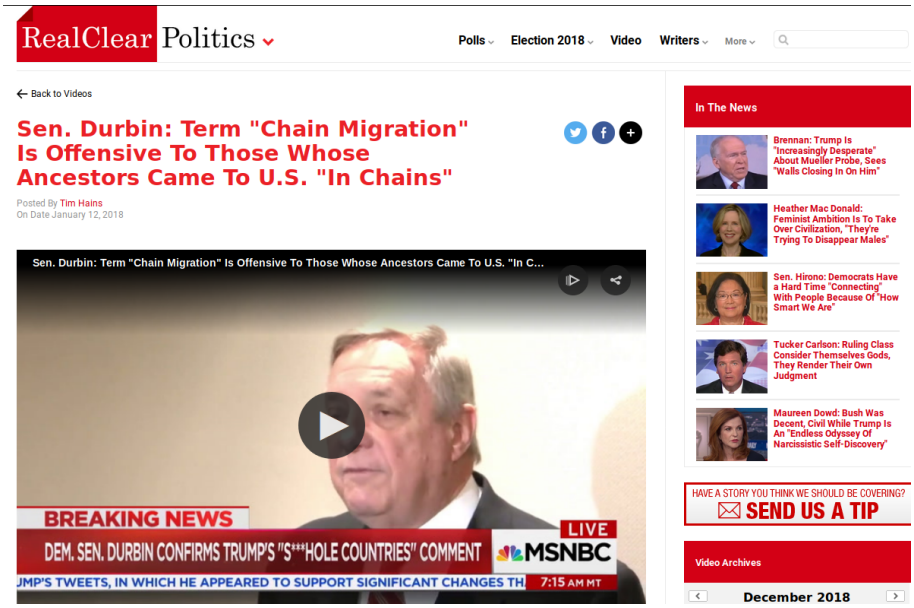

Figura 4.8: Noticia sobre las migraciones relacionada con el Senador Durbin

Dadas las noticias anteriores, se ha planteado la siguiente pregunta: ¿Cuál de los dos partidos políticos refleja un análisis de sentimiento positivo? Como era de esperar, en el partido demócrata correspondiente a la Figura 4.9 se observa que el sentimiento medio es mucho menor, esto se debe a los comentarios negativos que se repiten frente a la actual presidencia.

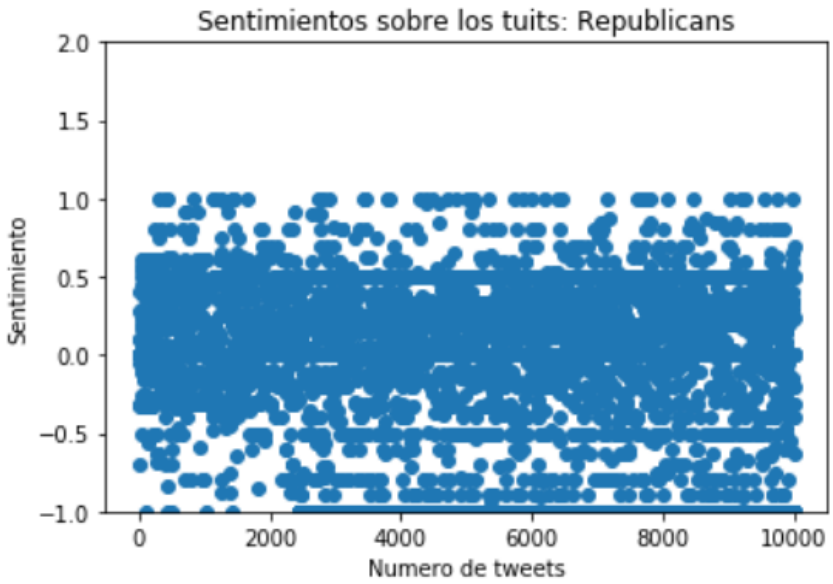

Figura 4.9: Tuits clasificados por sentimiento sobre el partido demócrata

En la Figura 4.10 y la Figura 4.9 se pueden comparar las dos gráficas, de las cuales se ha realizado un análisis de sentimientos PANA configurado en una escala de uno y menos uno, siendo uno sentimiento positivo y menos uno negativo.

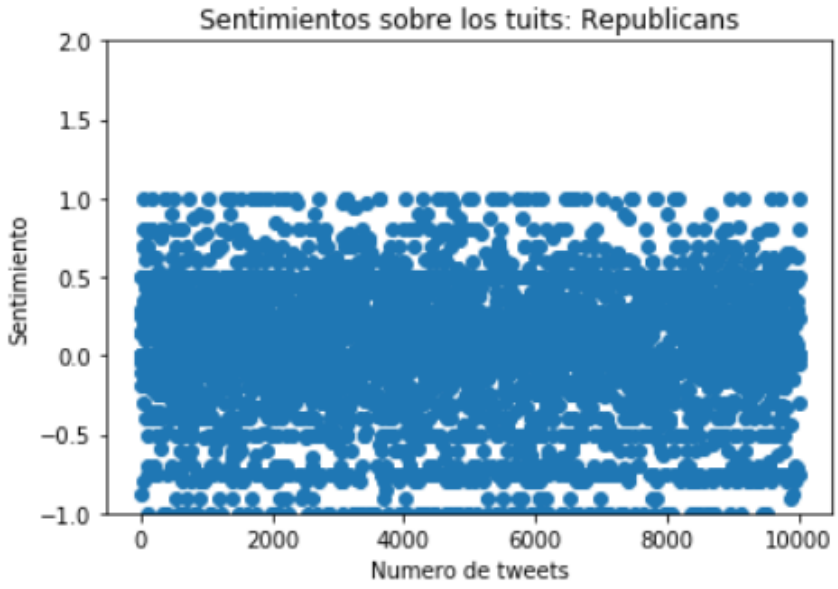

Figura 4.10: Tuits clasificados por sentimiento sobre el partido republicano

4.3. Análisis de la estructura comunicativa

En este apartado se muestran los resultados después de aplicar la propuesta planteada correspondiente al capítulo 3 de este documento. A continuación se describen los datos obtenidos tanto a nivel global como a nivel local o de nodo para el grafo que ha sido generado a partir de los tuits publicados por ambos partidos y también para los grafos de cada uno de los partidos, es decir, para los grafos que se han generado teniendo en cuenta sólo los tuits procedentes de cuentas de usuario relacionadas con ellos.

4.3.1. Análisis a nivel global

En las medidas estructurales de los grafos, podemos observar las diferencias de cada uno de los partidos, véase el Cuadro 4.2 y que en el mismo número de tuits existen más nodos en el partido demócrata, por lo que esto supone que hay más usuarios en esta red comparada con los republicanos. Cada nodo es un usuario que, o bien ha publicado un tuit, o bien ha sido

mencionado por otro. Al mismo tiempo, esto supone más interacciones ente ellos, por lo que el número de enlaces es mayor.

	Demócratas	**Republicanos**
Nodos	5280	6334
Enlaces	10718	13449

Cuadro 4.2: Comparación de nodos y enlaces de los grafos asociados a cada partido

Para ver claramente el análisis entre los partidos políticos, se ha propuesto representar los datos mediante grafos, donde los nodos representarán a los perfiles que han publicado algún tuit y/o han sido mencionados por otros y las conexiones a otros nodos (perfiles) representan relaciones establecidas por medio de referencias o menciones.

En la Figura 4.11 se puede observar el primer grafo obtenido donde se han considerado los tuits de los republicanos y demócratas juntos. En este grafo, el tamaño de los nodos viene determinado por su centralidad basada en intermediación y los colores representan comunidades.

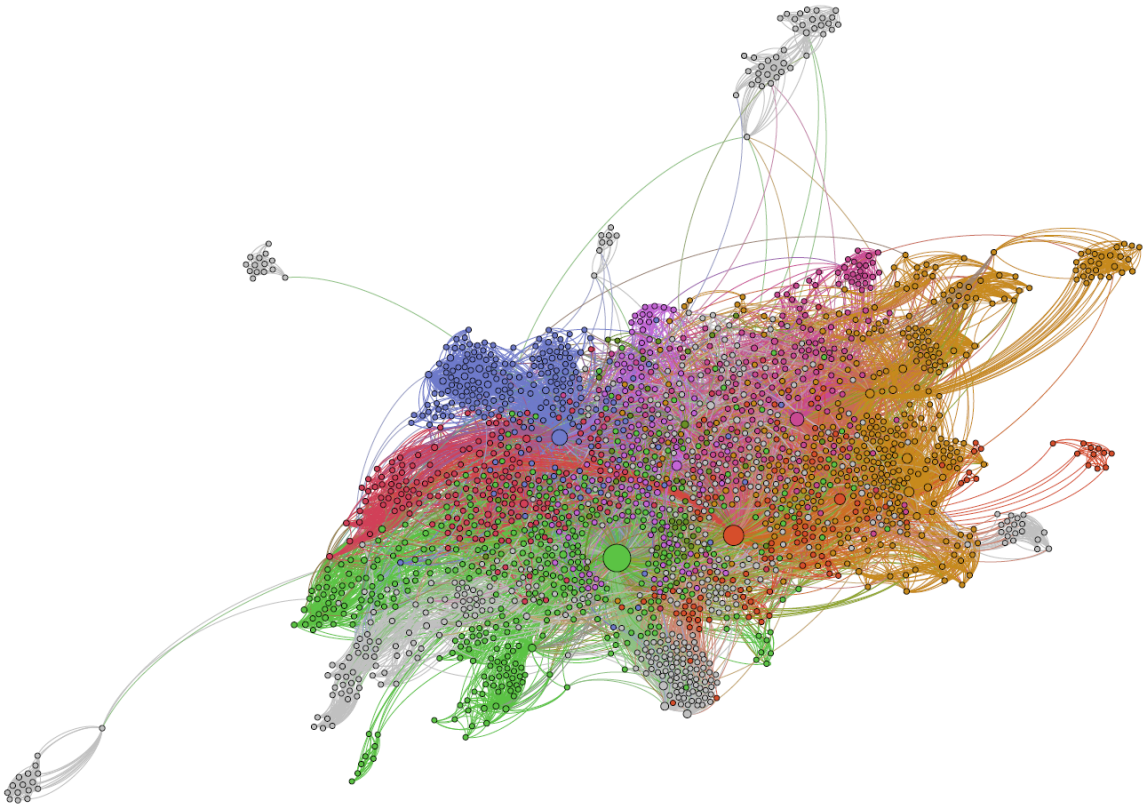

Figura 4.11: Tuits de los partidos republicanos y demócratas representados mediante un grafo

Modularidad. Se ha aplicado esta característica en el entorno de análisis por comunidades, si hay más de una comunidad en un supuesto único ámbito (el político).

	Demócratas	**Republicanos**
Numero de comunidades dentro del grafo	109	189

Cuadro 4.3: Modularidad en el grafo de los demócratas y en el de los republicanos

Vistos los resultados en el Cuadro 4.3, hay muchas comunidades diferentes dentro del grafo político, con comunidades centradas que pertenecen a un mismo tema y también mixtas donde intervienen nodos de diferentes comunidades y no hay una zona clara de cuál es la que prevalece.

Enlaces. Las interacciones de los usuarios son representados mediante los enlaces en el grafo. Para ver cómo fluye la información, se ha aplicado el análisis de sentimientos sobre los enlaces.

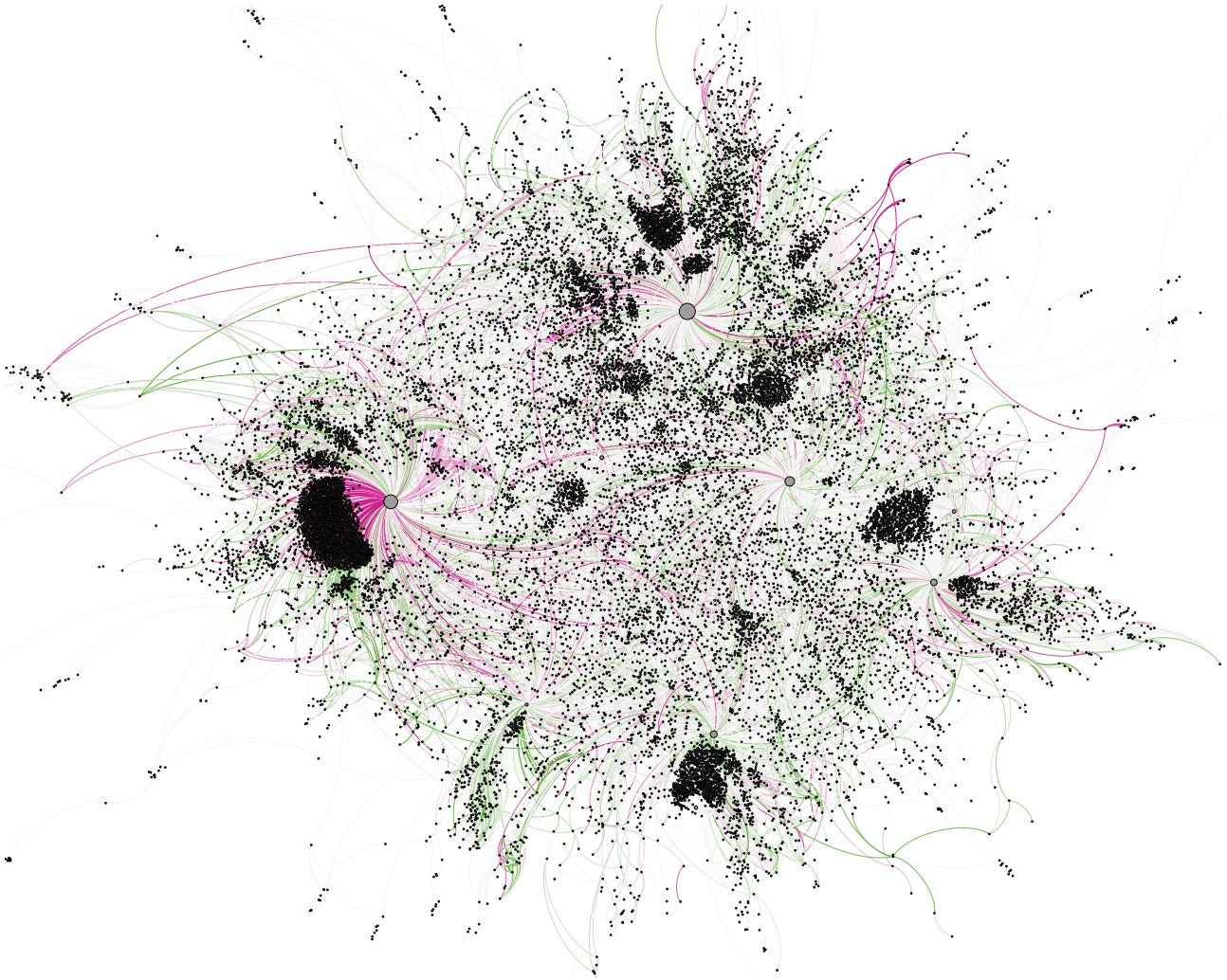

Figura 4.12: Grafo generado a partir de los tuits de ambos partidos. El color de los enlaces depende del sentimiento del tuit que haya generado ese enlace. El color verde indica un sentimiento positivo y el color rojo indica un sentimiento negativo. El tamaño de los nodos viene determinado por el grado.

Se puede observar en el grafo generado a partir de los tuits de ambos partidos (ver Figura 4.12) que tuits negativos mencionan perfiles de un grado más alto. Concretamente, en la parte izquierda del grafo, el nodo gris con los enlaces rosa conectados nos indica que el perfil ha sido mencionado en varios tuits negativos, o el perfil ha publicado tuits negativos.

Calculamos también la densidad del grafo generado para cada partido político. En el Cuadro 4.4 se puede observar que los valores son próximos a 0, eso es indicativo de que los dos grafos de los partidos políticos apenas tienen enlaces entre los nodos (existe poca interacción entre las cuentas de los usuarios), ya que un valor próximo al 1 seria un grafo completo.

	Demócratas	Republicanos
Densidad	0.00076	0.00066

Cuadro 4.4: Comparación de la densidad de los grafos de los partidos demócrata y republicano

Asortatividad. Para comprender los datos anteriores y si son relevantes o no, debemos calcular la asortatividad, es decir, averiguar si los nodos tienen preferencia por la unión con otros. Para ello es necesario especificar porqué atributo queremos realizar el análisis. En este ejemplo se ha elegido el grado.

Al hacer esto sobre el grado de un grafo no dirigido obtenemos un calculo global de los datos. En las Figuras 4.13 y 4.14 podemos ver que los dos partidos políticos son muy parecidos en este análisis.

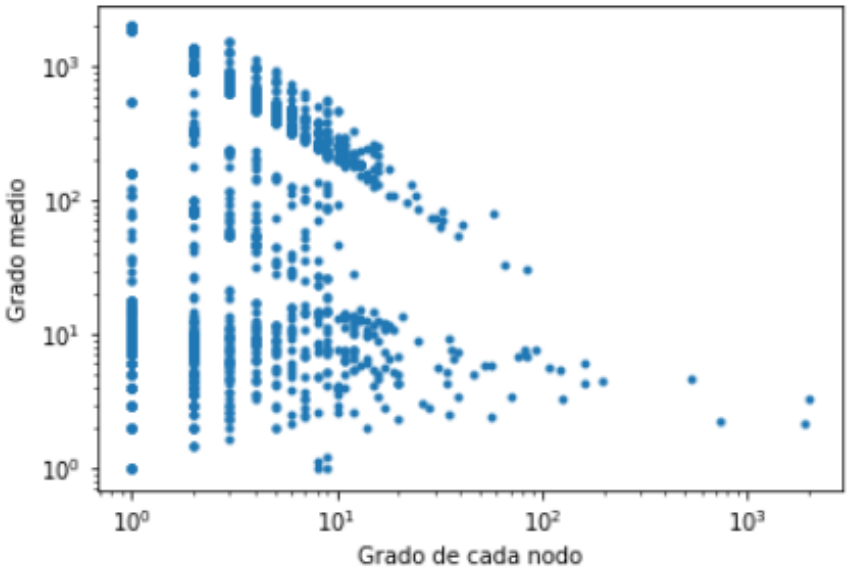

Figura 4.13: Gráfica de asortatividad en el partido demócrata

Los resultados son bastante similares en la Figura 4.14 junto con la anterior, por lo que los dos partidos políticos tienen parecidas conexiones.

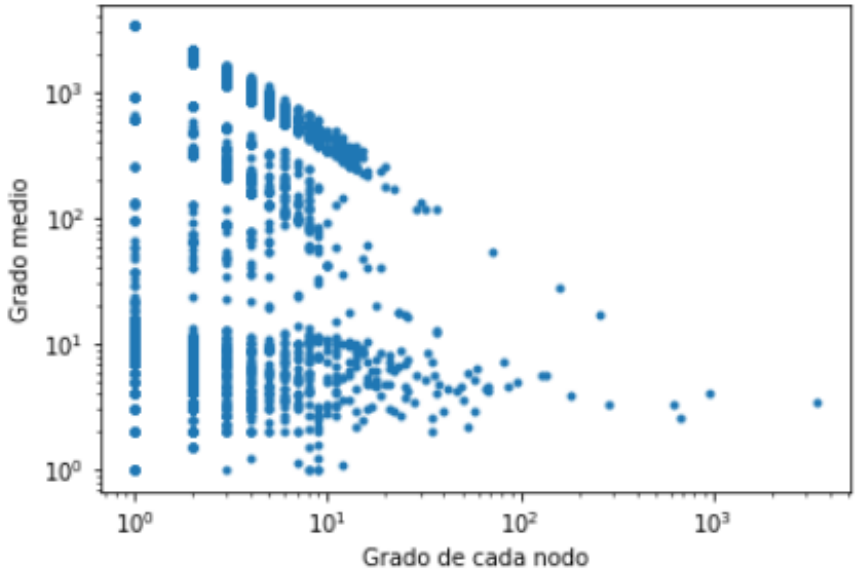

Figura 4.14: Gráfica de asortatividad en el partido republicano

En el Cuadro 4.5 se puede observar que al ser los valores negativos nos indica que en los dos partidos políticos cuyos nodos con grado alto están conectados con los de grado bajo. Esto es que existen relaciones entre usuarios con características diferentes, perfiles que tienen un alto grado de interacción (bien porque generan mensajes o bien porque reciben menciones) se comunican con otros que tiene poco grado de interacción.

	Demócratas	**Republicanos**
Asortatividad	- 0.297	- 0.206

Cuadro 4.5: Asortatividad del grafo

4.3.2. Análisis a nivel de nodo

En los resultados anteriores, se ha realizado un análisis de forma global. A continuación realizaremos el análisis a nivel local para detectar qué usuarios son relevantes en cada grafo atendiendo a distintos criterios. En este análisis se ha considerado el grafo dirigido para que todos los nodos tengan sus enlaces de entrada y salida.

En la Figura 4.15 se puede ver la gráfica de la centralidad basada en el grado sobre los enlaces de entrada en los nodos.

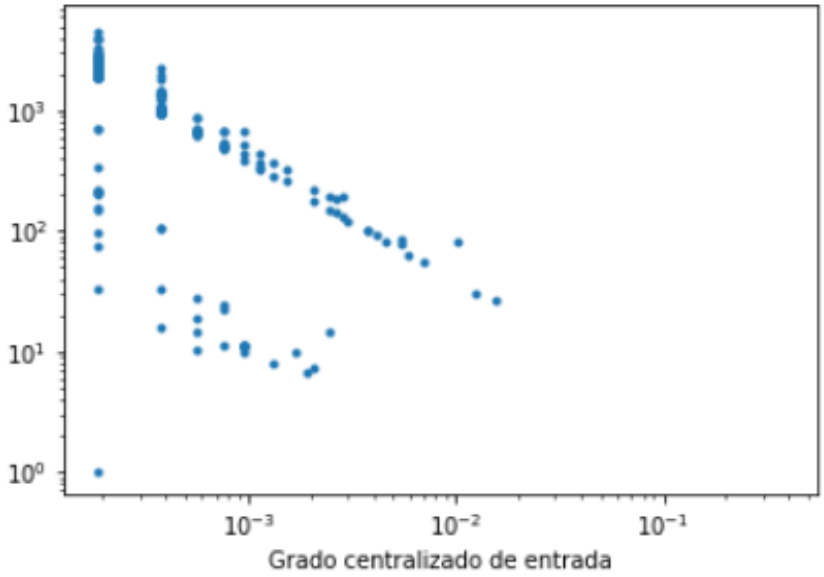

Figura 4.15: Gráfica del grado de centralidad de entrada en el partido demócrata

En la Figura 4.16 se observa que los nodos tienen más enlaces de entrada que la Figura 4.15, por lo que podemos decir que en el partido político republicano hay mucha más actividad que en el partido demócrata.

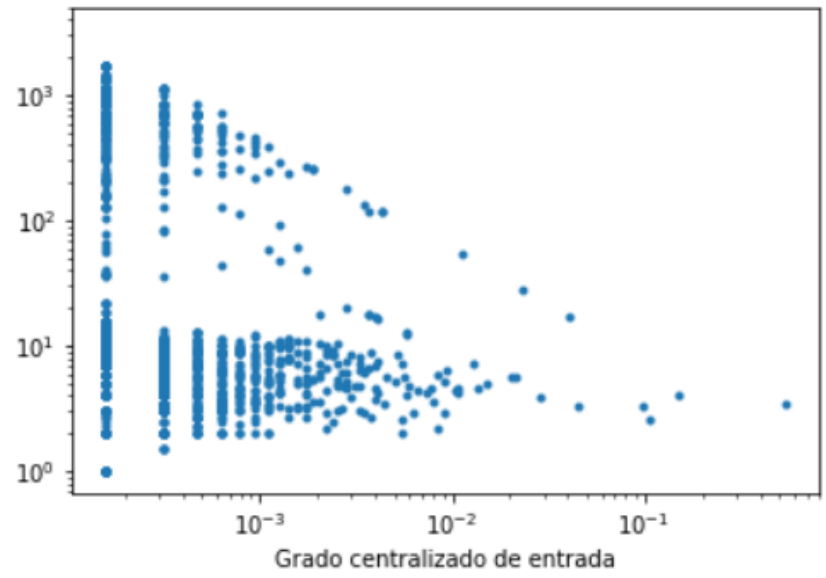

Figura 4.16: Gráfica del grado de centralidad de entrada en el partido republicano

A continuación realizamos el mismo análisis, pero esta vez con los enlaces de salida de cada nodo.

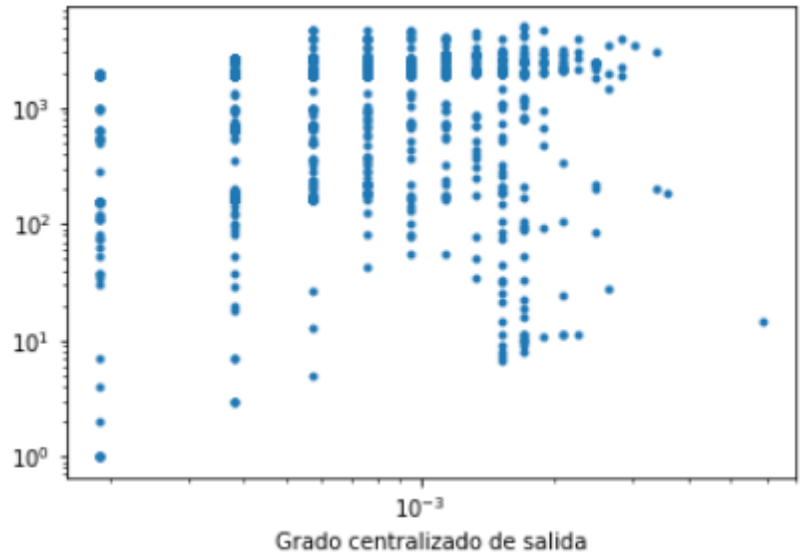

Figura 4.17: Gráfica del grado de centralidad de salida en el partido demócrata

Al comparar la Figura 4.17 con la siguiente 4.18 se puede observar que en esta segunda hay una ligera diferencia que se explica a continuación.

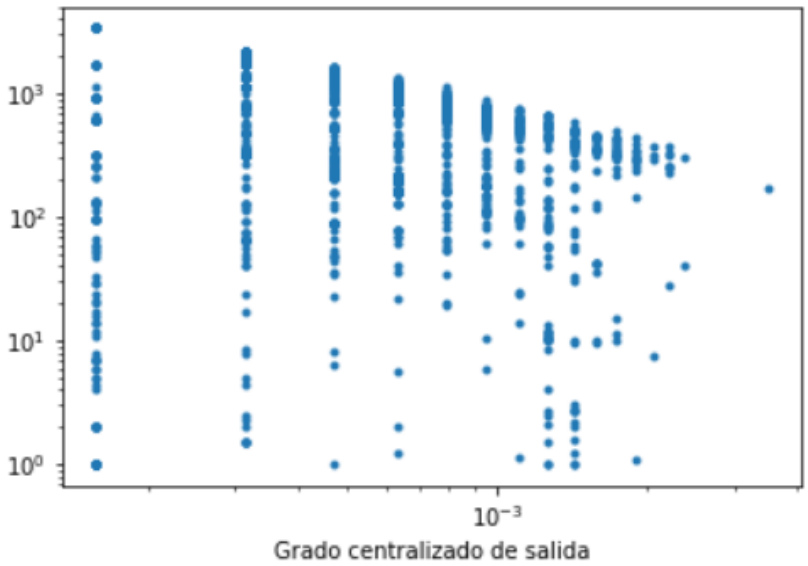

Figura 4.18: Gráfica del grado de centralidad de salida en el partido republicano

En la Figura 4.17 se mantienen los enlaces de salida, esto quiere decir que o bien afecta o bien son relevantes los tuits siempre para los mismos usuarios. Sin embargo, en la Figura 4.18 tiende a disminuir, esto quiere decir que tienen menos enlaces de salida, sea porque tuitean menos o porque no hacen referencia a ninguna persona.

Antes de realizar cualquier análisis, comentar que se han aplicado los filtros descritos en el capítulo 3 apartado de medidas a nivel global que son los siguientes, "K-core", "Componente gigante" y el "Atributo no-nulo" explicado a continuación.

Atributo no-nulo: Se trata de un filtro que no cuenta los datos nulos en el atributo que especificaremos. Esto es importante a la hora de analizar el sentimiento de un nodo. Existen nodos que no han publicado ningún tuit, por lo que no se puede calcular el sentimiento de éste. Estos nodos son creados porque han sido mencionados por otros usuarios.

En esta primer Cuadro 4.6 podemos observar la centralidad basada en intermediación del partido demócrata.

	Centralidad basada en intermediación	**Tipo de perfil**
usvetram	463.0	periodista
mikefarb1	224.0	periodista
_ROB_111	154.0	informático
SmallBiz4Trump	140.0	perfil de difusión de contenido
thomaskaine5	122.0	empresario y consultor político

Cuadro 4.6: Centralidad basada en la intermediación para el partido demócrata

El siguiente Cuadro 4.7 muestra el resultado de aplicar centralidad basada en intermediación. Sigue el orden de mayor a menor, comentar que se han seleccionado las cinco cuentas de usuario con mayores valores.

	Centralidad basada en intermediación	**Tipo de perfil**
President1Trump	1991.0	perfil de difusión de contenido
Jim_Peoples_	482.0	ingeniero aeroespacial
FoxNews	287.5	medio de comunicación
MAGARoseTaylor	243.5	editora (NewRightNetwork)
RNRKentucky	146.5	perfil de difusión de contenido

Cuadro 4.7: Centralidad basada en la intermediación para el partido republicano

Las siguientes tablas analizan la centralidad basada en la cercanía se muestra que los dos partidos tienen el valor máximo uno, es decir, están conectados a otros y lo interesante es los

diferentes perfiles de cada nodo.

	Centralidad basada en la cercanía	Tipo de perfil
mikefarb1	1.0	periodista
_ROB_111	1.0	informático
thomaskaine5	1.0	empresario y consultor político
Westxgal	1.0	perfil de difusión de contenido
Calypsia6978	1.0	usuario de Twitter

Cuadro 4.8: Centralidad basada en la cercanía para el partido demócrata

	Centralidad basada en la cercanía	Tipo de perfil
President1Trump	1.0	perfil de difusión de contenido
FoxNews	1.0	medio de comunicación
MAGARoseTaylor	1.0	editora (NewRightNetwork)
SusanBe35980777	1.0	cuenta suspendida
kupajo333	1.0	cuenta suspendida

Cuadro 4.9: Centralidad basada en la cercanía para el el partido republicano

A continuación se muestran dos cuadros en las que se ha calculado la centralidad basada en el vector de valores propios.

	Centralidad basada en la cercanía	Tipo de perfil
TheDemocrats	1.0	perfil del partido político
HillaryClinton	0.9430004754053958	senadora de “DOP”
DickDurbin	0.3693402116078976	político estadounidense en “DOP”
realDonaldTrump	0.27084216507861825	actual presidente de estado unidos
FoxNews	0.09761727749436049	medio de comunicación

Cuadro 4.10: Centralidad basada en el vector de valores propios para el partido demócrata

	Centralidad basada en la cercanía	Tipo de perfil
POTUS	1.0	perfil de difusión de contenido oficial
realDonaldTrump	0.2758878520729157	actual presidente de estado unidos
Scaramucci	0.1960672819786087	consultor político
GOP	0.1840568218569463	perfil de difusión de contenido del partido político
Resist0917	0.08412876583910478	cuenta suspendida

Cuadro 4.11: Centralidad basada en el vector de valores propios para el partido republicano

Estos resultados del Cuadro 4.10 nos indican de que nodos son importantes, es decir, que otros perfiles interactúan con nodos que a su vez son importantes.

Mediante la subjetividad y la polaridad se ha querido analizar la estructura de cada partido político. Para ello se han extraído las variables (subjetividad y polaridad) de cada tuit, para ver el resultado en una misma gráfica de cada partido, se han invertido los valores de subjetividad para el partido político republicano, así se puede ver si hay simetría.

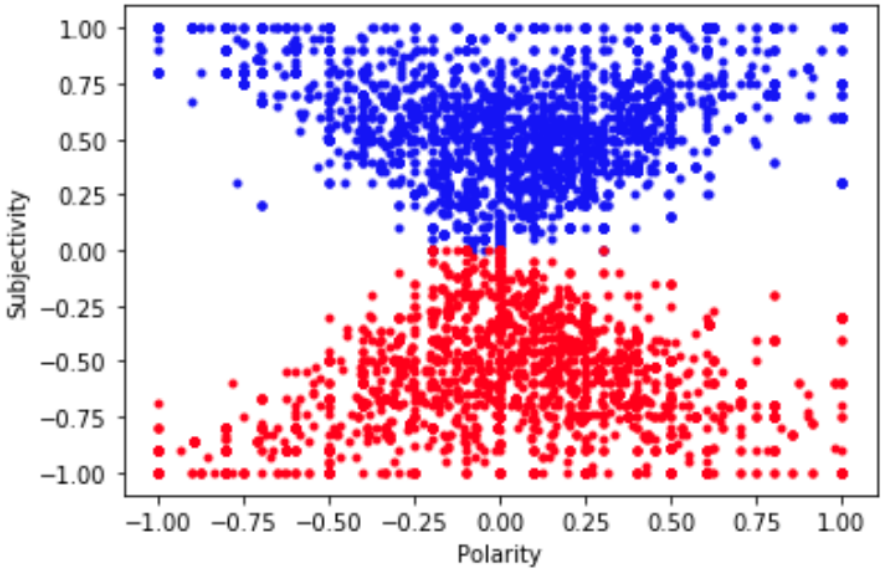

Figura 4.19: Subjetividad y polaridad de los partidos políticos

Sin embargo, de esta forma se pierde información, ya que los tuits unos con otros se sobreponen los nodos con el mismos valor en subjetividad y polaridad. Por ello se planteó la alternativa de representar los valores mediante un mapa de calor que se muestra en la Figura 4.20

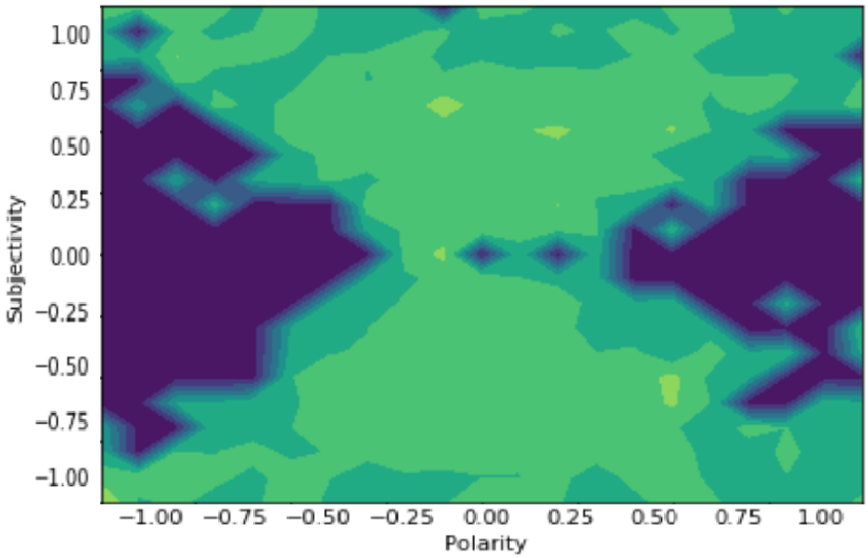

Figura 4.20: Mapa de calor de republicanos y demócratas

Al analizar el mapa de calor se puede observar varios puntos más claros por ejemplo en las posiciones (0.00, 0.00), también en (0.00, 0.75) o (0.50, -0.60).

Gracias a las Figuras 4.20 y 4.21 se observa que las estructuras son bastante diferentes, ya que mediante la undidad destacan el número de tuits con el mismo patrón (subjetividad y polaridad), cosa que no es posible visualizar a simple vista en la Figura 4.19.

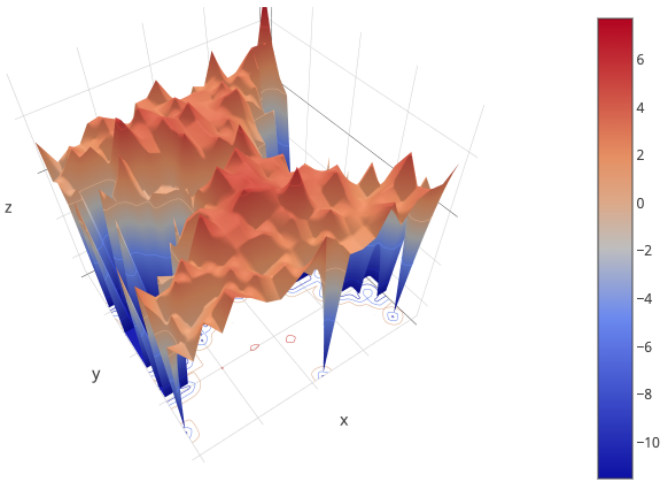

Figura 4.21: Visualización en 3D de republicanos y demócratas

4.4. Análisis híbrido

En este análisis, se va a considerar el grafo generado a partir de todos los tuits de ambos partidos. En este análisis el primer paso es aplicar los filtros comentados anteriormente “Componente gigante”, “K-core” con un valor de k = 2, valor no nulo en atributo del sentimiento de tuit y aplicando la modularidad que nos permite detectar comunidades.

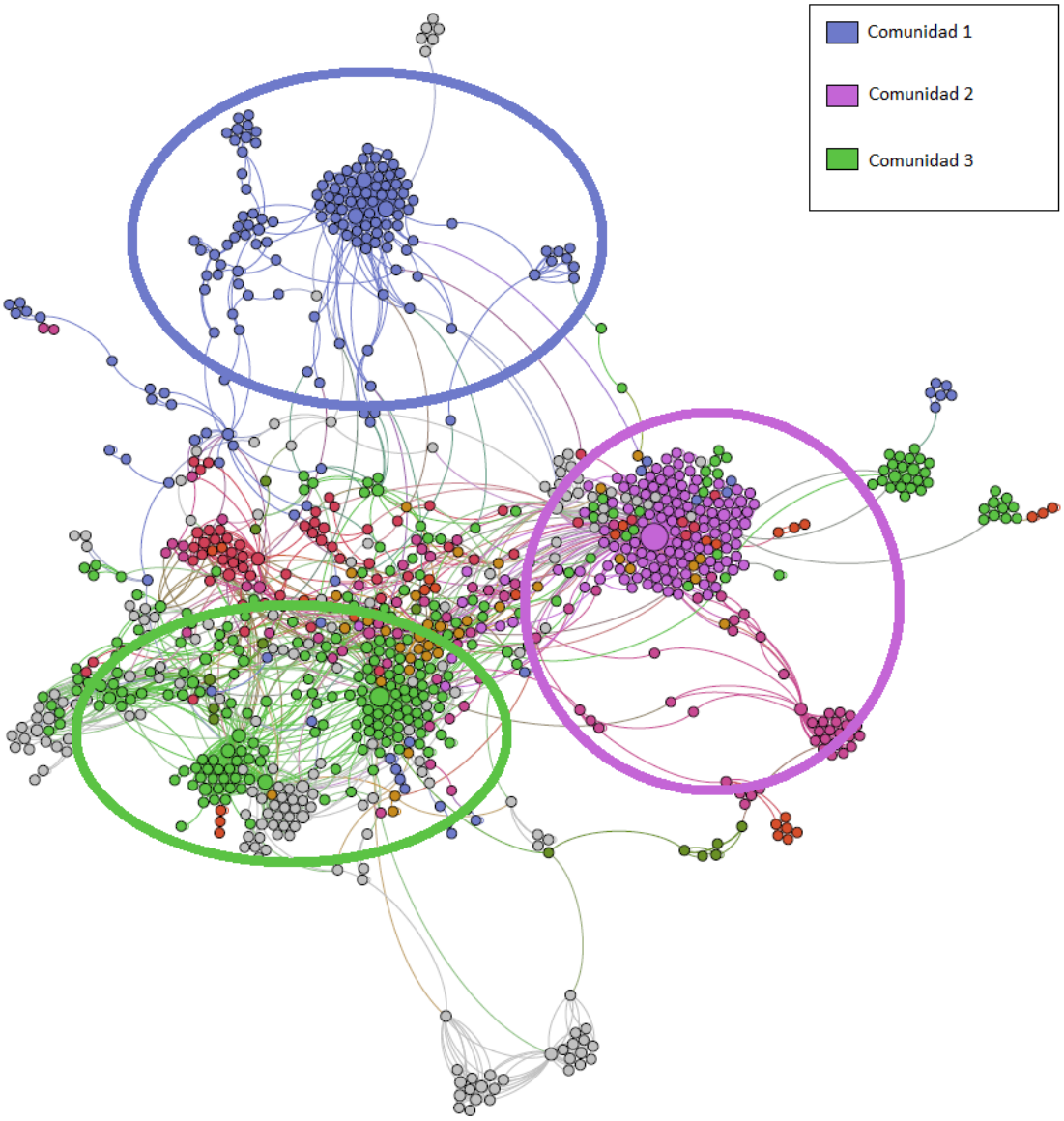

Figura 4.22: Grafo a partir de tuits de los demócratas y republicanos con filtros aplicados (componente gigante, k-core, atributo no nulo)

En la Figura 4.22 se puede observar que destacan tres comunidades claras que contienen nodos con un grado alto. En la comunidad 1 destaca el usuario “grandoftwo”, un perfil a favor del partido político demócrata. En las comunidad 2 destaca un medio de comunicación, concretamente el perfil de “FoxNews”. Por último, en la comunidad 3 destaca el nodo con el perfil de “president1Trump”, un perfil de difusión de mensajes declarado como republicano.

Se ha aplicado filtro por partido para ver qué nodos pertenecen a qué partidos políticos y si realmente se cumple que hay 3 comunidades diferentes que hablen de temas en común.

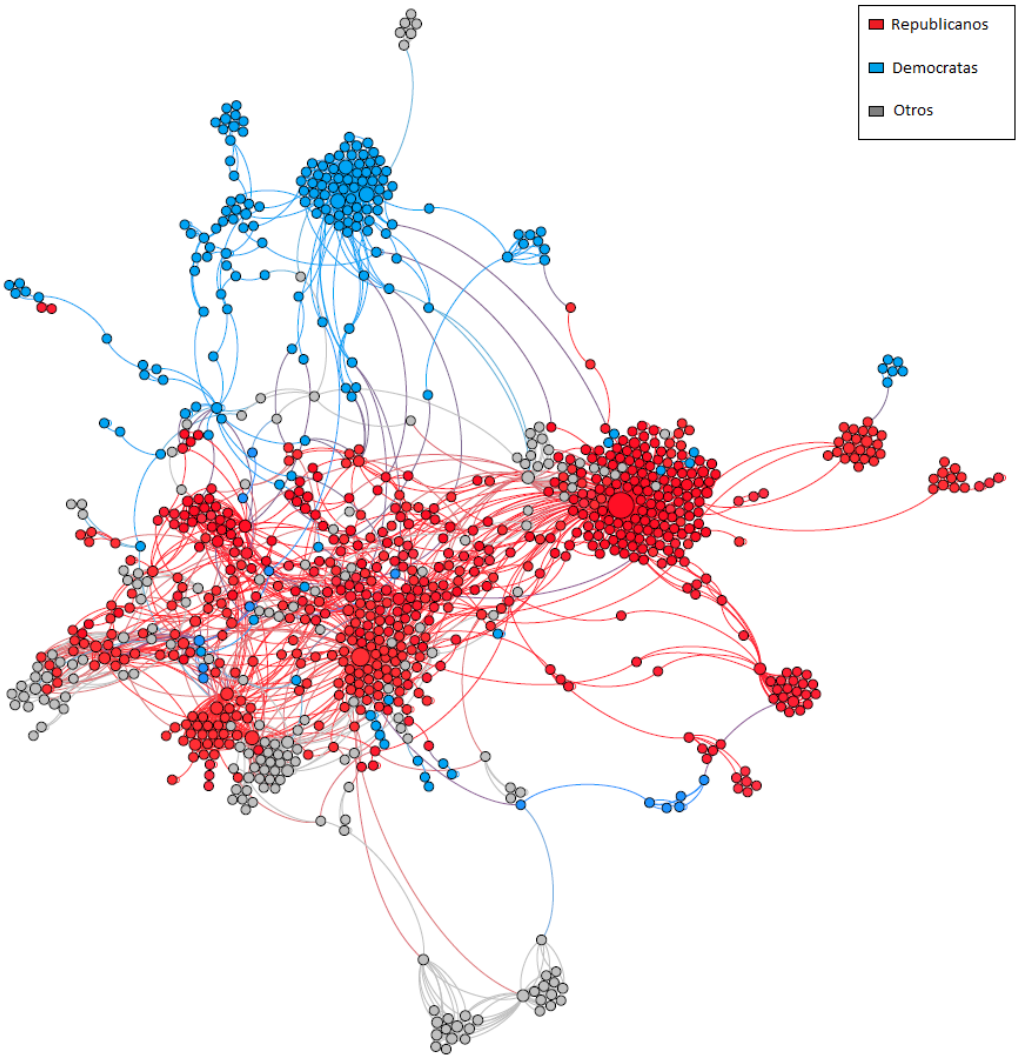

Figura 4.23: Grafo a partir de tuits de los demócratas y republicanos con filtros aplicados (componente gigante, k-core, atributo no nulo) y distinción política por colores

Sin embargo vista a los resultados en la Figura 4.23 podemos ver que la comunidad 2 con

la comunidad 3 no se distingue, por lo que puede esconder información relevante sobre la orientación política de los perfiles.

4.4.1. Sentimiento basado en comunidades

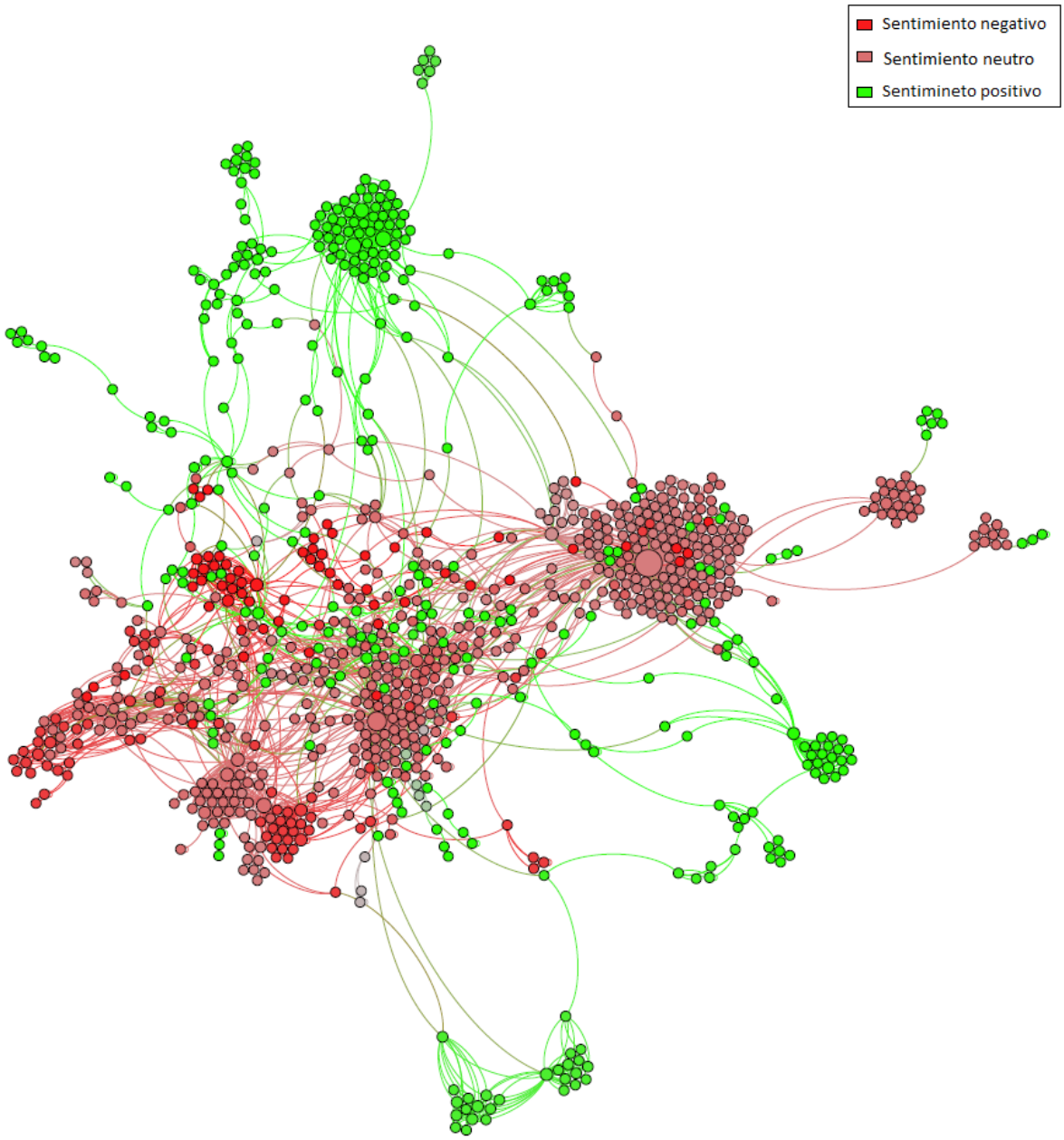

Figura 4.24: Grafo con filtros y el análisis de sentimientos

Después de aplicar una clasificación de los nodos por el sentimiento vemos que las comunidades anteriores que teníamos solo una de ellas concuerda con sentimiento positivo mientras que la comunidad 2 y 3 son nodos la mayoría neutros o negativos.

4.4.2. Asortatividad basada en el sentimiento

La función de asortatividad es la medida en que los nodos de un grafo se vinculan con otros del mismo grafo en base a la similaridad basada en ciertos atributos. En este caso se han tenido en cuenta el grado y el sentimiento.

	Republicanos y Demócratas
Asortatividad basada en el sentimiento	0.0154

Cuadro 4.12: Asortatividad basada en el sentimiento sobre republicanos y demócratas

Visto el resultado de la tabla 4.12 significa que hay muy pocos nodos que estén vinculados con otros teniendo el mismo grado y además el mismo sentimiento.

CAPÍTULO

5

Conclusiones

En este documento se ha realizado un análisis de diferentes trabajos relacionados con la extracción y el tratamiento de datos en la red social de Twitter. Mediante la estructura de la Figura 3.1 se ha seguido el procedimiento de la extracción de los datos, limpieza de estos y análisis para la propuesta, que es el punto donde el proyecto toma diferentes caminos de análisis.

Esto implica realizar un análisis de contenido de los mensajes que incluye dos subtareas. Por una parte el procesado de LDA para la extracción de topicos y comprobación de si son verídicos, es decir, búsqueda de noticias relacionadas con estos en la fecha de los tuits. Por otra parte realizar un análisis de sentimientos, esto nos ha demostrado que concuerda con las noticias si son positivas o negativas.

Seguido de esto se ha realizado el análisis de la estructura comunicativa, con un primer análisis para ver el comportamiento de los tuits en el grafo a nivel global. Muchos de los procesos no se pueden aplicar al grafo entero globalmente y no se puede ver el comportamiento de los nodos de éste, por ello se ha realizado un proceso de análisis a nivel local o de centralidad que abarca los comportamientos de los nodos del grafo.

Para comprobar si realmente estamos tratando toda la información en este documento, se ha propuesto como objetivo clasificar el grafo por diferentes comunidades, es decir, diferentes subgrafos que estén relacionados. Este apartado corresponde al análisis híbrido. En este apartado se ha comprobado si realmente concuerda la información de las comunidades con el grafo general y si actúan igual que éste.

Al acabar el análisis se ha podido comprobar que cuando analizamos un grafo general se oculta información y si no se realiza un análisis por partes no podremos ni aprovechar esta ni demostrar que realmente es verídica (la del grafo global).

Se plantea como proyectos futuros:

1. Realizar un análisis de los usuarios que pueden ser posibles perfiles influyentes conocidos como “influencers”.
2. Detectar usuarios que son poco activos o con mensajes automatizados (bots).
3. Análisis de publicaciones de tuits durante el mandato y comparar cuando no estuviera gobernando un partido político.

Bibliografía

[1] AHN, Y.-Y., HAN, S., KWAK, H., MOON, S., AND JEONG, H. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web* (2007), ACM, pp. 835–844.

[2] ALVAREZ RAMOS, L. Análisis de ciudades a través su actividad en redes sociales.

[3] ARCILA-CALDERÓN, C., ORTEGA-MOHEDANO, F., JIMÉNEZ-AMORES, J., AND TRULLENQUE, S. Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *http://www.elprofesionaldelainformacion.com* (2016).

[4] BANN, E. Y. *Undergraduate Dissertation: Semantic Clustering of Basic Emotion Sets*. PhD thesis, Master's thesis, The University of Bath, 2012.

[5] BARBERÁ, P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis 23*, 1 (2015), 76–91.

[6] BASTIAENSENS, S., VANDEBOSCH, H., POELS, K., VAN CLEEMPUT, K., DESMET, A., AND DE BOURDEAUDHUIJ, I. Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior 31* (2014), 259–271.

[7] BORONDO, J., MORALES, A., LOSADA, J. C., AND BENITO, R. M. Characterizing and modeling an electoral campaign in the context of twitter: 2011 spanish presidential election as a case study. *Chaos: an interdisciplinary journal of nonlinear science 22*, 2 (2012), 023138.

[8] BRADLEY, M., GREENWALD, M., PETRY, M., AND LANG, P. J. Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 18* (1992).

[9] CARMEL, D., ZWERDLING, N., GUY, I., OFEK-KOIFMAN, S., HAR'EL, N., RONEN, I., UZIEL, E., YOGEV, S., AND CHERNOV, S. Personalized social search based on the user's social network. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 1227–1236.

[10] CHANG, J.AND BOYD-GRABER, J., AND M., B. D. Connections between the lines: augmenting social networks with text. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009).

[11] CONCHA PÉREZ CURIEL, M. G. G. Influence politics and "fake" trend on twitter. 21d post-election effects within the procés in catalonia. *http://www.elprofesionaldelainformacion.com/contenidos/2018/sep/07.pdf* (2018).

[12] EBNER, M., AND REINHARDT, W. Social networking in scientific conferences–twitter as tool for strengthen a scientific community. In *Proceedings of the 1st International Workshop on Science* (2009), vol. 2, pp. 1–8.

[13] GOLBECK, J., ROBLES, C., AND TURNER, K. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (2011), ACM, pp. 253–262.

[14] GORDON, K. Social media statistics and facts. *https://www.statista.com/topics/* (2017).

[15] HILLMANN, R., AND TRIER, M. Dissemination patterns and associated network effects of sentiments in social networks. . *In IEEE International Conference on Advances in Social Networks Analysis and Mining, ASONAM* (2012).

[16] HONG, L., AND DAVISON, B. D. Empirical study of topic modeling in twitter. *Proceedings of the first workshop on social media analytics ACM* (2010).

[17] HUBERMAN, B. A., ROMERO, D. M., AND WU, F. Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045* (2008).

[18] KANNANGARA, S. Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (2018), ACM, pp. 751–752.

[19] KEMP, S. Global digital report. *https://search.proquest.com* (2019).

[20] LARSSON, A. O., AND MOE, H. Studying political microblogging: Twitter users in the 2010 swedish election campaign. *New Media & Society 14*, 5 (2012), 729–747.

[21] LESKOVEC, J., ADAMIC, L. A., AND HUBERMAN, B. A. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB) 1*, 1 (2007), 5.

[22] LOTAN, G., GRAEFF, E., ANANNY, M., GAFFNEY, D., PEARCE, I., ET AL. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication 5* (2011), 31.

[23] M. BLEI, D., Y. NG, A., AND I. JORDAN, M. Latent dirichlet allocation. *The Journal of Machine Learning Research* (2003).

[24] MA, Z., S. A. Y. Q., AND CONG, G. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 999–1008. ACM* (2014).

[25] MICHAEL CONOVER, JACOB RATKIEWICZ, M. F. B. G. A. F. F. M. Political polarization on twitter. *www.researchgate.net/publication/* (2011).

[26] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (2007), ACM, pp. 29–42.

[27] MORALES, A., LOSADA, J., AND BENITO, R. Structure and dynamics of emerging social networks from twitter's conversation# sosinternetve. *Int. J. Complex Systems in Science 1*, 2 (2011), 216–220.

[28] NASKAR, D., MOKADDEM, S., REBOLLO, M., AND ONAINDIA, E. Sentiment analysis in social networks through topic modeling. *https://www.semanticscholar.org/paper/Sentiment-Analysis-in-Social-Networks-through-Topic-Naskar-Mokaddem/00052e8713adf52290c260208cafec05f70035d6?navId=extracted* (2016).

[29] PEDRUELO, M. R., NOGUERA, E. D. V., CASAMAYOR, C. C., CHUST, A. P., AND SÁNCHEZ, F. P. Consensus over multiplex network to calculate user influence in social networks. In *International Journal of Complex Systems in Science* (2013), vol. 3, Interlude, pp. 71–75.

[30] PERLIGER, A., AND PEDAHZUR, A. Social network analysis in the study of terrorism and political violence. *PS: Political Science & Politics 44*, 1 (2011), 45–50.

[31] PLUTCHIK, R. The nature of emotions. *American Scientist.* (2011).

[32] POVEDA, C. M. Sistema multiagente para el análisis de interacciones entre usuarios en medios sociales.

[33] RICHARDSON, L.-J. Micro-blogging and online community. *Internet Archaeology* (2015).

[34] ROMERO, D. M., GALUBA, W., ASUR, S., AND HUBERMAN, B. A. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2011), Springer, pp. 18–33.

[35] ROMERO, D. M., MEEDER, B., AND KLEINBERG, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web* (2011), ACM, pp. 695–704.

[36] RUSSELL, J. Culture and categorization of emotions. *Psychological Bulletin 110* (1991).

[37] SÁNCHEZ, E. V. *Observatorio de interacciones en Twitter*. PhD thesis, 2017.

[38] SMITH, M. A., RAINIE, L., SHNEIDERMAN, B., AND HIMELBOIM, I. Mapping twitter topic networks: From polarized crowds to community clusters. *Pew Research Center 20* (2014), 1–56.

[39] TORRES-NABEL, L. C. Redes deseantes. tendencias político-emocional en redes sociales. *https://uvadoc.uva.es/bitstream/10324/19521/1/redesdeseantes.pdf* (2016).

[40] TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., AND WELPE, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm 10*, 1 (2010), 178–185.

[41] VIVANCO, E., PALANCA, J., DEL VAL, E., REBOLLO, M., AND BOTTI, V. Using geotagged sentiment to better understand social interactions. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (2017), Springer, pp. 369–372.

[42] WANG, Q., B.-J. H. S., AND LUO, B. Classification of private tweets using tweet content. *Semantic Computing (ICSC), 2017 IEEE 11th International Conference* (2017).

[43] WATSON, D. TELLEGEN, A. Toward a consensual structure of mood. *Psychological Bulletin. 98* (1985).