Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

La relación entre los casos de Covid-19 y su impacto en Twitter

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: Raúl Talens Ferrer

Tutor: César Ferri Ramírez

Co-tutor: Fernando Martínez Plumed

2019-2020

Resumen

Este trabajo plantea dos estudios que analizan la relación entre la COVID-19 y las menciones relacionadas con la pandemia en la red social Twitter. Más concretamente, en el primer estudio realizaremos un análisis de variación en la respuesta social en el lapso que se corresponde, en rasgos generales, con el periodo de confinamiento obligatorio. El resultado es que los mensajes enviados son mayormente positivos o neutrales y esto no varía significativamente a lo largo del tiempo. En el segundo estudio comprobaremos la existencia de vínculos y correlaciones entre los tuits publicados relacionados con la enfermedad y la propagación real de la enfermedad en distintos países. A diferencia de lo que podríamos esperar, el resultado obtenido de este análisis es que en todos los países incluidos en el estudio el número de tuits disminuye a lo largo del tiempo, mientras que los casos de contagiados continúan aumentando. Este proyecto ha sido realizado a partir de datos de Twitter y cifras reales de infectados por la COVID-19 proporcionados por entidades internacionales. Para el análisis de los datos y su representación se ha usado el lenguaje de programación Python junto con un cuaderno local de Jupyter y la plataforma web Google Colab.

Palabras clave: COVID-19, Twitter, Python, Análisis de datos, Redes sociales.

Resum

Aquest treball planteja dos estudis que analitzen la relació entre la COVID-19 i les mencions relacionades amb la pandemia a la xarxa social Twitter. Més concretamente, en el primer estudi farem un anàlisi de la variació a la resposta social en el lapse que es correspon, a trets generals, amb el periode de confinament obligatòri. El resultat és que els missatges enviats són majoritàriament positius o neutrals i açò no varia significativament al llarg del temps. Al segon anàlisi comprovarem l'existència de vincles i correlacions entre els tuits publicats relacionats amb l'enfermetat i la propagació real de la enfermetat a distints països. A diferencia del que podríe, esperar, el resultat obtingut en aquest anàlisi és que a tots els paisos inclosos en l'estudi el nombre de tuits disminueix al llarg del temps, mentre que els casos de contagiats continuen aumentant. Aquest projecte ha estat realitzat a partir de dades de Twitter i xifres reals de contagiats per la COVID-19 proporcionats per entitats internacionals. Per a l'anàlisi de les dades i la seua representación s'ha utilitzat el llenguatge de programación Python juntament amb un quadern local de Jupyter i la plataforma web Google Colab.

Paraules clau: COVID -19, Twitter, Python, Anàlisi de dades, Xarxes socials.

Abstract

This work proposes two studies that analyze the relation between COVID-19 and the mentions related to the pandemic on the social network Twitter. More specifically, in the first study, we will carry out an analysis of the variation in the social response during a period that corresponds, in general, with the time of mandatory lockdowns. The result is that people's messages stay mostly positive or neutral, and this fact does not change significantly as time passes. In the second study, we verify the existence of links and correlations between the published tweets related to the disease and its actual spread in different countries. Unlike what we might expect, the result obtained in this analysis is that in all of the countries included in the study the number of tweets decreases in time while the cases of infected people continue to increase. This project has been carried out using data from Twitter as well as real data regarding the spread of COVID-19 provided by international entities. Python, a local Jupyter notebook and Google Colab platform have been used to analyze and visualize data.

Keywords: COVID -19, Twitter, Python, Data analysis, Social Networks.

Tabla de contenidos

Tabla de figuras

1. Introducción

En los últimos años, la red social Twitter se ha convertido en una plataforma imprescindible para la comunicación. Tanto es así, que la mayoría de las instituciones oficiales del mundo la usan, incluso, para transmitir comunicados oficiales. Es por esto por lo que, al verse la sociedad envuelta en una pandemia producida por la enfermedad COVID-19 y sumirse todo el mundo en cuarentena, Twitter se convierte en un retrato de los sentimientos generales sobre la enfermedad.

La COVID-19 aparece a finales de 2019, cuando China avisa al mundo de la aparición de un virus altamente contagioso en la ciudad de Wuhan. Este nuevo virus pertenece a la misma familia de coronavirus que los anteriores MERS y SARS, que se identificaron en 2012 y 2002, respectivamente. El 23 de enero, dado el peligro que empezaba a posar el virus, la región entera a la que pertenece la ciudad de Wuhan quedó bajo cuarentena. Desafortunadamente, el virus ya se había extendido a otras partes del mundo.

En los meses posteriores, uno por uno, casi 100 países han establecido algún tipo de cuarentena para controlar la propagación del virus[1][2][3].

En este trabajo de final grado se presentar un estudio sobre las menciones a la COVID-19 en la red social Twitter. Concretamente, estudiaremos cuál ha sido la reacción social de los usuarios de esta plataforma, analizando las palabras más recurrentes y los sentimientos que transmiten. También intentaremos establecer si existe alguna relación entre el avance de la enfermedad y el número de ocasiones en las que se menciona en la plataforma. Todo esto se llevará a cabo usando el lenguaje de programación Python y diversos cuadernos para escribir código especializados en el análisis de datos.

1.1 Motivación

Durante las prácticas de SIE (Sistemas de información estratégica) se trataron dos de los temas que más me han entusiasmado de la carrera: el análisis y la minería de datos. Fue por esto por lo que, tras sufrir un problema en la empresa donde realizaba las prácticas curriculares que me impidió hacer allí el proyecto de final de grado, no dudé ni un segundo en ponerme en contacto con Fernando, quien fue mi profesor durante estas prácticas. Él mismo me propuso este tema, que me interesó desde el primer momento. También se puso en contacto con César, quien sería mi tutor principal durante el transcurso de este proyecto.

Desde un punto de vista técnico, este trabajo plantea una reflexión sobre la fiabilidad de las redes sociales como reflejo de la sociedad. Se intenta establecer cuanta información veraz y representativa se puede extraer de su análisis. Entrando más en detalle en el tema en cuestión, se ha elegido dada la actualidad del suceso y la relevancia de llevar a cabo un análisis desde un punto de vista más social. Además, se ha tenido en cuenta que, pese a la abundancia de análisis relacionados con el tema, no se ha encontrado ninguno que pueda responder a la pregunta en cuestión.

1.2 Objetivos

Este trabajo parte de un análisis general sobre la pandemia provocada por la enfermedad COVID-19 a través de su repercusión en la red social Twitter. Los objetivos principales son los siguientes:

- Generar un análisis del sentimiento general de la población durante un periodo de tiempo correspondiente al del primer confinamiento obligatorio utilizando las palabras de los tuits con menciones sobre la enfermedad.

- Establecer la relación entre el crecimiento de casos positivos reales de COVID-19 y el aumento en las menciones sobre la enfermedad en la red social Twitter.

1.3 Metodología

El primer paso que se tomó en la realización de este proyecto fue investigar y buscar otros proyectos similares.

En la propuesta del proyecto del profesor ya aparecía una tarea en la plataforma Kaggle que contenía archivos con los tuits que necesitaba para el análisis y, adjuntos a esta tarea, había más de una docena de cuadernos de personas que habían usados estos datos para hacer sus propios estudios. Al ver esto, decidí empezar por ahí y buscar guía entre los mejor valorados.

Después de esto, se estudiaron las distintas plataformas desde las que podía escribir el código y se experimentó con varias de ellas, ya que no poseía experiencia previa en ninguna de ellas.

Por último, se empezaron a generar los análisis empezando por el de sentimiento de los tuits y continuando con la relación entre estos y la expansión de la enfermedad. Estos análisis fueron la parte que llevó más tiempo y constituyen el cuerpo principal del trabajo.

1.4 Estructura

Este trabajo de fin de grado empezará con el capítulo del estado del arte, en el que se hará una recopilación y análisis de los distintos estudios relacionados con el tema en cuestión mencionados brevemente en el apartado anterior. Este apartado también contendrá una descripción y reflexión sobre la ciencia de datos.

A continuación, se verán las herramientas que suelen usarse en este tipo de trabajos y se comentará y justificará la elección de estas. Esto es relevante debido a que en esta área existen multitud de formas de realizar este tipo de análisis de datos, así como de herramientas disponibles.

El siguiente capítulo será el que contenga todo el desarrollo del estudio; desde el planteamiento inicial hasta la conclusión, pasando por los dos análisis que se corresponden con los dos objetivos descritos anteriormente.

Finalmente, reflexionaremos sobre posibles ampliaciones a este trabajo y trabajos relacionados que podrían ayudar a esclarecer gran parte de los resultados.

2. Estado del arte

La recogida y el análisis de datos siempre ha sido un área muy importante en el campo de la ciencia de datos (y la minería de datos), pero se ha hecho aún más relevante desde la aparición y popularización de la informática. A partir de ese momento empiezan a aparecer muchos estudios y análisis realizados con grandes cantidades de datos. Esto es debido a la gran cantidad de datos de la que disponemos hoy en día y al aumento y abaratamiento de potencia computacional. De este tema hablaremos en detalle más adelante.

Si nos centramos ahora en estudios hechos sobre el virus, nos encontramos con multitud de ellos: artículos sobre cómo afecta la propagación de la enfermedad a los sistemas sanitarios[4], informes sobre las consecuencias psicológicas de la COVID-19 y el confinamiento[5], análisis sobre el origen del virus[6], etc. Pero, en concreto, nos interesa ver los que analizan datos similares a los de este proyecto. Es decir, estudios que analizan el impacto de la COVID-19 en las redes sociales. La mayoría de estos artículos e informes se centran en cómo ha aumentado el uso de internet y de las redes sociales durante la pandemia. Con esta premisa, encontramos los siguientes casos:

Estudio viralidad del CORONAVIRUS en las redes sociales – Top Position[7]

Este análisis se centra, principalmente, en cómo se ha disparado la actividad en las redes sociales durante el periodo de la cuarentena. El primer dato que se muestra es la variación de uso de diversas redes sociales en España, habiendo un aumento en Instagram de 22.7%, en Facebook del 36.5% y en Twitter del 56.1%. Después, se han analizado las 10 publicaciones más virales que contienen el término “coronavirus” para las mismas tres plataformas.

Al terminar con los análisis, el estudio concluye que Instagram es la red social más viral con relación a la enfermedad y que el formato que más se repite entre las publicaciones más vistas es el video (57%), seguido de la imagen (33%), enlaces (7%) y, por último, texto (3%).

El uso de redes sociales en España aumenta un 55% en la pandemia de coronavirus[8]

Este es un artículo publicado por el diario ABC en el que, al igual que en el anterior, se muestra cómo la población está aumentado su uso de internet y, en concreto, de las redes sociales. También se contrastan los datos del aumento en las redes sociales en España con los de otros países. Nuestro país se coloca líder en este aumento con un 55%.

Dentro de este artículo también se menciona un estudio en el que se analiza el aumento en el uso de dispositivos móviles, siendo de un 38.3%[9].

El articulo atribuye este aumento en el uso de internet a la necesidad de comunicarse con amigos y familiares y a la búsqueda de una fuente de entretenimiento.

Por último, se menciona que, a pesar del aumento tan significativo en el tiempo de uso, empresas como Twitter han visto reducidos sus ingresos a causa de la reducción en inversiones de anunciantes.

Además de los estudios ya mencionados, encontramos otros trabajos que desarrollan análisis exhaustivos tanto de la propagación de la enfermedad en sí, como de la respuesta social encontrada en redes. Al ser mucho más similares a este trabajo, van a ser los que analicemos con más detalle.

Los dos trabajos están publicados en la plataforma de recopilación y análisis de datos Kaggle[1]. Kaggle es una comunidad online donde tanto principiantes como profesionales en la ciencia de datos comparten su trabajo y compiten para resolver desafíos en este campo.

Antes de empezar a analizar estos trabajos, vamos a hablar un poco de como aparece la ciencia de datos, qué es y qué conocimientos son requeridos para trabajar en este campo.

2.1 La ciencia de datos

Con la llegada de los ordenadores y la informatización de todos los procesos empresariales, aparece la posibilidad de almacenar, organizar y analizar una gran cantidad de datos relevantes para estos procesos. Esta nueva realidad, junto con el aumento y abaratamiento de potencia computacional de este siglo, ha conllevado la creación de esta nueva actividad o profesión.

El término «ciencia de datos» aparece por primera vez en 1962 en un artículo titulado "*The Future of Data Analysis*" de John W. Tukey, quien define el significado de esta expresión. Más tarde, en 1974, el científico Peter Naur utilizó el termino como sustituto de las ciencias computacionales en su libro *Concise Survey of Computer Methods*. Es en este momento cuando el termino pasa a usarse y estudiarse plenamente en el mundo académico. A medida que nos acercamos a los 2000, el termino va ganando fuerza, siendo utilizado por primera vez en un título de una conferencia en 1996 llamada *Ciencia de Datos, clasificación y métodos relacionados*. Actualmente, a causa sobre todo del aumento del uso de las redes sociales, la ciencia de datos ha cobrado muchísima importancia y está considerada una de las profesiones más interesantes del momento. Así lo dijo el economista en jefe de Google, Hal Varían, y lo publicó en un artículo Thomas H. Davenport llamado "*Data Scientist: The sexiest job of the 21st Century"*

La ciencia de datos trata de emplear técnicas de programación para analizar datos. Concretamente, un profesional formado en este campo será capaz diseñar la captura del dato en cualquier entorno, procesar y analizar estos datos, así como extraer el conocimiento y comunicar de manera efectiva cómo usarlo para facilitar la toma de decisiones (UPV, s.f.).

[1] https://www.kaggle.com/

Para llevar a cabo este tipo de trabajos, se aplica el conocimiento de cuatro áreas específicas[10]:

Programación: La programación cumple la función de explicar al computador qué se necesita de él y qué queremos hacer. Para esto se hace uso de un lenguaje de programación específico. Para las tareas de análisis de datos, se puede usar cualquier lenguaje de programación, pero algunos de ellos son mucho más útiles e interesantes que otros. En el siguiente capítulo, los trataremos en más profundidad.

Estadística: Esta área es también fundamental para un buen análisis, ya que en la mayoría habrá que utilizar conceptos como medias, medianas, desviaciones, etc.

Comunicación: Para que todos los resultados y mediciones tengan sentido, sobre todo para un público amplio, es imprescindible una buena visualización de estos resultados. Cualquier persona debería ser capaz de leer a simple vista las tablas y los gráficos y entender qué intentan transmitir.

Conocimiento del dominio: Por último, resulta de muchísima ayuda conocer el campo del que se intenta extraer la información. Esto ayudará a encontrar preguntas interesantes que estudiar, así como a identificar resultados incoherentes.

2.2 Análisis específicos en Kaggle

Volviendo a los análisis, en Kaggle encontramos gran cantidad de ellos dada la propia naturaleza de la plataforma, que es abierta y libre para todos los usuarios. En este trabajo vamos a analizar dos de ellos, uno muy completo pero muy general, y otro que se centra solamente en hacer un análisis de los tuits relacionados con el virus.

Los dos trabajos se han hecho utilizando herramientas parecidas. En concreto los dos usuarios han utilizado el lenguaje de programación Python para la extracción y el análisis de los datos. También han usado algunas librerías extra como pandas, numpy, etc. Como veremos más adelante, este formato de trabajo será el mismo que utilizaremos en nuestro propio análisis. A continuación, describiremos los dos trabajos en detalle.

COVID-19 Analysis, Visualization, Comparison and Predictions – Tarun Kumar

Empezando por el trabajo más general y el que tiene mayor cantidad de *likes* y visualizaciones, encontramos este trabajo de Tarun Kumar[2]. Se trata un análisis general de los efectos y la propagación de la enfermedad tanto a nivel mundial como en distintos países.

[2] https://www.kaggle.com/tarunkr/covid-19-case-study-analysis-viz-comparisons

	Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality Rate (per 100)
0	23456597	809349	15155418	7381606	69137.00	3.45

Figura 1: Datos de casos reportados hasta la fecha (Tarun Kumar, 2020)

El estudio empieza con un análisis general mostrando los valores a nivel mundial de casos confirmados, muertes, recuperados, casos activos, tasa de índice y ratio de mortalidad. Esto sirve para establecer una base desde la cual generar una división por continentes. De esta forma, viendo la tabla, se puede observar que la tasa de mortalidad es mucho más elevada en Europa que en el resto de los continentes (triplica a algunos otros).

	Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality Rate (per 100)
continent						
Africa	1188715	27800	906407	254508	6277.89	2.34
Asia	6296065	130271	4983670	1182124	21491.52	2.07
Australia	27028	544	21394	5090	140.37	2.01
Europe	3382745	205437	1903384	1163707	20076.66	6.07
North America	6772280	255634	2792327	3724319	8548.06	3.77
Others	32893	619	21412	10855	2552.03	1.88
South America	5756871	189044	4526824	1041003	10050.47	3.28

Figura 2: Casos por continente (Tarun Kumar, 2020)

A continuación, se muestran los mismos datos, pero divididos por países. Aquí se pueden apreciar aún más las diferencias, sobre todo en las tasas de mortalidad. Por ejemplo, en algunos países de Europa como Reino Unido, Francia o Italia, la tasa de mortalidad sobrepasa el 10%, mientras que en países del este de Europa ronda el 2-3%. También cabe destacar el caso de Yemen, donde el virus ha provocado una situación alarmante con sistemas sanitarios colapsados y medios más que insuficientes [8]. Esta realidad se refleja en la tabla con una tasa de mortalidad de casi el 30%.

Yemen	1911	553	1086	272	6.41	28.94

Figura 3: Datos de Yemen (Tarun Kumar, 2020)

El análisis general termina con unas tablas que muestran los países con más casos confirmados [figura 4], más muertes [figura 5], más casos activos [figura 6] y más recuperados [figura 7].

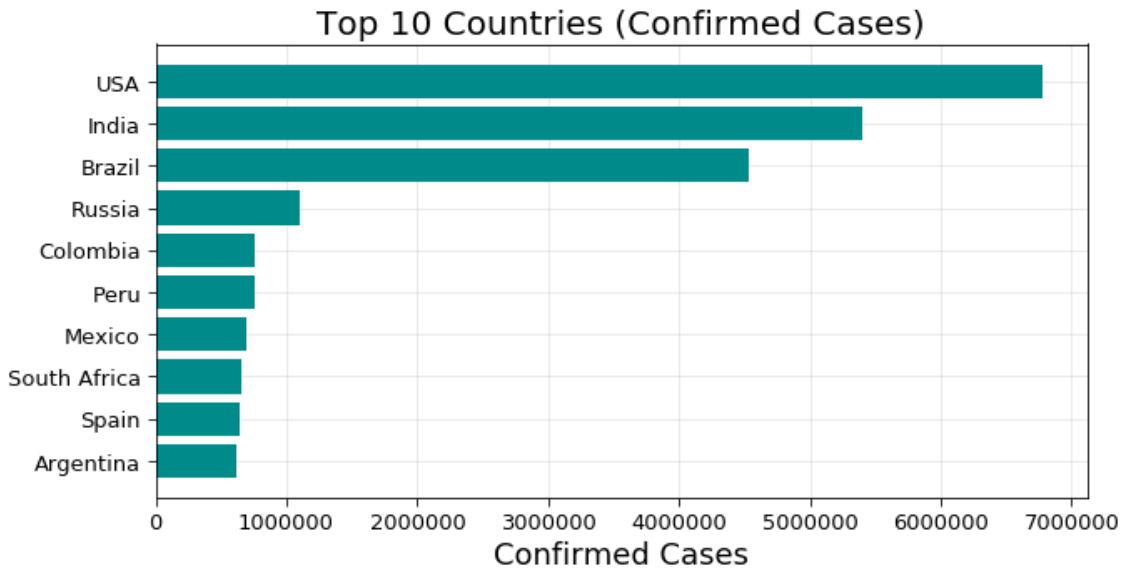


Figura 4: Clasificación de países con más casos confirmados (Tarun Kumar, 2020)

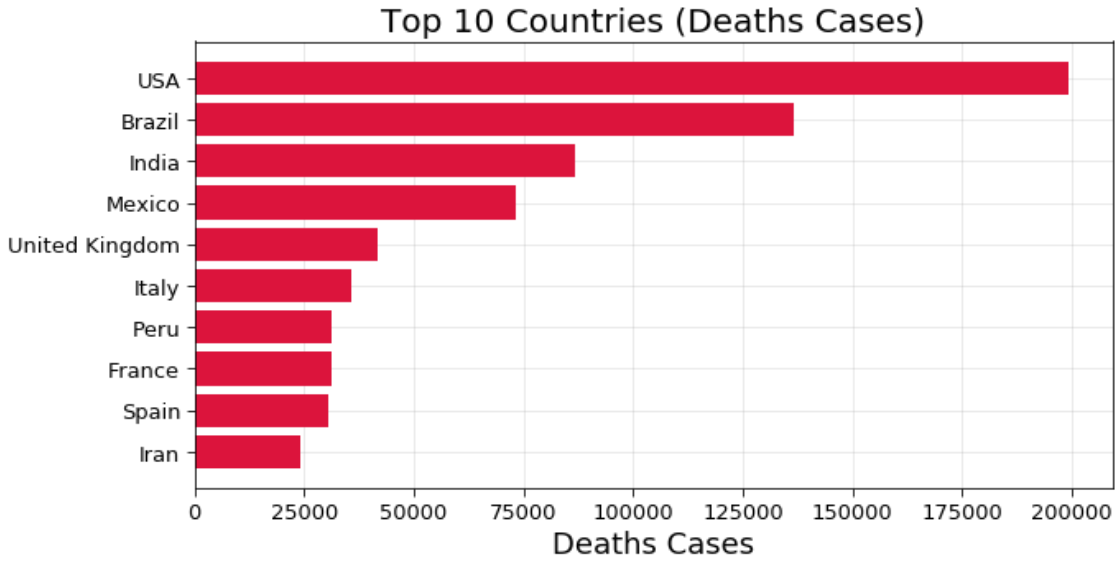


Figura 5: Clasificación de países con más muertes (Tarun Kumar, 2020)

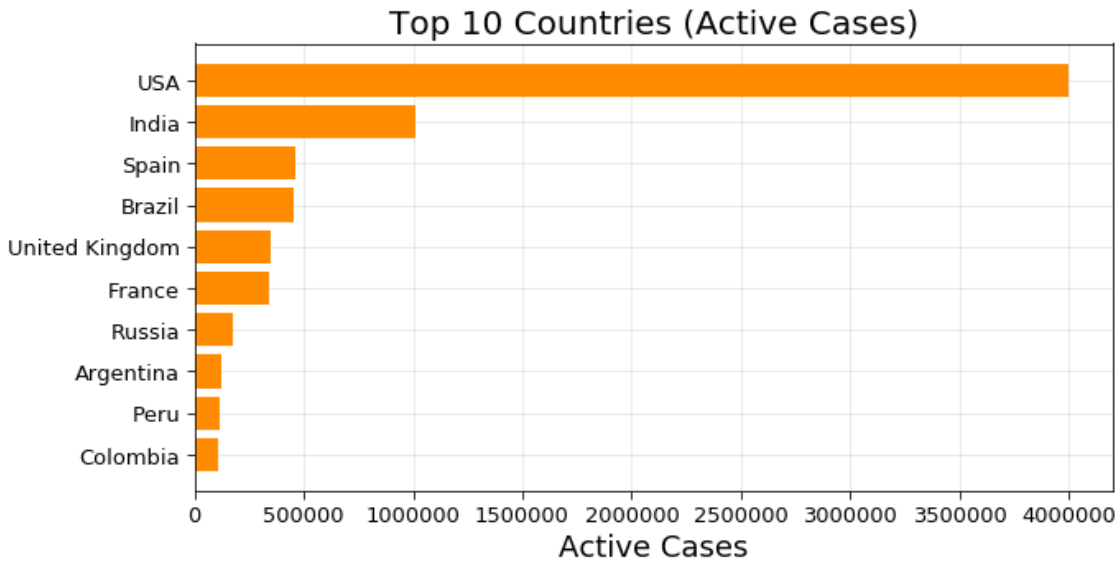


Figura 6: Clasificación de países con más casos activos (Tarun Kumar, 2020)

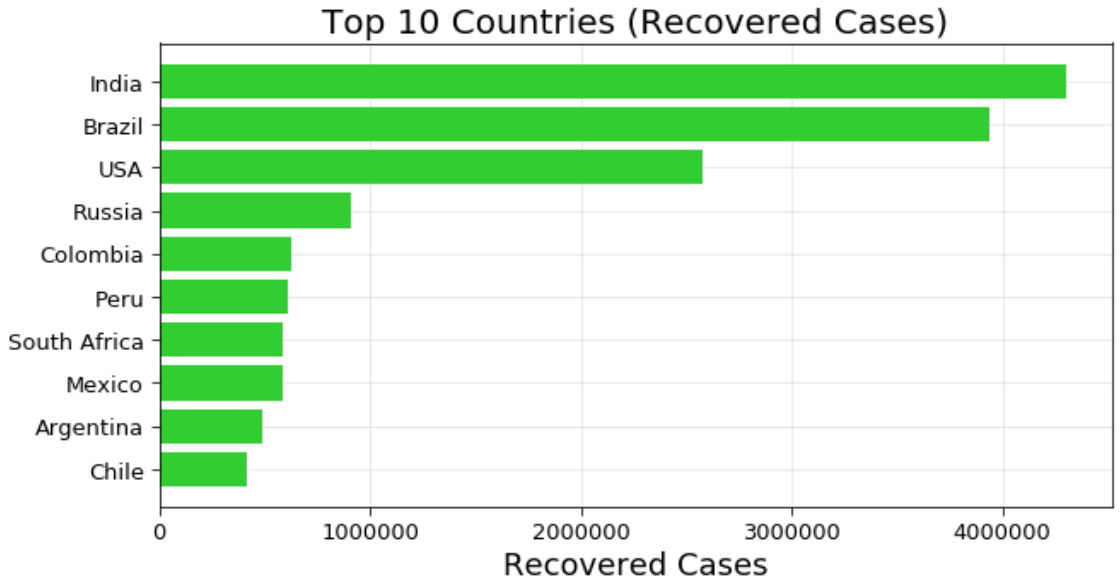


Figura 7: Clasificación de países con más recuperados (Tarun Kumar, 2020)

En estas imágenes podemos ver que Estados Unidos, Brasil y India son los países que lideran las tablas, sobre todo en casos confirmados totales. Llama la atención que, pese a que los tres países tienen un número similar de casos confirmados, Estados Unidos es el que más casos activos tiene en este momento.

Aquí vamos a terminar nuestro repaso al estudio, pero este continúa con más visualizaciones y comparativas de todo tipo de una gran cantidad de países.

Twitter Sentiment Analysis of Covid-19 – Kartik Mohan

Para terminar, estudiaremos el proyecto de Kartik Mohan[3]: un análisis de los sentimientos expresados en la red social Twitter mediante tuits relacionados con la COVID-19. Este ha servido de guía para la primera parte de nuestro trabajo, y es que gran parte de los modelos de programación se han extraído de aquí.

En el trabajo se usan los tuits enviados el día 16 de abril de 2020 y, después de cargarlos, se filtran para usar solamente los tuits enviados desde India y en inglés. Después, se limpia el formato de los tuits para dejar solamente los parámetros deseados cómo la fecha de creación, el texto, el recuento de favoritos y el número de retuits.

A continuación, se muestran distintas mediciones. La primera de ellas es un listado con los tuits que más veces han sido marcados como favorito y los más retuiteados. Después, se muestra un gráfico que clasifica los tuits por la hora a la que se envían [figura 8]. Seguidamente, aparece plasmado un *Wordcloud*, que es un tipo de gráfico que suele usarse para mostrar las palabras que más veces aparecen en cierta muestra. En este caso, las palabras más recurrentes en los tuits[figura 9].

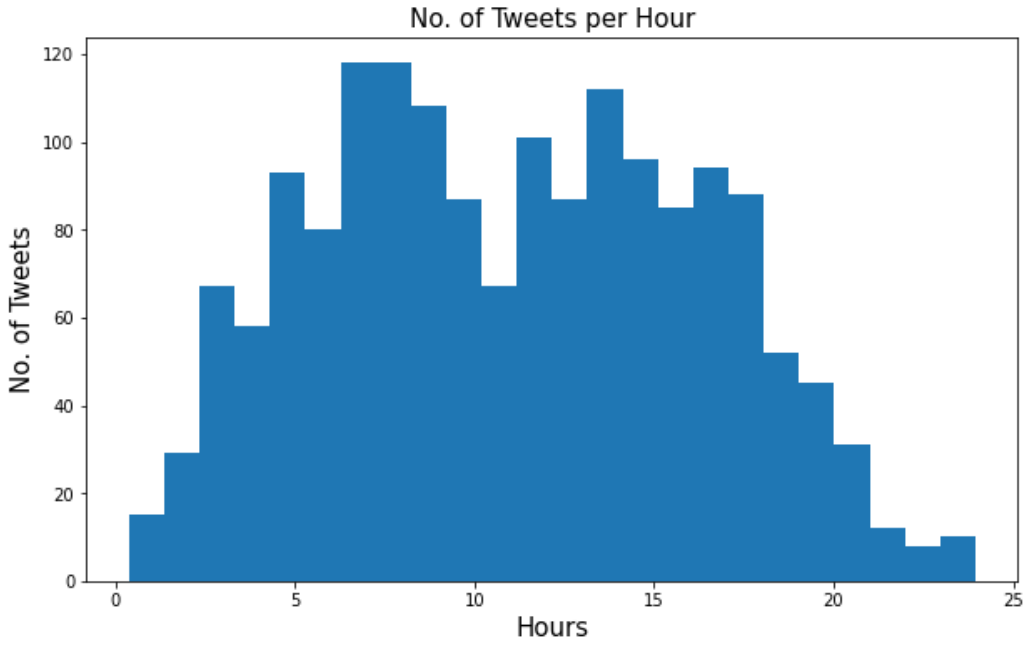


Figura 8: Nº de tuits enviados por hora (Kartik Mohan, 2020)

[3] https://www.kaggle.com/kartikmohan1999/covid19-sentiment-analysis

Figura 9: Wordcloud con las palabras que más veces aparecen en los tuits con menciones al virus (Kartik Mohan, 2020)

Por último, llegamos a la parte principal, el análisis de sentimientos. Antes de empezar con el análisis, se utiliza una función para eliminar las *stopwords*, es decir, palabras de uso muy común que no tienen gran significado propio y que, si se dejasen puestas, diluirían muchísimo el análisis de las demás palabras, que pueden ser mucho más interesantes.

Para el análisis se usa *TextBlob*, que es una librería de Python que dispone de diversas herramientas para el procesamiento de texto. De entre ellas, la que nos interesa en este caso es la llamada "*sentiment",* que puede usarse para añadir un parámetro "*polarity" a cada palabra.* A continuación, se explicará el funcionamiento del código escrito [figura 10].

Primero, se crean dos nuevas columnas en el archivo tweet llamadas *sentiment* y *polarity.* Después, para clasificar los tuits, el autor recorre la lista de los tuits quedándose con el texto. Luego, convierte el texto en tipo *TextBlob* llamándolo «blob» y le asigna a la columna *polarity* el valor del mismo nombre determinado por el algoritmo de *TextBlob*. Para finalizar, simplemente clasifica los tuits en positivos, negativos o neutrales dependiendo del valor de *polarity.*

```
tweet['sentiment'] = ' '
tweet['polarity'] = None
for i,tweets in enumerate(tweet.text) :
    blob = TextBlob(tweets)
    tweet['polarity'][i] = blob.sentiment.polarity
    if blob.sentiment.polarity > 0 :
        tweet['sentiment'][i] = 'positive'
    elif blob.sentiment.polarity < 0 :
        tweet['sentiment'][i] = 'negative'
    else :
        tweet['sentiment'][i] = 'neutral'
tweet.head()
```

Figura 10: Fragmento de código para la clasificación de los tuits por sentimientos (Kartik Mohan, 2020)

Los resultados de este análisis se muestran tanto en un gráfico de barras [figura 11] como en otro que muestra la frecuencia con líneas y barras. [figura 12]

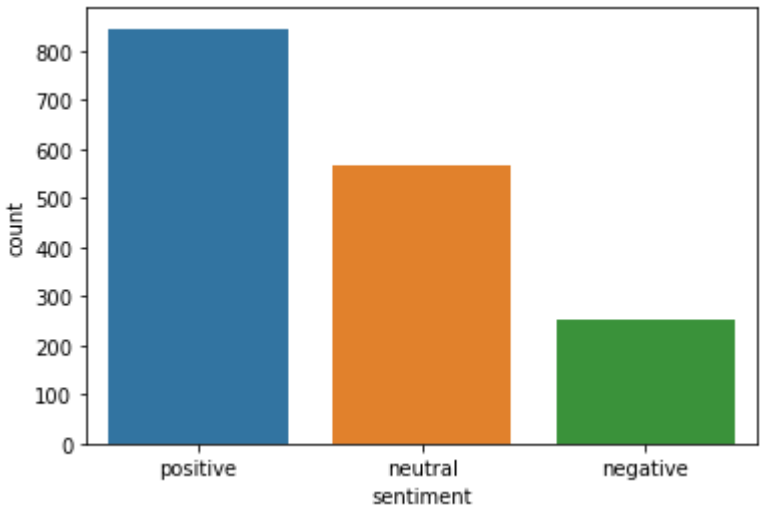


Figura 11: Clasificación de los tuits por sentimiento que transmiten (Kartik Mohan, 2020)

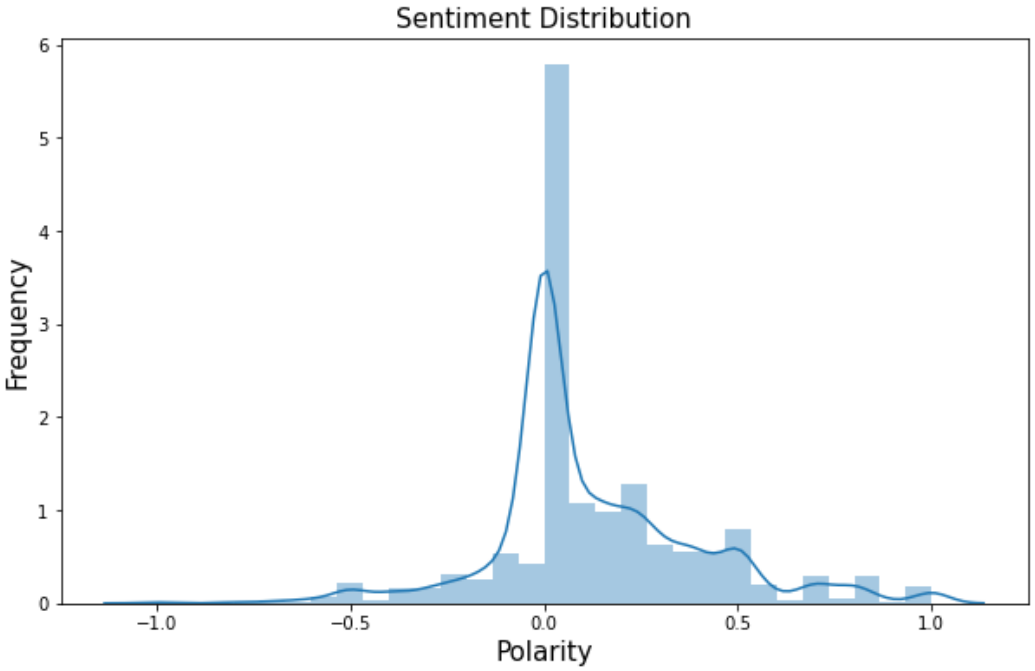


Figura 12: Distribución de los tuits en cuanto a su polaridad (Kartik Mohan, 2020)

El trabajo termina mostrando *wordclouds* con las palabras más usadas en cada tipo de tuit (positivos, negativos y neutrales) y un gráfico de barras con las palabras que más aparecen en todos los tuits.

Después de todas las mediciones y análisis, el autor concluye, a raíz de los resultados, que la mayoría de las personas tienen actitudes y sentimientos positivos o neutrales durante esta fase de la cuarentena.

3. Tecnología utilizada

En este capítulo hablaremos de las herramientas utilizadas para la realización del análisis. Empezaremos explicando cuáles eran las posibles opciones en cuanto a lenguajes de programación y por qué se eligió Python. Después hablaremos de las librerías adicionales y de la elección de una libreta para la escritura y elaboración del análisis.

3.1 Lenguajes de programación

El primer elemento que vamos a tratar son los lenguajes de programación. En concreto, nos vamos a centrar en aquellos más utilizados en el área específica de los análisis de datos.
Teniendo esto en cuenta, los lenguajes de programación que hemos decidido exponer serán R, SAS y Python.

Lenguaje R

R es uno de los lenguajes de programación más utilizados en investigación científica, y está particularmente enfocado hacia el análisis de datos. Fue desarrollado en Bell Laboratories en 1993 y, por su similitud, se puede considerar una implementación diferente del lenguaje S. Hay muchas diferencias entre los dos, pero gran parte del código escrito para S se ejecuta sin ningún cambio en R.

Algunas de las técnicas estadísticas que nos puede proporcionar R son: modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento, etc. Además, también cuenta con técnicas gráficas y es altamente extensible. Esto último conforma uno de los puntos fuertes de R, y es que se pueden conseguir gráficos de muy buena calidad. También incluye símbolos matemáticos y fórmulas.

R se pude encontrar como software libre bajo licencia GNU GPL y está disponible para Windows, Mac, Unix y Linux.

Lenguaje SAS

SAS es un lenguaje de programación desarrollado por SAS Institute a finales de los años 60 que cuenta con dos intérpretes: el desarrollado por SAS Institute, y otro de la empresa World Programming. Las dos versiones han ampliado el lenguaje con módulos, especialmente para el análisis de datos. A continuación, nos vamos a referir a las características base que se poseen las dos versiones.

SAS opera principalmente con tablas de datos; las puede leer, modificar y crear informes a partir de ellas. Como resumen, estas son algunas de sus funciones:

- Pasos *data* que permiten hacer operaciones sobre las filas de un conjunto de datos.
- Procedimientos de manipulación de datos que permiten ordenar tablas, enlazarlas, etc.
- Un intérprete de SQL.
- Un *superlenguaje* de *macros.*

Python

Al contrario que los dos anteriores, Python es un lenguaje generalista, lo cual significa que cuenta con herramientas y funciones para programar cualquier cosa. Aun así, es uno de los lenguajes más utilizados para tareas de análisis de datos.

Python fue creado a principios de los años noventa por Guido van Rossum en el CWI (Stichting Mathematisch Centrum) como sucesor al lenguaje ABC. Python es un lenguaje de programación interpretado, orientado a objetos de alto nivel y con semántica dinámica. Pone gran

énfasis en la simplicidad y legibilidad del código y ofrece la potencia y la flexibilidad de los lenguajes compilados, pero con una curva de aprendizaje suave.

Python ha sido el lenguaje elegido para llevar a cabo este análisis. Su sencillez, el hecho de que sea de uso libre y su gran comunidad y soporte han hecho muy clara la elección. Además, lo que hace a Python un candidato perfecto para el análisis de datos son sus librerías, ya que cuenta con gran cantidad de ellas. Por ejemplo, hay librerías que nos proporcionan herramientas científicas, numéricas, de análisis y estructura de datos o de algoritmos de Aprendizaje Automático. De esto hablaremos en más detalle a continuación. Otro aspecto que se tuvo en cuenta para su elección, sobre todo comparándolo con R, es que Python es un lenguaje generalista. Es por esto por lo que, al no tener experiencia con ninguno de los dos, consideramos preferible dedicar tiempo a aprender Python porque puede ser útil en futuros trabajos que no estén tan centrados en la ciencia de datos.

Librerías complementarias de Python

Como ya hemos dicho, una de las grandes ventajas de usar Python es la gran cantidad de librerías complementarias que existen. De entre todas ellas, las que se han usado para el trabajo son las siguientes:

- **Numpy**: Esta es la librería más importante en lo que respecta al análisis de datos en Python. No obstante, no se usa por sí misma, sino que usamos librerías que derivan y dependen de ella. En concreto, todas las que aparecen a continuación menos TextBlob dependen de Numpy para su funcionamiento.
- **Pandas**: Pandas es una herramienta para análisis y manipulación de datos. Su función principal ha sido leer los datos de un archivo y transformarlos en un *Dataframe* para poder trabajar con ellos.
- **Matplotlib:** Librería para la creación de gráficos y otros tipos de visualizaciones.
- **Seaborn**: Derivada de Matplotlib. Se han usado las dos librerías en conjunto para la creación de todas las representaciones.
- ***WordCloud* y Stopwords:** Estas no son librerías sino más bien complementos del propio Python. *Wordcloud* nos permite plasmar las palabras más usadas en imágenes generadas automáticamente y Stopwords es una lista con las palabras más usadas en cierto idioma y que el programa debe ignorar (palabras como artículos o conjunciones).
- **TextBlob:** Librería que aporta herramientas para trabajar con texto. En concreto, gracias a ella se ha podido llevar a cabo el análisis del sentimiento de los tuits.

3.2 Plataforma

Después de elegir Python como lenguaje, debemos elegir en qué plataforma escribir el código. Para trabajos de análisis de datos, se aconseja usar cuadernos ya que de esta forma puedes ejecutar el código y ver el resultado en el momento. También puedes añadir comentarios entre distintos fragmentos de código. Todo esto, junto a las visualizaciones de los resultados en forma de gráficos

y demás, forma un documento que incluye comentarios, código y resultados. Este formato es el perfecto para compartir con otras personas que trabajen en el mismo sector.

Hoy en día, encontramos muchas opciones de cuadernos, de entre todos ellos hemos elegido tres que explicar en detalle.

Jupyter

La primera alternativa y una de las más famosas es Jupyter Notebook. Esta es una aplicación de código abierto que cuenta con todas las características descritas anteriormente. El proyecto Jupyter surgió en 2014 a partir de IPython, que fue la primera plataforma de este tipo que existió, desarrollada por la propia empresa detrás de Python. Desde ese momento, los componentes de IPython siguieron funcionando, pero contenidos bajo el sistema Jupyter.

En resumen, Jupyter es un cuaderno que se ejecuta de forma local (utilizando el *hardware* de nuestro propio ordenador) y nos permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones y comentarios.

Kaggle Kernel

Otra alternativa interesante, y más teniendo en cuenta el uso de la plataforma Kaggle para la investigación de otros proyectos similares, sería hacer uso del cuaderno de esta misma entidad. En 2015 y también a partir del proyecto IPython, Kaggle implementó este software en su propia web, llamándolo Kaggle Kernels. Estos cuadernos son, en esencia, iguales que un cuaderno de Jupyter pero utilizan *hardware* prestado por la propia empresa. Esto incluye la memoria RAM y la GPU.

La principal ventaja de este tipo de implementación es que, al no utilizar el *hardware* de nuestro propio ordenador, no necesitamos una máquina potente para poder realizar los cálculos. Tampoco necesitamos instalar nada en nuestro ordenador ya que funciona todo por web. Además, al ser un servicio de Kaggle nos puede facilitar mucho la tarea de compartir el documento en la propia web, si es eso lo que queremos.

Google Colab

La última propuesta y la que hemos utilizado casi exclusivamente es Google Colab. Google Colab es muy similar al sistema de Kaggle Kernels contenido dentro del entorno Google. Al igual que la versión de Kaggle, dispone también de servidores remotos para realizar las ejecuciones. Sin embargo, la ventaja más significativa respecto a las anteriores alternativas es que los cuadernos se guardan directamente en nuestra cuenta de Google, teniéndolos así disponibles en todos nuestros dispositivos de forma automática.

3.3 Instalación y detalles

Como resumen de esta sección, explicaremos qué herramientas se han utilizado y por qué.

Después de elegir Python 3.0 como lenguaje de programación se empezó a buscar cómo instalarlo. La opción más evidente era instalar la versión oficial de Python desde su propia página web, pero no es la mejor para esta tarea.

Python, como cualquier lenguaje de programación, cuenta con distribuciones ajenas a la versión oficial. Estas opciones, además del software para el lenguaje en sí, cuentan con distintas extensiones que mejoran o hacen más fácil su funcionamiento. De entre estas se decidió instalar Anaconda.

Anaconda es un kit de herramientas pensado para la ciencia de datos que, entre otras cosas, cuenta con la mayoría de las librerías utilizadas ya instaladas. Entre ellas NumPy, Pandas y Matplotlib. También con el cuaderno de Jupyter.

Este cuaderno Jupyter, dentro de Anaconda, fue donde se empezó a trabajar. Más adelante, se abandonó casi por completo en favor de la versión de Google Colab. El motivo por el cual se tomó esta decisión fue que, al estar trabajando con una gran cantidad de datos, la mayoría de las ejecuciones son muy largas. Esto provocaba tener que dejar el ordenador muchas horas sin poder utilizarlo y con miedo a que el proceso terminase con algún error a causa de la falta de memoria local.

Una vez habiéndose pasado a utilizar Colab, la mayoría de los problemas se solucionaron. Solo se volvió a utilizar un cuaderno local de Jupyter cuando había que leer algún archivo local que no conseguía subir a Colab.

4. Desarrollo del análisis

Por último, llegamos al capítulo principal de este trabajo de fin de grado, donde explicaremos en detalle el desarrollo del análisis. Empezaremos la sección viendo cuál era la situación inicial antes del trabajo. Continuaremos viendo en qué partes se ha dividido el estudio y por qué, y terminaremos analizando los resultados.

4.1 Situación inicial

Antes de empezar a describir el proceso del estudio, vamos a recordar cuáles eran los análisis relacionados que podíamos encontrar y por qué se decidió realizar este trabajo.

En el segundo capítulo del trabajo, hemos descrito varios proyectos que hacen una muy buena labor analizando distintos aspectos de la pandemia. El primero de ellos hacía un estudio general por países en el que se indicaba cuáles de ellos estaban en peor o mejor situación según distintas medidas (contagios, tasa de incidencia, etc.) El segundo, en cambio, sí se centra en los tuits relacionados, pero no responde a la pregunta de si hay alguna conexión entre la propagación del virus y el aumento o disminución del número de tuits con menciones a la enfermedad.

Además de los análisis vistos en este trabajo, todavía hay muchísimos más que no hemos tratado, y es que, al encontrarnos aún en medio de la pandemia provocada por el virus, la cantidad de información relacionada con ella es abrumadora. Es por esto por lo que, pese a poder encontrar una gran cantidad de estudios similares, era muy complicado encontrar uno que respondiera a la pregunta concreta que nos hacíamos: si el aumento en casos de la enfermedad provocaba un aumento en el número de tuits. Así pues, con esta pregunta en mente empezamos el estudio.

4.2 Recopilación de datos

La primera y una de las partes más importantes en un estudio es la recogida de datos. En nuestro caso, los primeros archivos que obtenemos son los tuits. Estos los descargamos de una tarea de Kaggle publicada por el usuario Shane Smith[4] . De aquí descargamos los archivos en cuestión, que contienen datos de tuits de usuarios que usaron los *hashtags*: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19, #epitwitter, #ihavecorona, #StayHomeStaySafe, #TestTracelsolate. Los datos disponibles abarcan desde el día 29 de marzo del 2020 hasta el 1 de mayo de 2020 y contienen toda la información de la que dispone Twitter. Entre toda esta información, los campos sobre los que hemos trabajado son:

- Fecha de creación
- Texto del tuit
- País desde el que se envía
- Número de favoritos
- Número de retuits

Estos datos se encuentran divididos en archivos con formato .CSV (*Coma Separated Values*), un archivo por cada día [figura 13] y en distintas tareas de Kaggle en las que se van expandiendo los *hashtags* utilizados. En el siguiente apartado veremos cómo hemos abierto y trabajado con esos datos.

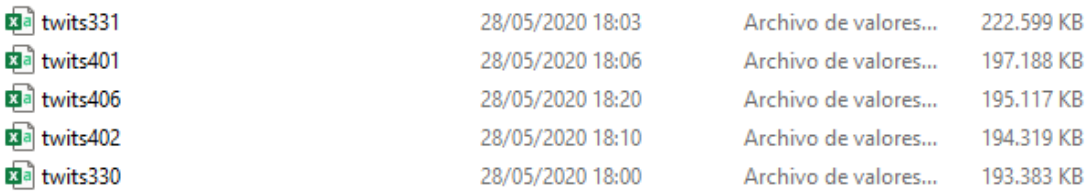


Figura 13: Archivos de tuits separados por día

Después de obtener los tuits es el turno de los datos sobre los contagios. Estos, a diferencia de los tuits, son mucho más fáciles de encontrar y los podemos conseguir de infinidad de fuentes.

[4] Enlace del sitio web de la tarea de Kaggle https://www.kaggle.com/smid80/coronavirus-covid19-tweets

En nuestro caso, hemos utilizado la misma fuente que utilizaba el primer estudio que hemos comentado en el capítulo 2, Worldometers[5]. Esta web contiene, entre otras cosas, una sección dedicada a la pandemia en la que podemos encontrar los datos de contagiados, muertos y recuperados a nivel mundial y por países [figura 14].

Date	Country	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	WHORegion
22/01/2020	Afghanistan	0	0	0	0	0	0	0	Eastern Mediterranean
22/01/2020	Albania	0	0	0	0	0	0	0	Europe
22/01/2020	Algeria	0	0	0	0	0	0	0	Africa
22/01/2020	Andorra	0	0	0	0	0	0	0	Europe
22/01/2020	Angola	0	0	0	0	0	0	0	Africa
22/01/2020	Antigua and Barbuda	0	0	0	0	0	0	0	Americas
22/01/2020	Argentina	0	0	0	0	0	0	0	Americas
22/01/2020	Armenia	0	0	0	0	0	0	0	Europe

Figura 14: Formato del archivo que contiene los datos de casos producidos por la COVID-19

4.3 Análisis

En ambos de los siguientes apartados sobre los análisis se va a explicar paso a paso el procedimiento para la extracción de resultados y a mostrar las tablas y gráficos más importantes para su lectura. Cabe decir que, pese a que la mayor parte del estudio se ha hecho con Google Colab, algunas operaciones están hechas con un cuaderno Jupyter local. Es por esto por lo que antes de hacer un cambio de plataforma esto se indicará en el texto. Todos los cuadernos de los análisis estarán disponibles para su visualización en los anexos en forma de vínculo.

Durante la explicación de los análisis aparecerán los resultados del primer día analizado, en el caso del análisis de tuits, y de España en el caso de la comparativa entre tuits y contagios. Los resultados de los demás días y países estarán disponibles en los anexos.

Análisis de tuits

Este primer análisis se centrará en los tuits extraídos. Concretamente, en el texto de estos tuits y en los sentimientos que transmiten. Para establecer una progresión en el tiempo, se han estudiado seis días distintos en el periodo de aproximadamente un mes que contenían los archivos. Cada uno de estos días tendrá su propio cuaderno, gráficos y resultados, aunque aquí solo nos centraremos en explicar el procedimiento y mostrar los resultados de cada día analizado para compararlos entre ellos y establecer si existe o no un cambio significativo. Este análisis se ha realizado utilizando exclusivamente un cuaderno de Google Colab.

El primer paso para empezar a trabajar es crear un cuaderno. Para hacer esto, simplemente visitamos la web de Google Colab y pulsamos en el menú Archivo y Nuevo Cuaderno, tal y como se muestra en la siguiente figura.

[5] Worldometers.info

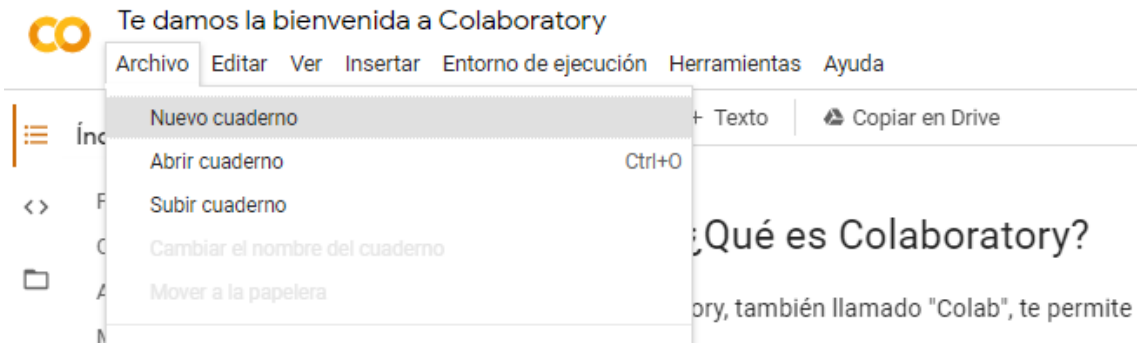


Figura 15: Indicaciones para usar Google Colab

Una vez dentro del cuaderno que acabamos de crear ya podemos empezar a escribir código. En este tipo de cuadernos, el código se distribuye por bloques o celdas. Cada una de estas celdas se puede ejecutar de forma independiente. Además, se pueden crear tanto celdas de código como de texto. Esto resulta conveniente para poder explicar los bloques de código en el mismo documento.

El primer bloque de código que se introducirá será el correspondiente a las librerías adicionales que importaremos para añadir funciones extra. De estas ya se ha hablado en el capítulo anterior, pero se explicarán en mayor detalle cuando se haga uso de ellas.

Una vez hecho esto, usaremos la librería «pandas» para leer el archivo de los tuits y convertirlo en un *dataframe*, que es un tipo propio de pandas con el que será mucho más sencillo trabajar.

Después, eliminamos las columnas innecesarias y filtramos los tuits para quedarnos solamente con aquellos escritos en inglés. Este paso es necesario porque solamente disponemos de las herramientas necesarias para realizar un análisis de sentimientos con las palabras escritas en inglés. En un proyecto futuro se podría ahondar en este aspecto.

A continuación, usamos el módulo de Python «re» (*regular expressions*) para obtener a partir del campo de los tuits «texto», una lista de *strings* que no contenga caracteres especiales (paréntesis, puntos, comas, etc.).

Como último paso para limpiar la lista, podemos eliminar las palabras más comunes sin significado propio. Son las denominadas *stopwords*. Para hacer esto disponemos de una lista predefinida proporcionada por la propia librería de Python. Un ejemplo del resultado final lo podemos ver en la figura 16.

```
0    coronavirus news alert dr vladimir zelenko boa...
1    something doesn t add global panic extreme act...
2    wuhan residents estimate based calculations cr...
3    adani foundation humbled contribute rs 100 cr ...
4    media t contain glee amp delight reporting u s...
Name: text, dtype: object
```

Figura 16: Ejemplo del texto de los tuits limpio

Con el atributo texto preparado, podemos empezar a obtener mediciones y resultados. El primero de ellos es un *Wordcloud*, que es un tipo de gráfico que muestra las palabras más usadas

en la muestra que le proporcionamos. En nuestro caso, vamos a reunir las palabras que más veces aparecen en tuits con menciones a la COVID-19 o términos relacionados. El resultado es el siguiente [figura 17]:

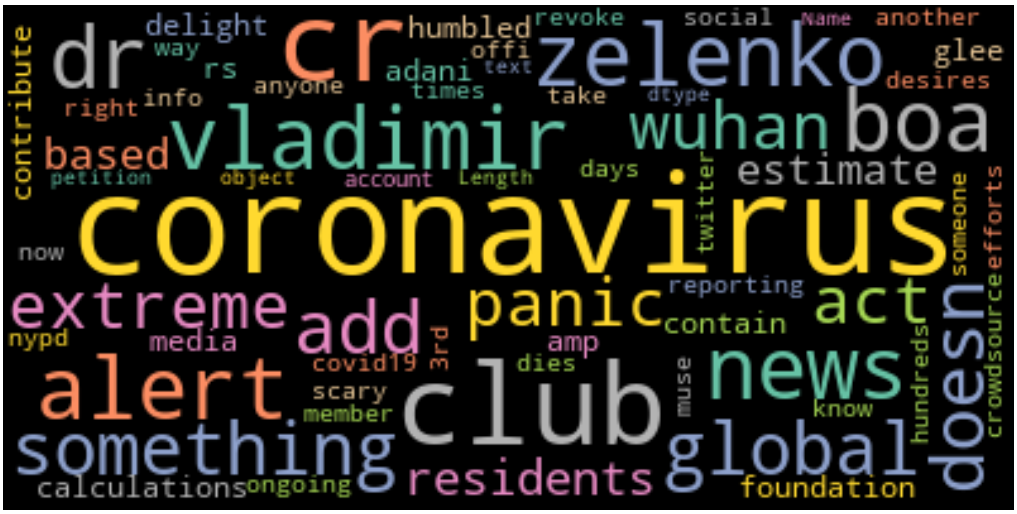


Figura 17: Wordcloud con las palabras que más veces aparecen en los tuits con menciones al virus

En la imagen destaca sobretodo la palabra "*coronavirus*" tanto por tamaño como por la posición central que ocupa. Además de esta, también podemos leer "*club*", "*vladimir*" "*zelenko*" y "*wuhan*" entre otras.

No obstante, una cosa que debemos tener en cuenta al observar estos resultados es que el algoritmo que usa *Wordcloud* no es estricto en cuanto a los tamaños de las palabras. Con esto queremos decir que, aunque en general las palabras que aparecen más grandes en la imagen son las más usadas, cuando la diferencia no es muy grande puede ser que aparezca una palabra que tiene menor frecuencia en lugar de otra para que quepa en el gráfico. Esto hace que este tipo de gráficos aparezca cada vez que lo ejecutamos de una forma distinta, por lo tanto, no los podemos usar para comparar correctamente la frecuencia de aparición de las palabras. Para solventar este problema, también podemos mostrar estas palabras en un gráfico más tradicional. [figura 18]

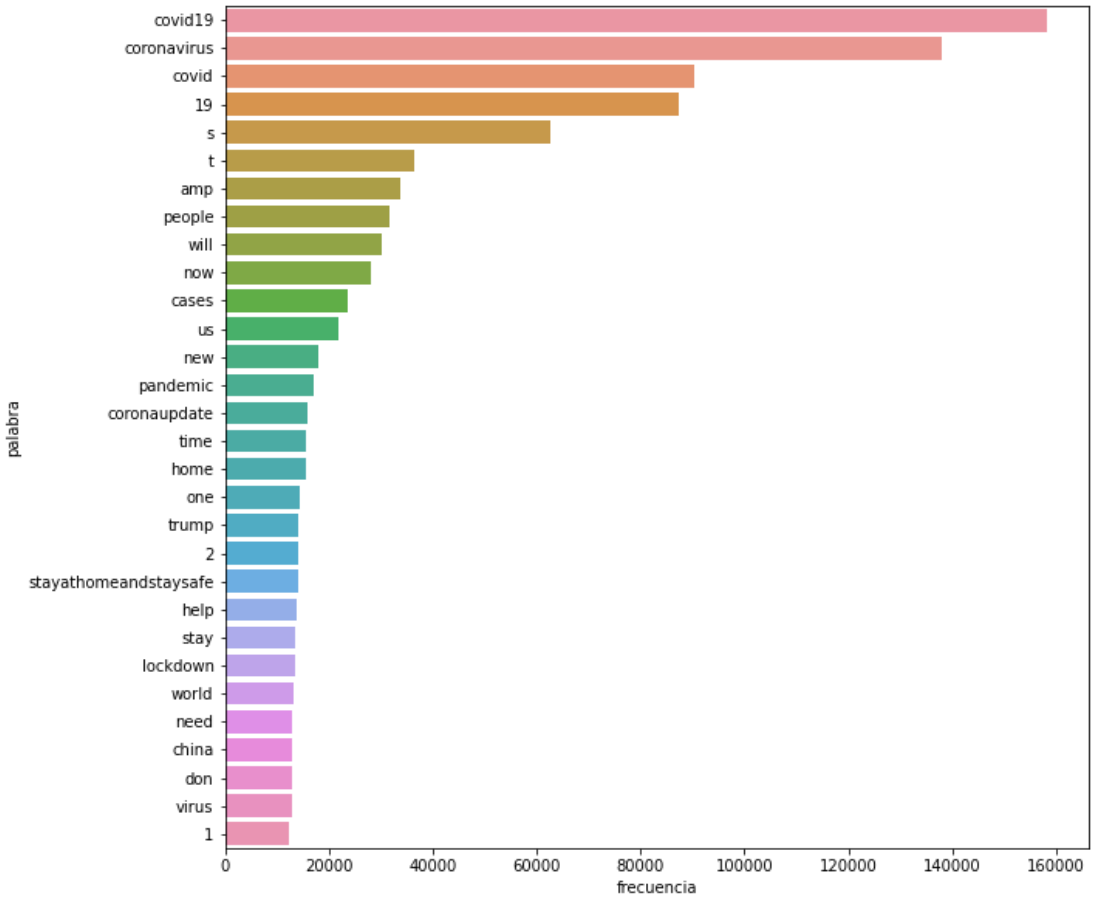


Figura 18: Palabras más veces usadas en los tuits con menciones

Este resultado sí sería mucho más fiable en cuanto a la frecuencia de aparición de las palabas. En este caso, los resultados[6] de todos los días analizados son muy similares, por lo tanto, no los vamos a mostrar aquí.

Seguidamente, empezamos con la parte principal de este primer análisis, que es la relacionada con los sentimientos. Cabe recordar que nuestro objetivo con este análisis es ver cuál es la respuesta social frente a la pandemia y a la cuarentena obligatoria y ver cómo evoluciona a lo largo de un mes.

El primer paso consiste en añadir los campos de "sentimiento" y "polaridad" al *dataframe* que contiene los tuits, que recordemos es como una tabla con distintos atributos. El primero de ellos tendrá el valor positivo, negativo o neutral y el segundo, un número que será negativo, positivo o cero dependiendo de las palabras que aparezcan en el texto. Este valor lo podemos conseguir gracias a la librería *TextBlob*, que utiliza un algoritmo para determinar qué polaridad tienen ciertas palabras y les otorga un número entre el -1 y el 1. En la siguiente imagen tenemos un ejemplo para entenderlo mejor [figura 19]. En ella vemos que a la palabra *hate* se le asigna una polaridad de -0.8, por lo tanto, será clasificada como negativa. En cambio, a la palabra *love* se le asigna un 0.5, es decir, positiva. También vemos que le asigna un parámetro de subjetividad pero que para nuestro análisis no lo hemos utilizado.

[6] Para ver los resultados de esta y todas las mediciones del análisis, Anexo A (resultados del análisis de tuits)

```
blob_negativo = TextBlob('hate')
print(blob_negativo.sentiment)
```

```
Sentiment(polarity=-0.8, subjectivity=0.9)
```

```
blob_positivo = TextBlob('love')
print(blob_positivo.sentiment)
```

```
Sentiment(polarity=0.5, subjectivity=0.6)
```

Figura 19: Ejemplo clasificación por sentimiento

Después de aplicar esa numeración a cada tuit, los clasificamos en positivo si el número es mayor que 0, negativo si es menor o neutral si el texto no contiene ninguna palabra con valor para el algoritmo. El resultado de esta clasificación lo podemos ver en los gráficos de barras a continuación:

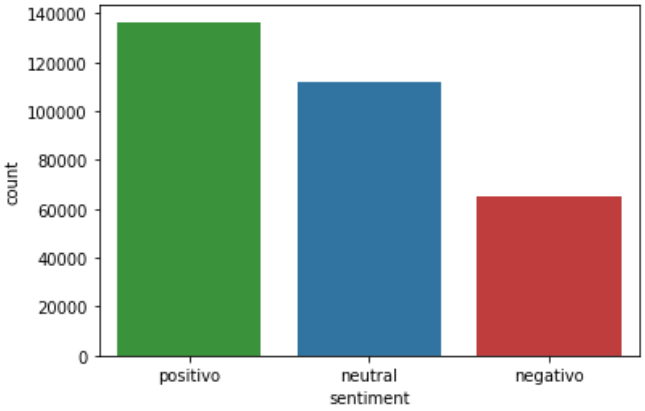


Figura 20: Clasificación de tuits por sentimiento. Dia 29 de marzo de 2020

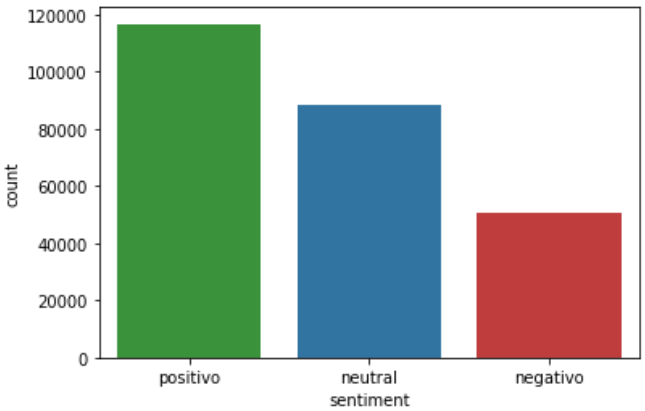


Figura 21: Clasificación del 5 de abril

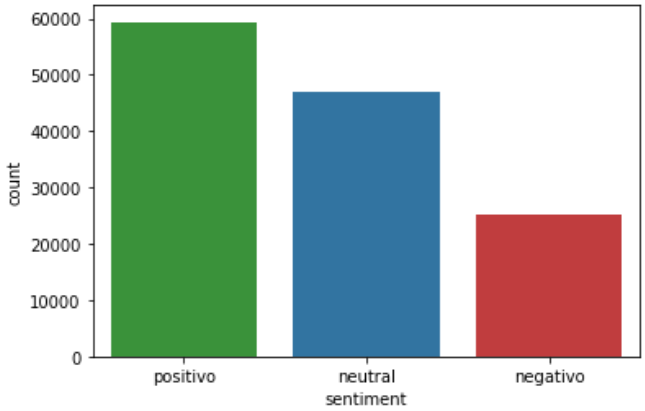


Figura 22: Clasificación del 12 de abril

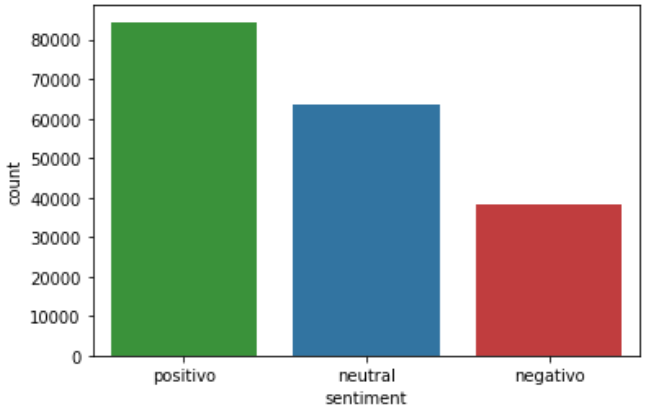


Figura 23: Clasificación del 19 de abril

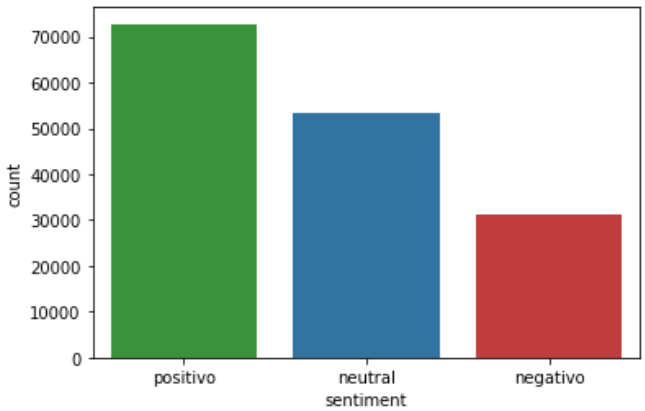


Figura 24: Clasificación del 26 de abril

En concreto, en la muestra de tuits del día 29 de marzo de 2020 escritos en inglés, detectamos 136 549 tuits positivos, 111 645 neutrales y 64 842 negativos.

Podemos ver que en todas las muestras se observa el mismo resultado; los tuits mayoritarios son los clasificados como positivos, seguidos de los neutrales y los negativos son los menos comunes. También podemos mostrar los resultados de otra forma. Para eso vamos a utilizar un gráfico de barras que utiliza el valor de polaridad en lugar de la clasificación final [figura 25]. De esta manera se aprecia mucho mejor que, aunque la clasificación ubica a la mayor parte de los tuits en la categoría de positivos, la mayoría lo son por muy poco. Sería más acertado decir que el sentimiento que está más presente en los tuits es el de neutralidad.

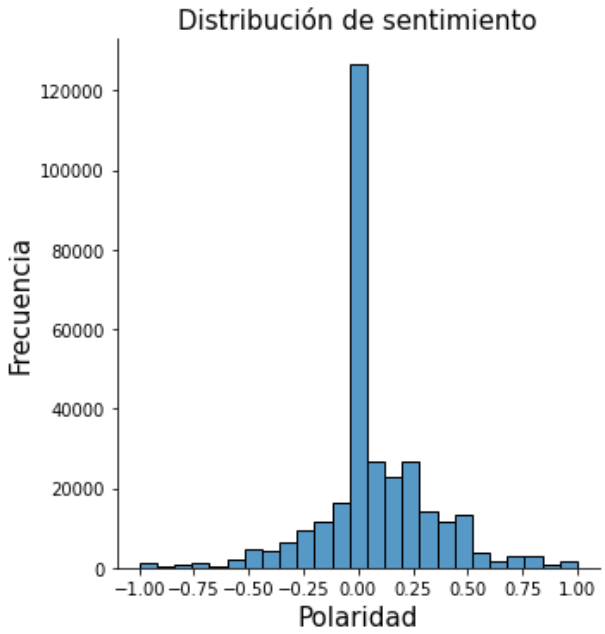


Figura 25: Distribución de los tuits en cuanto a su polaridad. Dia 29 de marzo de 2020

A partir de esta clasificación, hemos vuelto a utilizar *Wordcloud*, esta vez para mostrar las palabras más utilizadas en cada tipo de tuit (positivo, negativo o neutral). Los resultados del primer día los podemos ver en las siguientes figuras.

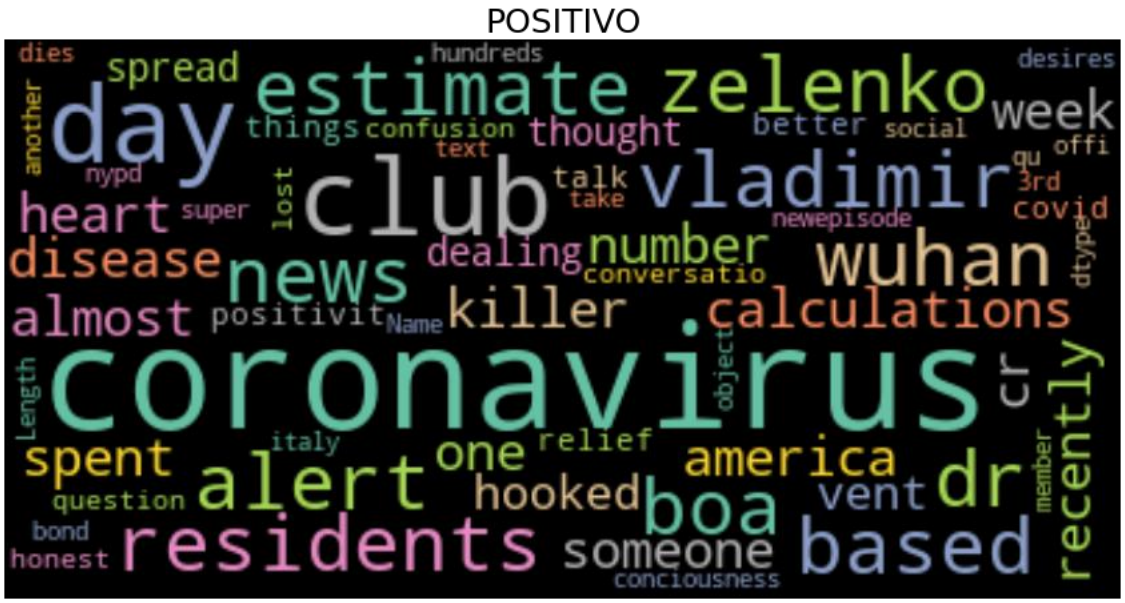


FIgura 26: Wordcloud con las palabras más usadas en los tuits clasificados como positivos

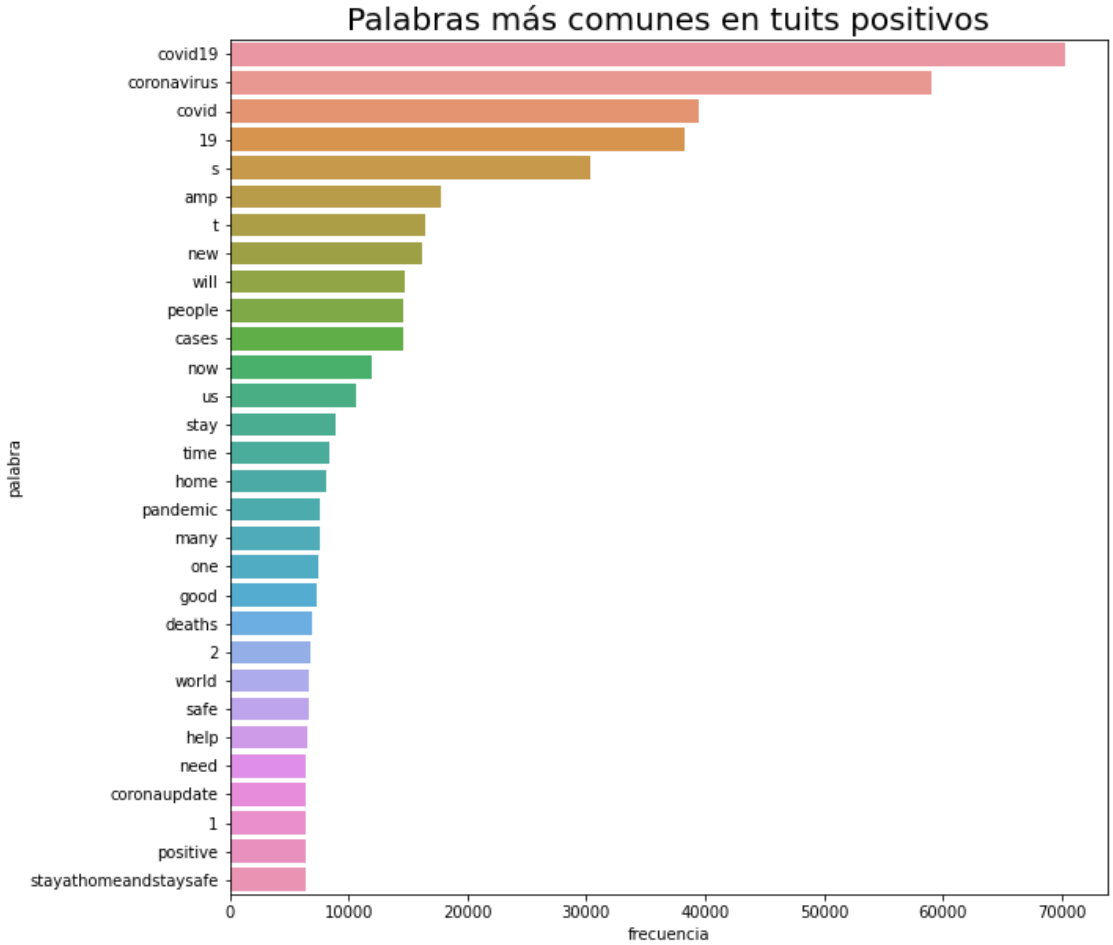


Figura 27: Palabras más usadas en tuits positivos

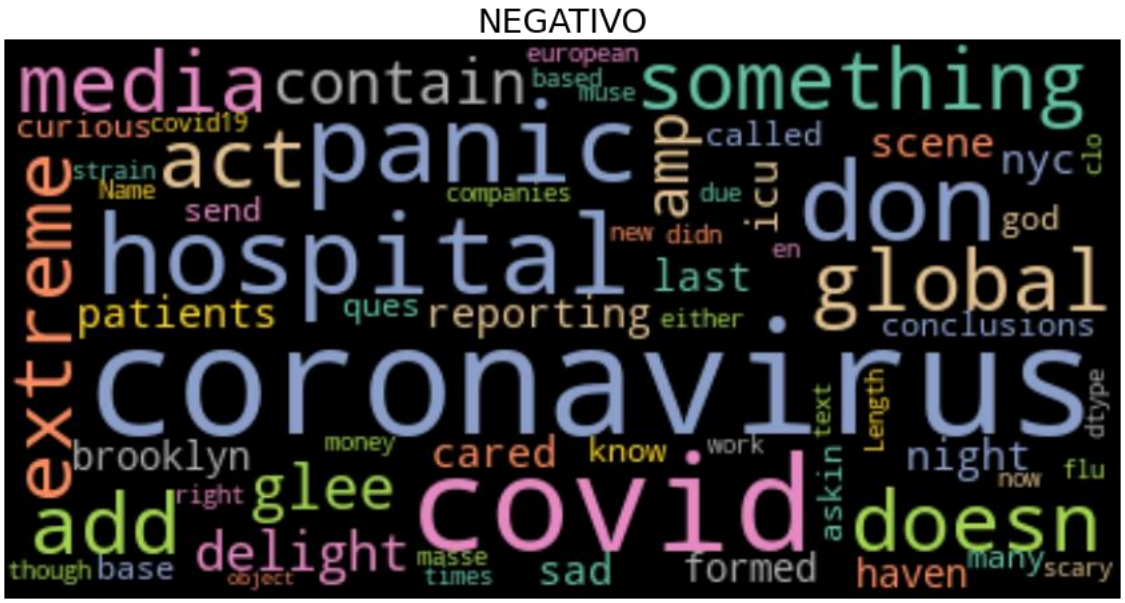


Figura 28: Wordcloud con las palabras más usadas en los tuits clasificados como negativos

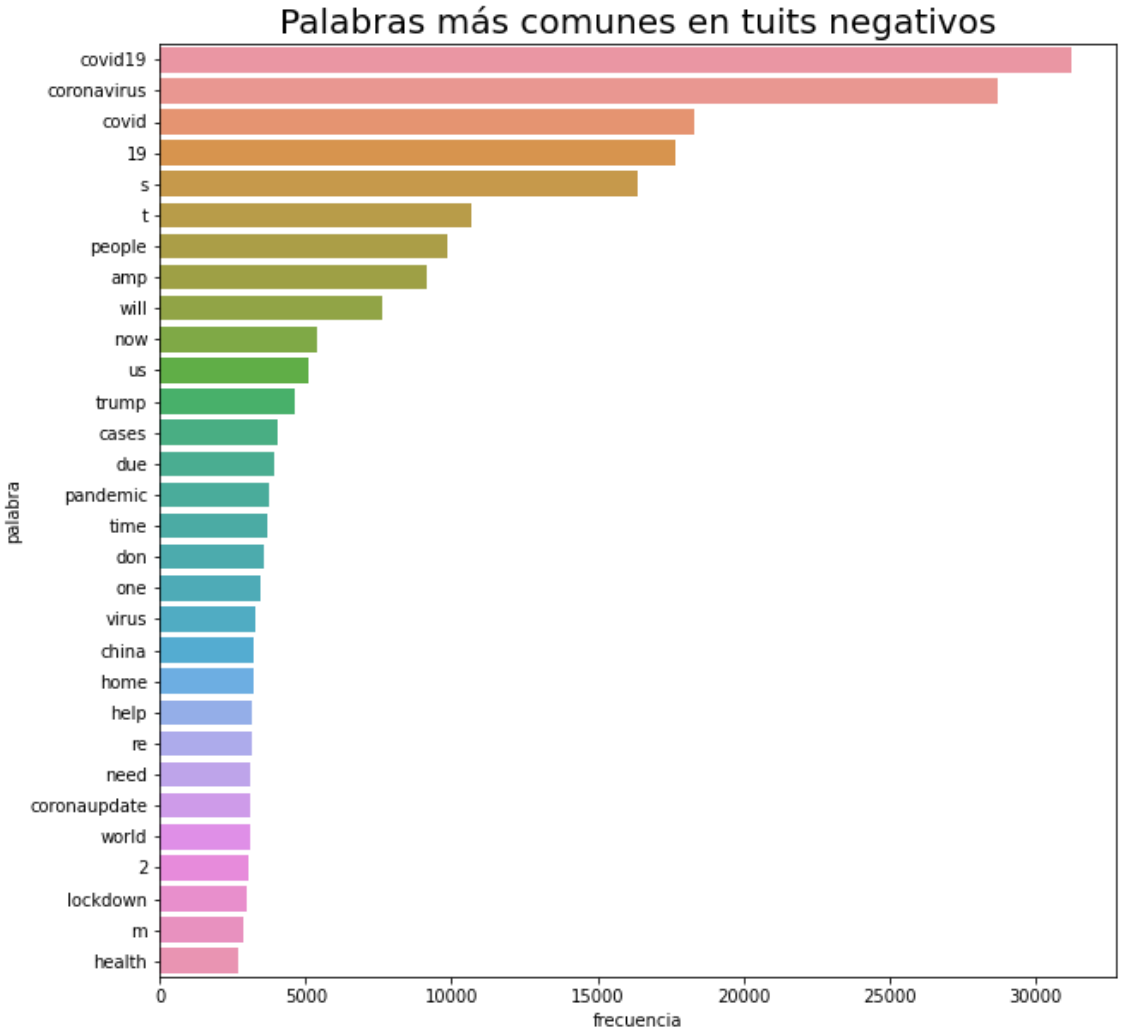


Figura 29: Palabras más usadas en tutis negativos

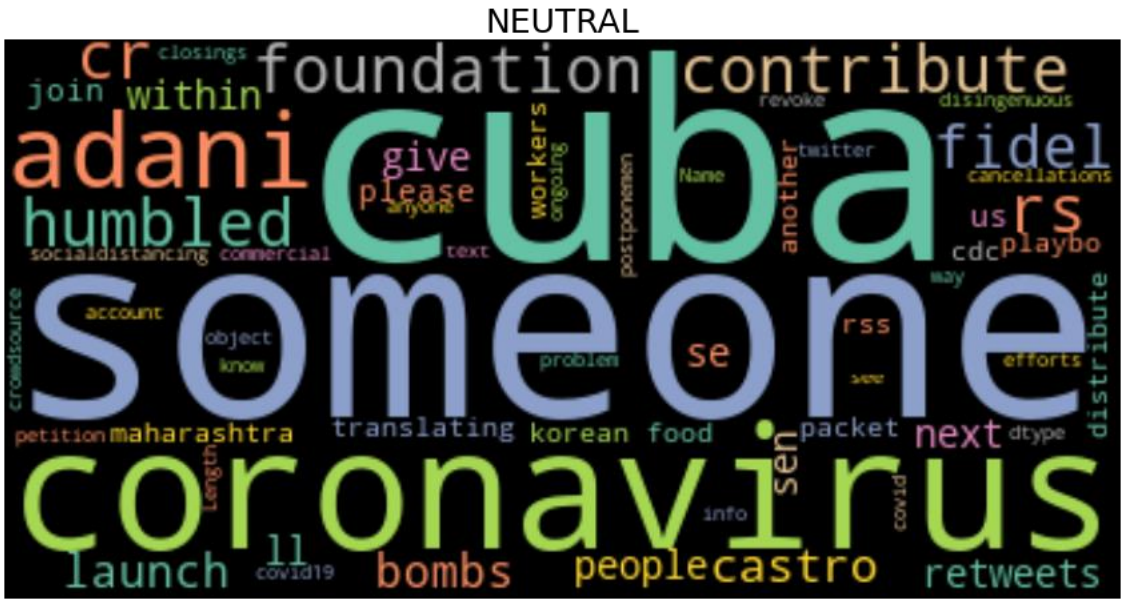


Figura 30: Wordcloud con las palabras más usadas en los tuits clasificados como neutrales

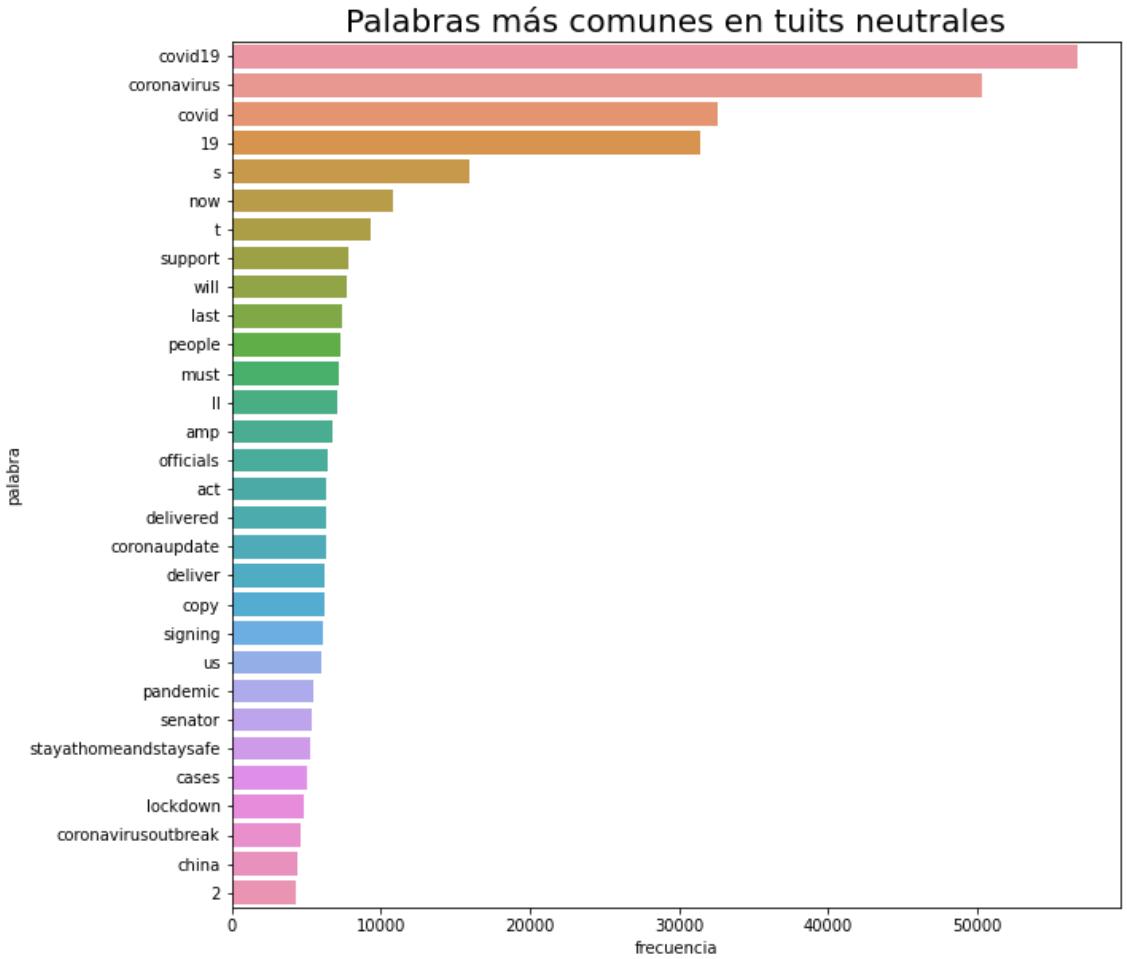


Figura 31: Palabras más usadas en tuits neutrales

Si nos fijamos primero en los *Wordcloud*, es interesante observar que en los tres tipos de tuit aparecen palabras que podríamos clasificar como negativas. Así pues, vemos que aparecen la palabra «*killer*» en los positivos y «*bombs*» en los neutrales. Esto se debe a que estas imágenes no muestran una clasificación de las palabras en sí, sino de las palabras que aparecen en los tuits marcados con ese sentimiento. Además, como ya hemos dicho cuando hemos hablado de *Wordcloud* por primera vez, el algoritmo no es fiable en cuanto a la frecuencia de las palabras.

Si nos centramos ahora en los gráficos que sí son fiables de las figuras 27, 29 y 31 vemos que la mayoría de las palabras que aparecen se repiten en los tres sentimientos y no aportan demasiada información sobre la clasificación. Aun así, sí hay algunas como *good* y *positive* que aparecen solo en los positivos o *Trump* que solo aparece en los negativos.

Otra medición interesante y que no aparece en el análisis de Kartik Mohan que hemos estado utilizando como guía es la ratio de tuits marcados como favoritos dependiendo del sentimiento del tuit. Para encontrar este dato, hemos calculado cuál es la suma total de favoritos para cada sentimiento y la hemos dividido entre el número total de tuits que se corresponde con ellos. El resultado es que hay un número considerablemente superior de tuits marcados como favoritos entre los negativos (16656.5) respecto a los neutrales (13887.9) y positivos (13718). Este resultado se mantiene en mayor o menor medida durante todos los días analizados. Lo podemos ver mejor en un gráfico del día 29 de marzo [figura 32].

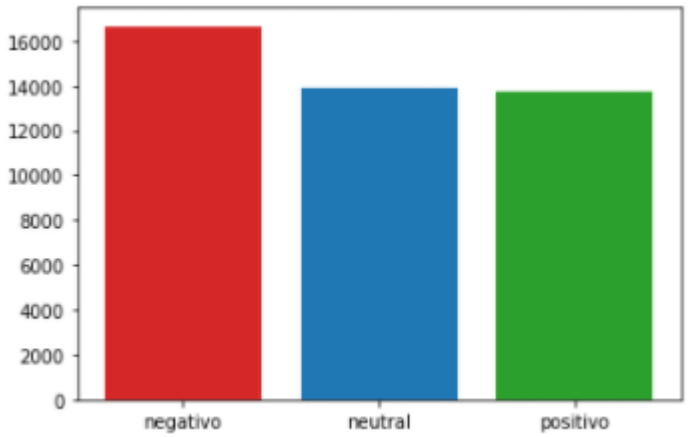


Figura 32: Ratio de favoritos para cada sentimiento

Con esto terminamos el primer análisis. De este estudio podemos concluir que la respuesta social de la población que usó Twitter durante el mes de abril era mayoritariamente neutral y positiva. Aun así, cabe destacar que los tuits que transmitían sentimientos negativos recibían un mayor número de favoritos por tuit de media. Este inesperado resultado puede dar paso a otro análisis en el que determinar si esta es una conducta generalizada en la plataforma o si solamente está presente en cierto tipo de temas.

Por otro lado, y antes de seguir con el segundo análisis, hemos decidido realizar este mismo análisis, pero utilizando los tuits de países específicos para encontrar diferencias entre ellos. Para hacer esto hemos usado el atributo "*country_code*" de los tuits. Este atributo es uno de los que más veces aparece vacío en los tuits recopilados, por lo que al aplicar el filtrado la cantidad de tuits de la que disponemos disminuye considerablemente. Es por esto por lo que en esta sección

solo compararemos los países de Estados Unidos, Reino Unido y India, ya que en estos casos disponemos de un número de tuits bastante elevado tras el filtrado.

Seguidamente veremos en qué aspectos del análisis varían estos países respecto al análisis general o entre ellos. Aquí, por lo tanto, no aparecerán todas y cada una de las mediciones, sino que solamente aquellas que tienen alguna relevancia por su variación. El resto de los resultados aparecerán en su propia sección del Anexo A.

Tras analizar los tres países, las únicas diferencias destacables respecto al análisis general aparecen en la medición de la ratio de favoritos por sentimiento. Antes hablábamos de que en general los tuits negativos obtenían una mayor cantidad de favoritos por tuit, no obstante, al analizar estos países de forma individual esta medición varía considerablemente dependiendo del día escogido. A continuación, veremos algunos ejemplos de este fenómeno.

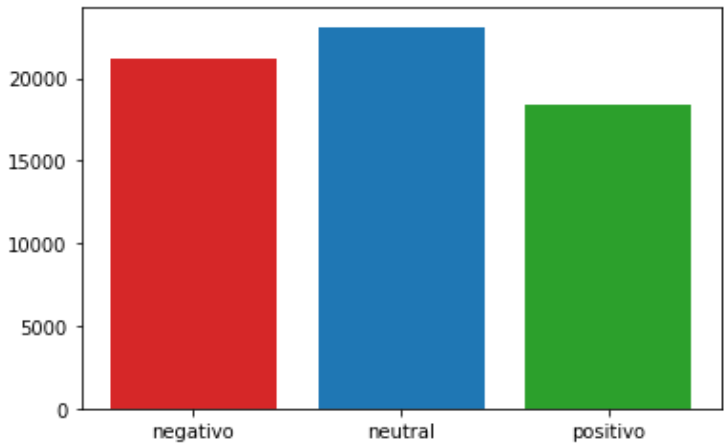


Figura 33: Ratio de favoritos 5 de abril Estados Unidos

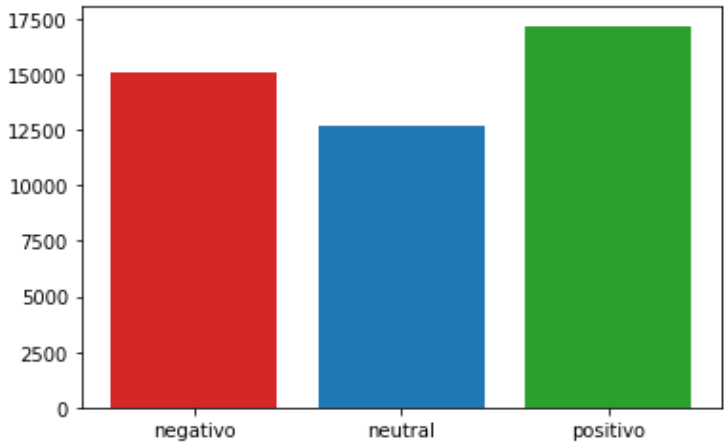


Figura 34: Ratio de favoritos 12 de abril Reino Unido

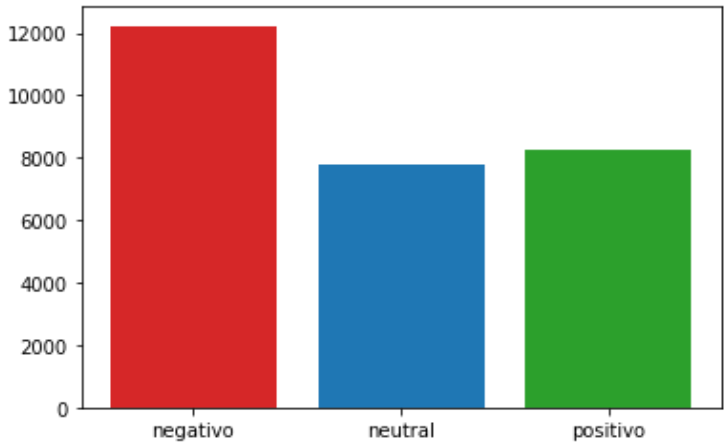


Figura 35: Ratio de favoritos 12 de abril India

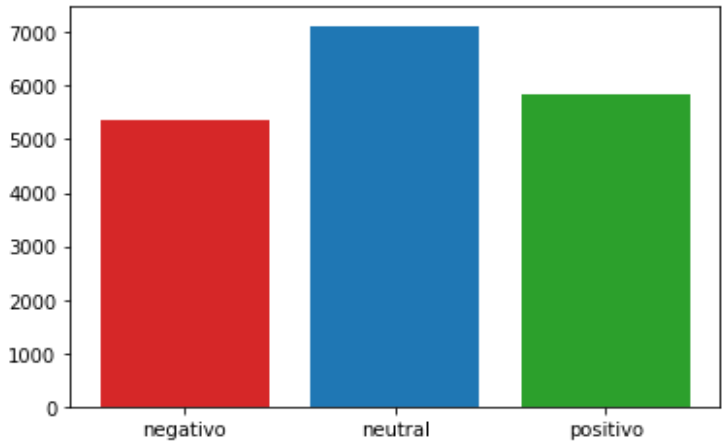


Figura 36:Ratio de favoritos 26 de abril India

En estas imágenes vemos como en algunos días los tuits negativos dejan de ser los que más favoritos tienen mientras que en otros el resultado que veíamos en el análisis general se acentúa aún más. En cualquier caso, no sería prudente darles la misma validez a estos resultados que a los extraídos del análisis general, ya que esta muestra es mucho más pequeña (entre 1000 y 5000 tuits comparado a cientos de miles en el análisis general) y por lo tanto mucho más susceptible a variaciones, pero no deja de ser interesante ver estos cambios.

Relación entre tuits y casos

Este segundo análisis será en el que busquemos una relación entre los tuits y los casos reales de COVID-19. Concretamente, intentaremos concretar si aumentan los tuits sobre el virus conforme aumentan los casos.

Para empezar, el primer paso ha sido juntar todos los archivos de los tuits, que, como hemos dicho antes, estaban divididos por días. Para ello, hemos empezado a trabajar con el cuaderno de Jupyter disponible gracias a la distribución Anaconda. Aquí, con una sencilla función hemos concatenado las filas de cada uno de los archivos en uno solo. Este proceso fue uno de los más

costosos, pues conllevó muchas horas el completarlo. Esta operación se hubiese podido llevar a cabo igual o más fácilmente usando un cuaderno de Google Colab, pero en ese momento del proyecto aún no se había considerado esa posibilidad.

Una vez conseguido el archivo que contiene todos los tuits, lo que vamos a hacer es dividirlos por países, para así poder extraer mejores conclusiones. Al empezar esta tarea apareció el primer y único problema que hubo con Google Colab.

El archivo conjunto que habíamos creado que contenía todos los tuits no aparecía en formato .csv como todos los demás, lo hacía como formato "Archivo" genérico. Esto hacía imposible subirlo a Colab para trabajar con él. Así pues, se decidió completar este proceso con el cuaderno local de Jupyter que, por el hecho de ser local, no tenía problemas para obtener este archivo.

Una vez abierto y leído el archivo con el cuaderno de Jupyter, el proceso para el filtrado no tiene mucha complicación. Cómo la tabla de tuits tiene un atributo "country_code", simplemente cogemos los tuits con el código que deseamos. Un detalle que es importante mencionar sobre esto es que, como ya habíamos mencionado al final del análisis anterior, solo un segmento muy pequeño de los tuits dispone de este código de país, por lo tanto, la información de la que disponemos no es del todo representativa de todos los tuits. Para hacerse una idea, el archivo del primer día, el 31 de marzo de 2020, contiene un total de 564.141 tuits. Pues solamente 30.526 de estos disponen de algún tipo de código de país. Esto ha supuesto que no hayamos podido analizar la mayoría de los países por falta de datos.

Después de disponer del archivo con los tuits divididos por países, solamente necesitamos los datos de los contagios de esos países para hacer el análisis. Como ya contamos con un archivo con los datos de todos los países por fecha, y como este archivo sí lo puede leer Google Colab, volvemos a trabajar con esta plataforma.

Una vez leído este archivo, lo primero que hacemos es limpiar los datos. Primero eliminamos las columnas que no vamos a utilizar, como los casos de recuperados, los activos o la región de la OMS y después los filtramos para quedarnos solamente con los datos del país que queremos analizar y durante las fechas que coinciden con los datos de tuits. Así, por ejemplo, con los datos de España nos quedaría esta tabla [figura 37]:

	Date	Country	Confirmed	Deaths
12686	2020-03-29	Spain	80110	6803
12873	2020-03-30	Spain	87956	7716
13060	2020-03-31	Spain	95923	8464
13247	2020-04-01	Spain	104118	9387
13434	2020-04-02	Spain	112065	10348

Figura 37: Ejemplo de formato de datos de contagios

Con esto ya podemos dibujar el primero de los gráficos, uno que muestre la evolución de los contagios durante estas fechas [figura 38]

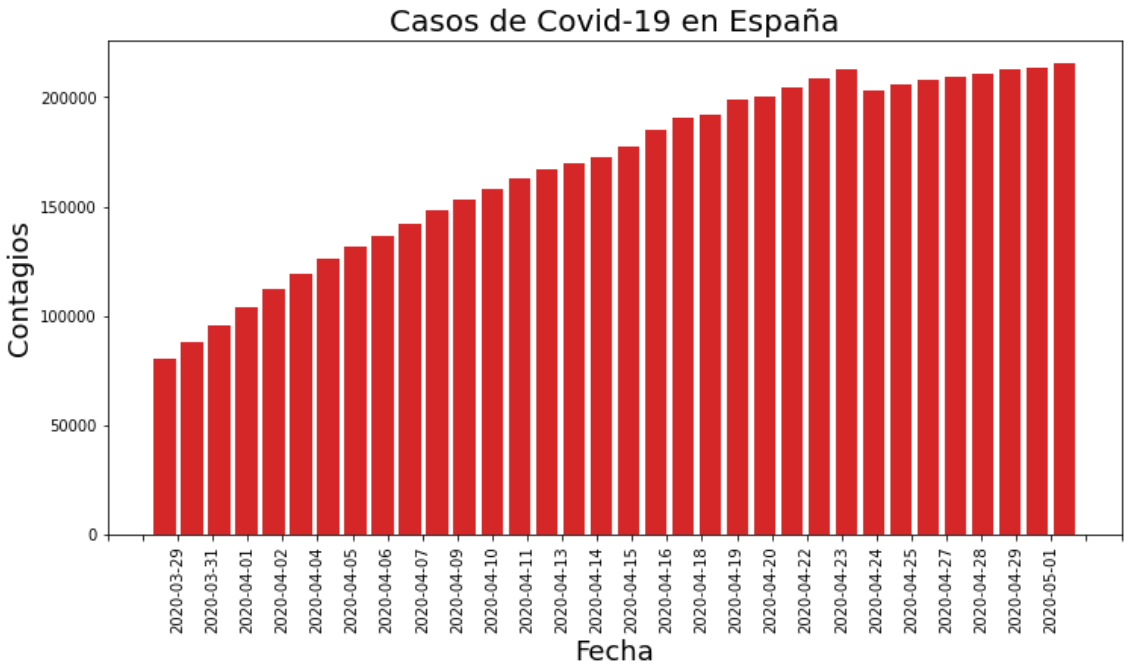


Figura 38: Casos de COVID-19 en España

Una vez obtenido esto, solo nos falta saber los tuits que se enviaban durante el mismo periodo de tiempo para poder hacer la comparación. Para averiguarlo, el primer paso ha sido leer el archivo de los tuits. De España, en este caso. Después hemos cambiado el formato del atributo "created_at" para convertirlo en una fecha sin hora y para normalizar el formato a mes/día/año. Este paso ha sido necesario porque había una parte de los tuits con este formato y otra con el formato día/mes/año. Por último, y como los tuits ya abarcan el mismo periodo de tiempo que los contagios, ya podemos dibujar los otros gráficos. Primero el que muestra solamente el número de tuits [figura 39] y finalmente el que muestra la comparativa entre el número de tuits y los casos confirmados de COVID-19 [figura 40].

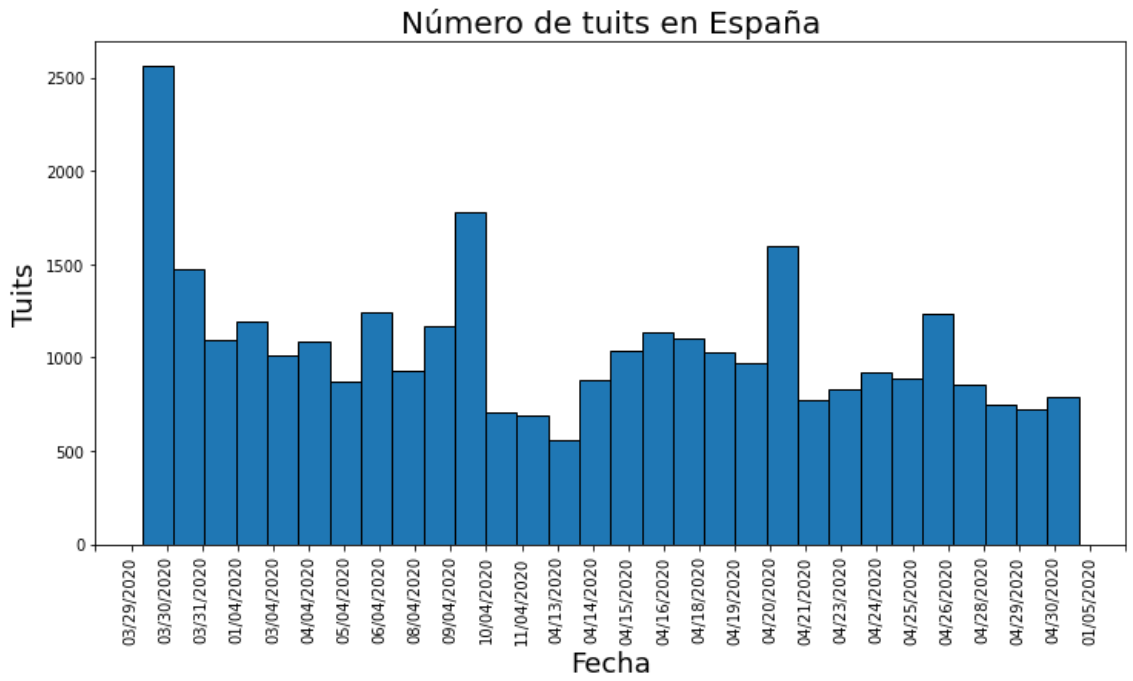


Figura 39: Número de tuits con menciones a la COVID-19 en España

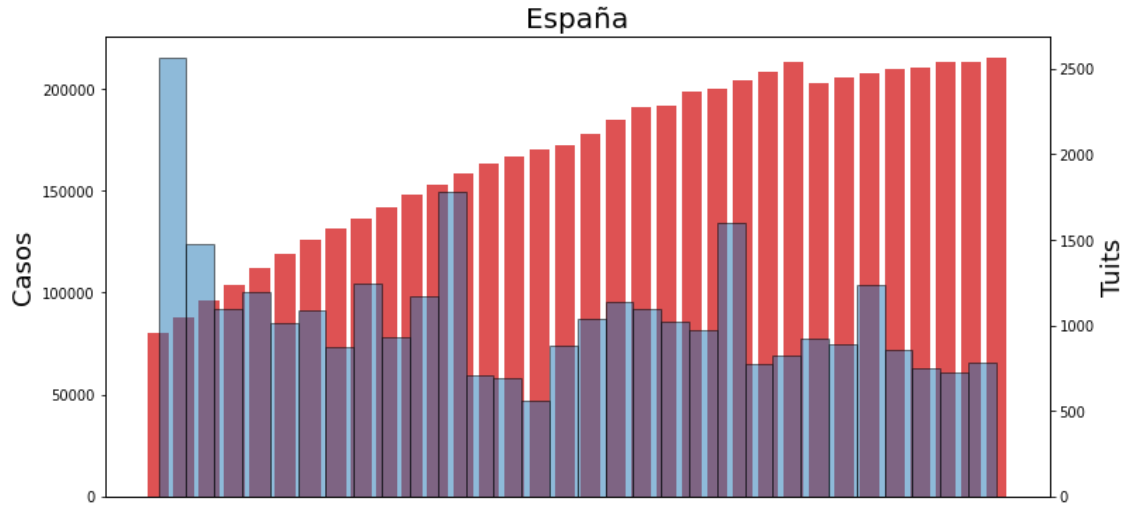


Figura 40: Comparativa entre casos de COVID-19 y nº de tuits con menciones al virus en España

Gracias a este último gráfico ya podemos sacar conclusiones. Podemos ver claramente que, pese a que los casos de contagios siguen subiendo, los tuits empiezan a descender poco a poco. Este efecto no es algo propio solamente del caso de España, sino que se repite, en mayor o menor medida, en todos los países analizados. Así lo podemos ver en las siguientes imágenes.

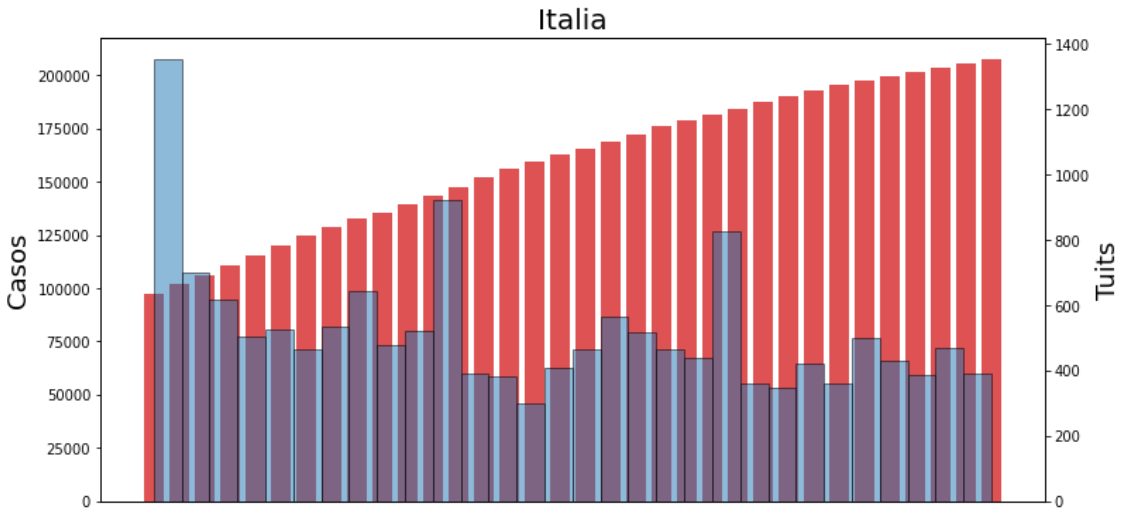


Figura 41: Comparativa de Italia

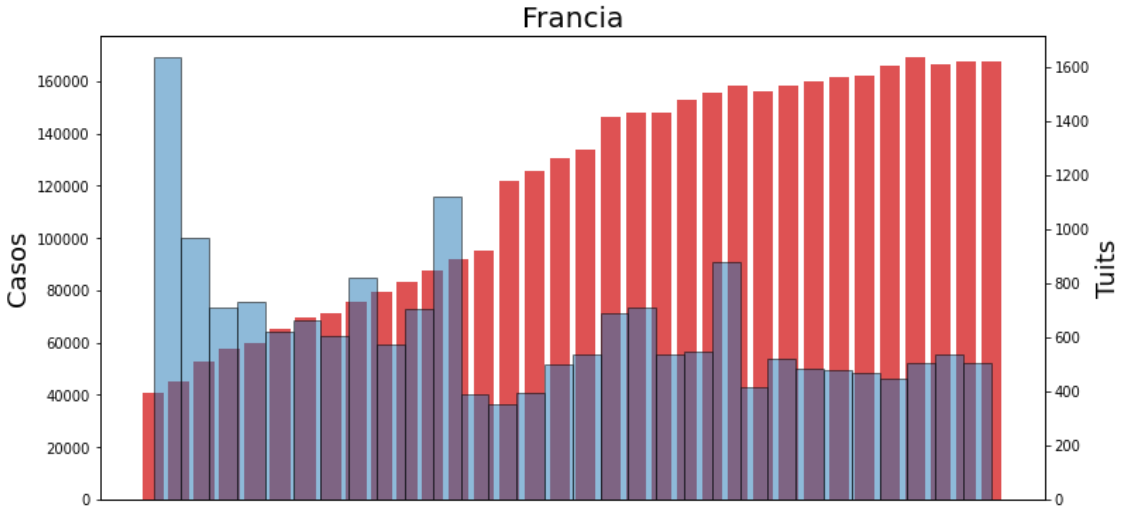


Figura 42: Comparativa de Francia

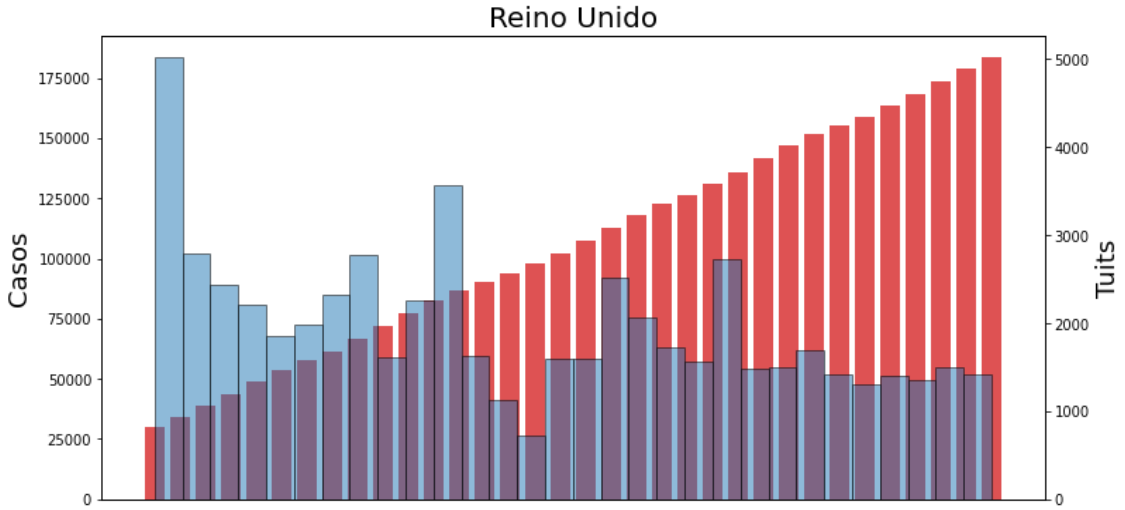


Figura 43: Comparativa de Reino Unido

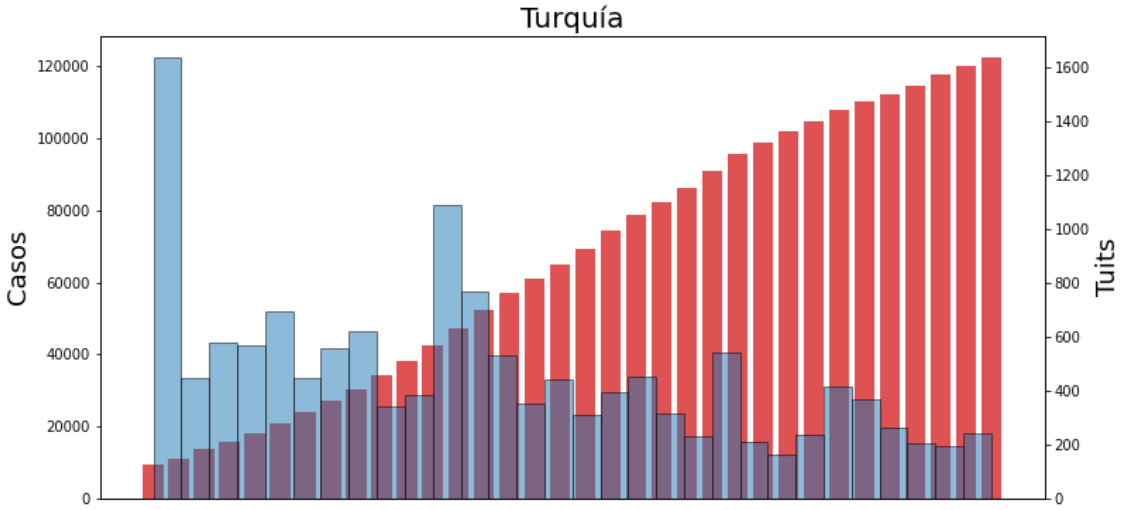


Figura 44: Comparativa de Turquía

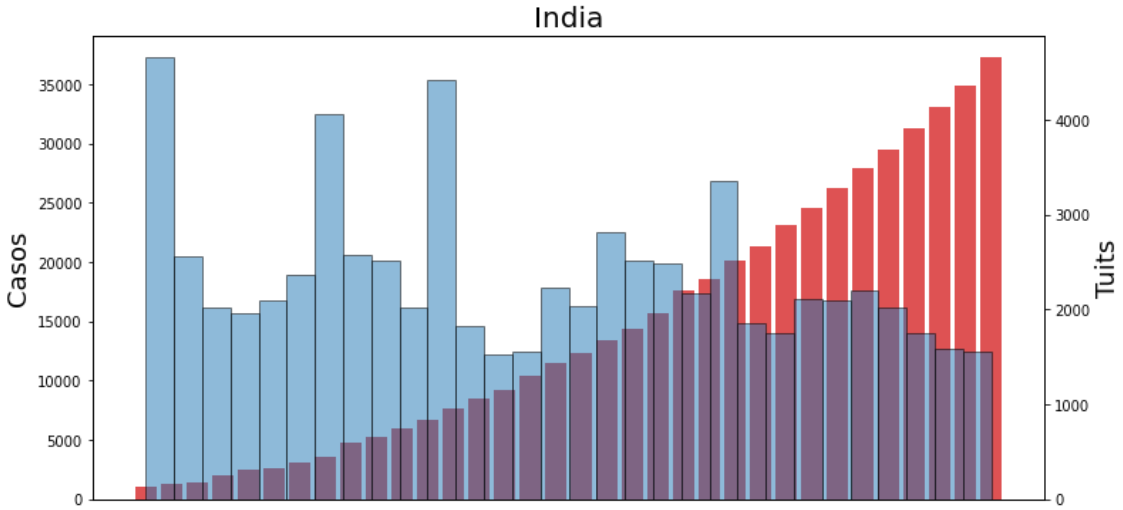


Figura 45: Comparativa de India

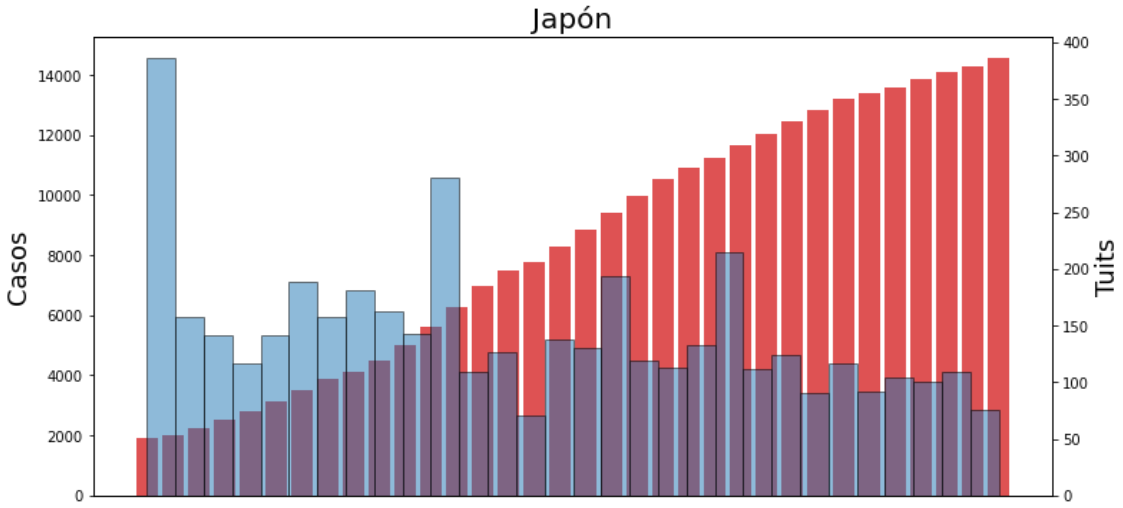


Figura 46: Comparativa de Japón

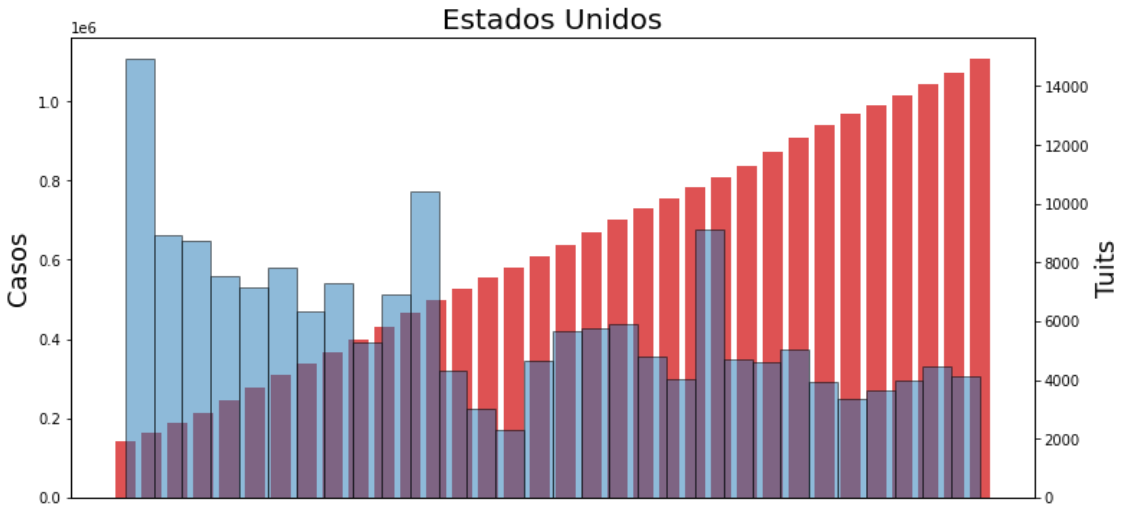


Figura 47: Comparativa de Estados Unidos

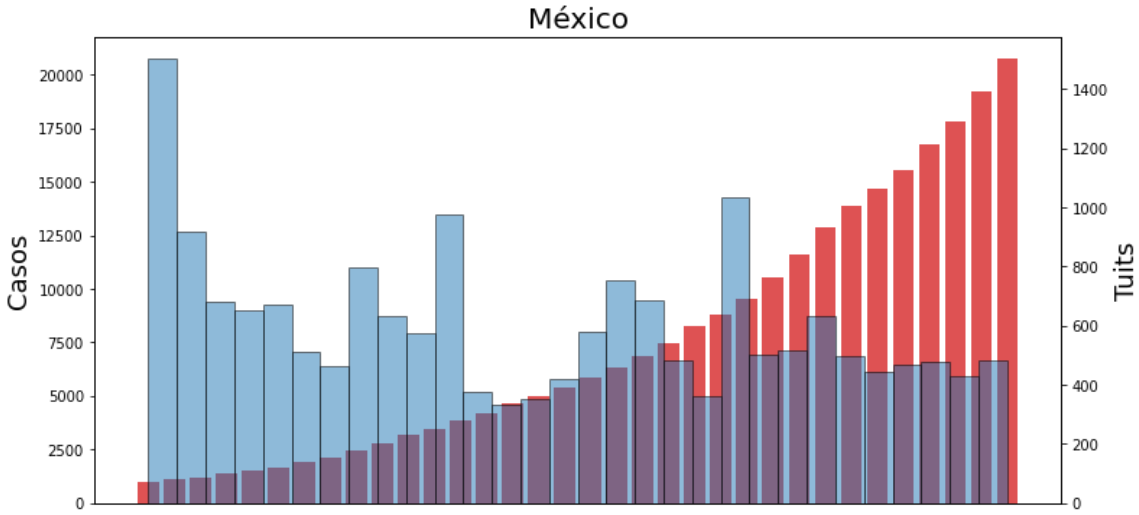


Figura 48: Comparativa de México

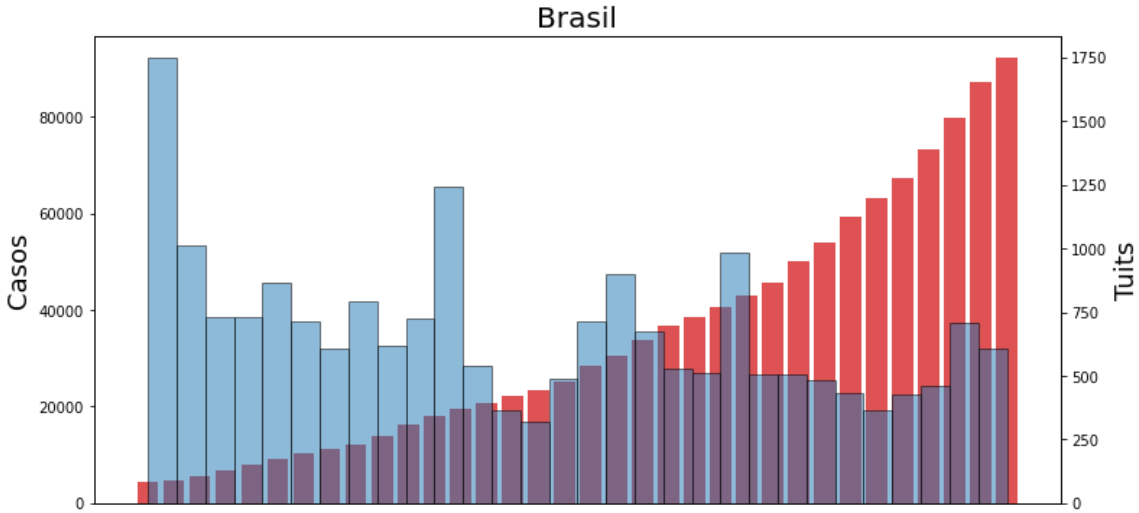


Figura 49: Comparativa de Brasil

Si observamos estos gráficos de distintos países vemos que los resultados son muy similares, sobre todo en cuanto a la progresión en el número de tuits. Este efecto es más pronunciado en algunos países como Estados Unidos [figura 47] y casi inexistente en India [figura 45] donde se mantiene un número muy elevado de tuits de forma constante.

Resulta realmente curiosa la similitud en el número de tuits en algunos países, sobre todo el pico inicial en el gráfico y, personalmente, no opinamos que se deba a una coincidencia. Podría deberse a algún error en la recogida de los tuits por el usuario que los publicó o a algún efecto producido al hacer los gráficos, pero es imposible saberlo a ciencia cierta.

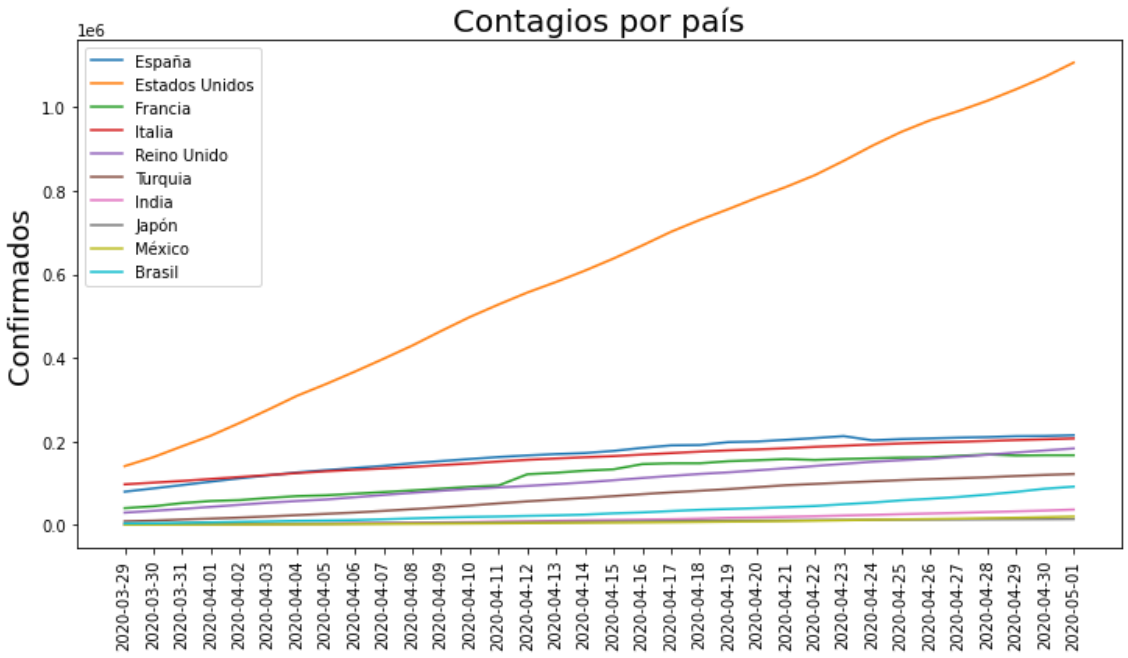


Figura 50: Comparativa de casos de COVID-19 por países

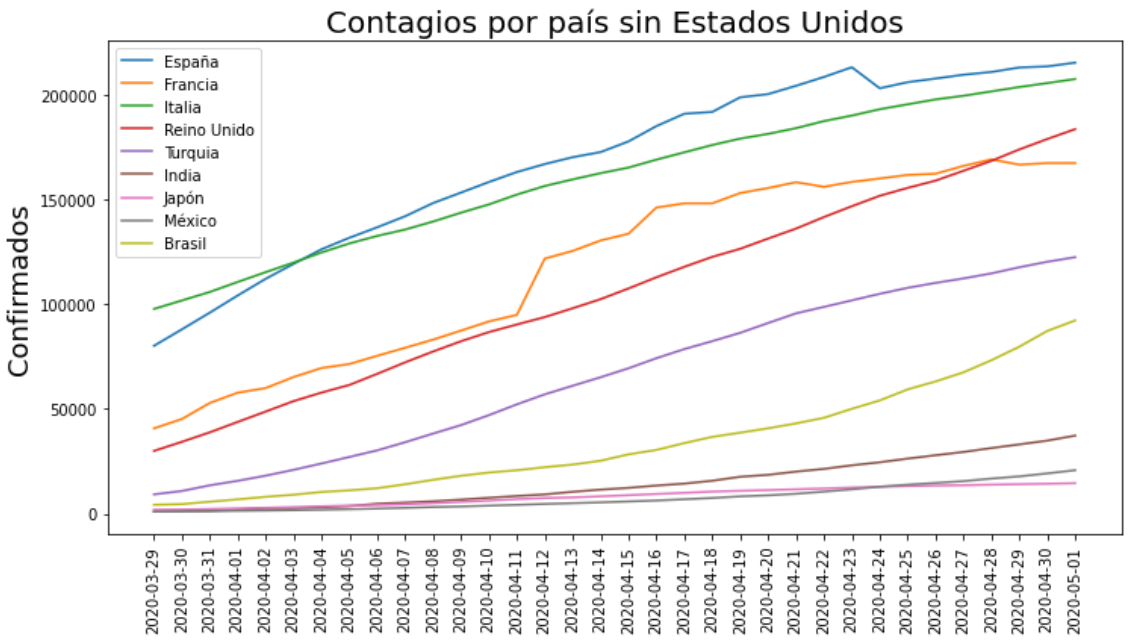


Figura 51: Comparativa de casos de COVID-19 por países sin Estados Unidos

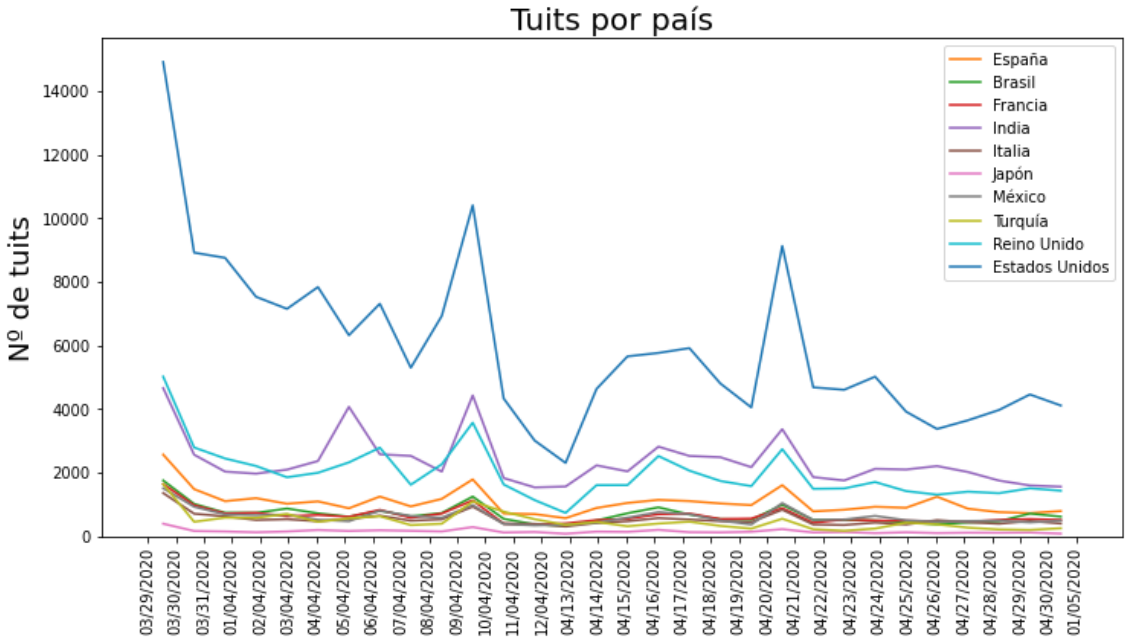


Figura 52: Comparativa de nº de tuits con menciones a la COVID-19 por países

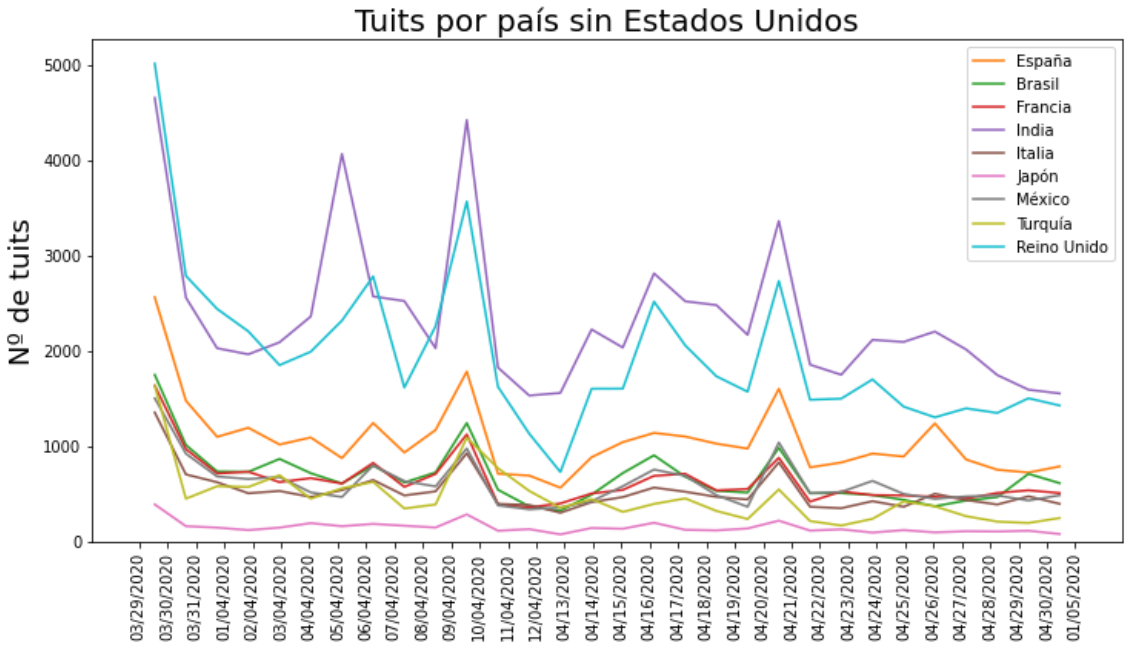


Figura 53: Comparativa de nº de tuits con menciones a la COVID-19 por países sin Estados Unidos

Por último, hemos decidido incluir con unos gráficos de líneas la comparación directa entre los países analizados de contagios y de número de tuits con menciones a la COVID-19. Se ha incluido una versión sin la aparición de Estados Unidos porque, al ser sus cifras mucho mayores a la mayoría de otros países, no deja ver bien como se comparan los demás entre sí. En las figuras 50 y 51 podemos ver que algunos países como España, Italia, Francia y Japón empezaban a mostrar mejoras en sus curvas de contagios mientras que otros como Brasil, que aún estaban en las primeras fases de la pandemia, mostraban una tendencia muy alarmante.

En las figuras 52 y 53 vemos aún mejor las similitudes de las que hablábamos antes en cuanto a la progresión en el número de tuits. Dejando de lado las cifras totales, que son más elevadas en algunos países que en otros, en general en todos los países descienden ligeramente y todos presentan variaciones en los mismos puntos.

Con esto terminamos este segundo análisis. La conclusión a la que podemos llegar con este estudio es que, independientemente del país analizado, los tuits sobre la pandemia y el virus han ido decreciendo desde el principio de esta (al menos respecto al periodo de tiempo analizado).

5. Conclusiones

Una vez completado el estudio, vamos a exponer en este capítulo las conclusiones a las que se ha llegado. Empezaremos resumiendo los análisis y contrastando sus resultados con los objetivos planteados al inicio del proyecto. Después, haremos una pequeña reflexión sobre el trabajo en sí y de cómo ha sido su desarrollo.

Primero, vamos a resumir y examinar los resultados y a decidir si estos son suficientes para cumplir con los objetivos establecidos al iniciar el trabajo. Empezando por el primer análisis, nuestro objetivo principal era observar la respuesta social de los usuarios de Twitter durante el transcurso del mes de abril. Para ello, hemos utilizado los tuits enviados entre el 29 de marzo de 2020 y el 1 de mayo de 2020 y se ha realizado un estudio de su contenido. Nos hemos centrado en los tuits escritos en inglés, ya que de estos podemos extraer más información.

Hemos comenzado viendo qué palabras han sido las más utilizadas. En este caso, no ha habido grandes sorpresas, ya que, dejando aparte las palabras usadas para la extracción de los tuits (coronavirus, pandemia, etc), los otros términos que más aparecen son palabras muy relacionadas, como *people, cases, home, world…*

Después, hemos clasificado los tuits por los sentimientos que transmiten, viendo así que la mayoría de ellos eran neutrales o ligeramente positivos. Este resultado se mantenía contante a lo largo de los días analizados. De este hecho podemos concluir que la respuesta social durante el periodo analizado se ha mantenido ligeramente positivo.

A continuación, hemos plasmado con *wordclouds* las palabras que aparecían en cada tipo de tuit. En este paso nos hemos dado cuenta de que había tuits que eran clasificados como positivos o neutrales pese a contener palabras claramente negativas.

Finalmente, hemos visto que la ratio de favoritos de los tuits negativos era constantemente superior al de los positivos y neutrales. De esto podemos deducir que, pese a que hay más personas que escriben tuits positivos, los negativos suelen gustar más o ser más apoyados.

El hecho de que los resultados no varíen significativamente a lo largo del mes, sumado al último dato analizado, hace que terminemos el primer análisis con una sensación agridulce. Aun así, podemos afirmar que hemos cumplido con el primer objetivo establecido.

En cuanto al segundo estudio, el objetivo principal era encontrar (si es que existía) la relación entre el número de tuits escritos relacionados con el virus y la pandemia y los datos reales de esta.

Para llevar a cabo esta tarea, hemos utilizado los datos de tuits en su conjunto y datos de contagios de todo el mundo.

Primero hemos dividido los tuits por países y después los hemos comparado con los datos de contagiados. El resultado ha sido que, pese a que los contagios continuaban subiendo en todos los países, los tuits descendían. Este fenómeno se repetía en todos los países analizados. Este resultado nos lleva a la conclusión de que la mayor o menor presencia de cierto tema en las redes sociales no es un fiel indicador de si afecta a más o menos personas y, por lo tanto, las redes sociales no son un reflejo fiel de la realidad.

Para terminar, vamos a explicar cómo ha sido la experiencia de completar este trabajo y qué conclusiones personales podemos extraer de él.

Realizar un proyecto como este ha sido todo un reto, tanto por el proceso de investigación y recopilación de datos, como por el hecho de que nunca había usado Python ni ninguna de las plataformas con las que he trabajado. No obstante, todo esto lo ha convertido en un proceso muy interesante, ya que cada día he aprendido algo nuevo. Uno de los elementos que más impulsó el avance en el trabajo y el desarrollo del mismo fue el cambio de plataforma. Dejar de usar un cuaderno local de Jupyter y pasar a usar Google Colab supuso una mejora extraordinaria en cuanto a comodidad y velocidad.

En conclusión, trabajar en este proyecto ha supuesto un inmenso ejercicio de aprendizaje en el que se han utilizado gran parte de los conocimientos y capacidades aprendidas durante el grado. Ha sido provechoso ver reflejadas en un mismo trabajo las nociones que se han ido adquiriendo durante los años y poder completar un trabajo de este tipo. Además, también se ha experimentado con herramientas con las que no se había tratado aún, lo que ha aportado un mejor dominio en este campo.

6. Ampliaciones y trabajos futuros

En este último capítulo del trabajo analizaremos de qué formas podría ampliarse y cuáles podrían ser algunos otros estudios relacionados o que lo complementen de alguna forma.

Para empezar, la ampliación más evidente de este trabajo sería seguir recopilando tuits hasta que terminase la pandemia o, al menos, de principio a fin de la primera ola. Para conseguir esto habría que extraer los tuits utilizando la API de Twitter o usar algún programa como Hydrator[7] y conseguirlos usando los identificadores de los tuits, a los que sí se puede acceder. Esto, como ya hemos explicado antes, es necesario desde hace relativamente poco, porque Twitter cambió su política para aumentar la privacidad de sus usuarios, haciendo imposible compartir el contenido de los tuits directamente del modo en que lo hicimos para este trabajo.

Esto nos permitiría ver una evolución mucho mayor desde que aparecían las primeras menciones del virus hasta el final de la primera ola, cuando los contagios bajaban considerablemente.

Otra ampliación interesante sería realizar un análisis de sentimientos en otros idiomas, principalmente en español. Para hacer esto y utilizando el mismo método usado por *TextBlob* para el algoritmo que hemos usado en inglés, podríamos utilizar reseñas de películas, restaurantes u hoteles para entrenar un algoritmo que clasificara las palabras entre negativas, neutrales o positivas.

Otro proyecto relacionado que nos podría ayudar a entender mejor algunos de los resultados obtenidos, sería uno que explorara por qué los tuits negativos obtuvieron un mayor número de favoritos. Sería interesante ver si esto es un hecho aislado o algo recurrente en Twitter o, incluso, en otras redes sociales o comunidades en internet.

Por último, sería interesante estudiar cuanto tiempo suele ser un tema relevante en las redes sociales, en especial Twitter y como puede haber afectado esto al descenso en el número de tuits relacionados con la COVID-19.

[7] https://github.com/DocNow/hydrator

Bibliografía

[1] DW, "+++ Coronavirus, minuto a minuto: América y Europa bajo cuarentena y con fronteras cerradas +++ (17.03.2020) | El Mundo | DW | 16.03.2020," *DW*, 2020. https://www.dw.com/es/coronavirus-minuto-a-minuto-américa-y-europa-bajo-cuarentena-y-con-fronteras-cerradas-17032020/a-52802033 (accessed Jun. 17, 2020).

[2] TeleSUR, "Países de América Latina refuerzan seguridad ante el Covid-19 | Noticias | teleSUR," *teleSUR*, 2020. https://www.telesurtv.net/news/america-latina-medidas-seguridad-contencion-coronavirus-20200316-0001.html (accessed Jun. 17, 2020).

[3] J. Yeung, "Así responden los países del mundo al brote de coronavirus | CNN," *CNN*, 2020. https://cnnespanol.cnn.com/2020/03/04/asi-responden-los-paises-del-mundo-al-brote-de-coronavirus/ (accessed Jun. 17, 2020).

[4] A. R. Anggraini and J. Oliver, "La COVID-19 afecta significativamente a los servicios de salud relacionados con las enfermedades no transmisibles," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019, doi: 10.1017/CBO9781107415324.004.

[5] D. N. B. Lasa *et al.*, *LAS CONSECUENCIAS PSICOLÓGICAS DE LA COVID-19 Y EL CONFINAMIENTO*. 2020, p. 210.

[6] X. Boni, M.F., Lemey, P., Jiang, *Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic*. 2020, p. 93.

[7] Top Position, • *Estudio viralidad del CORONAVIRUS en las redes sociales*. 2020.

[8] J. M. Sánchez, "• El uso de redes sociales en España aumenta un 55% en la pandemia de coronavirus," *ABC*, 2020.

[9] Smartmeanalytics, "ESTUDIO DEL IMPACTO DEL CORONAVIRUS EN EL USO DEL MÓVIL," 2020.

[10] A. Vazquez Brust, "Ciencia de Datos para Gente Sociable," 2020. https://bitsandbricks.github.io/ciencia_de_datos_gente_sociable/que-es-la-ciencia-de-datos.html.

Anexos

Anexo A - Análisis de tuits

Cuadernos

Análisis general

1. 29 de marzo de 2020: https://colab.research.google.com/drive/1uT89W3m_WPbjC-JaC1zSzf6_ckmOxooF?usp=sharing
2. 5 de abril de 2020: https://colab.research.google.com/drive/1qI-c-M47-EVyl_lbMdglX3evEbmzTT6f?usp=sharing
3. 12 de abril de 2020: https://colab.research.google.com/drive/1PUZGh1jxyaJLCoHS4KFgyWHuinjXgQRY?usp=sharing
4. 19 de abril de 2020: https://colab.research.google.com/drive/1SYZKHZV_otkLX_81Bu_DjWPdx9j2VZSh?usp=sharing
5. 26 de abril de 2020: https://colab.research.google.com/drive/1GCn-yWxp-4iHEegJsq87DSUS3DXNf8ok?usp=sharing
6. 30 de abril de 2020: https://colab.research.google.com/drive/1xTA2CY16ghDED8j6sPNrMhmK3GAJ1SE9?usp=sharing

Análisis por países

Estados Unidos

1. 29 de marzo de 2020: https://colab.research.google.com/drive/196b4xyE5HCpN-iZJmil2UPtl5c_6sk1n?usp=sharing
2. 5 de abril de 2020: https://colab.research.google.com/drive/12Ha2YSFou8CVNujBIoy4N-nV3IYgred1?usp=sharing
3. 12 de abril de 2020: https://colab.research.google.com/drive/1zh2Iuq5DQh_KjULMnXTq8tqIqbDFZoH6?usp=sharing

4. 19 de abril de 2020: https://colab.research.google.com/drive/1SNF_F-LeOJv8FeNPHLa6niNMvsUg4pKW?usp=sharing
5. 26 de abril de 2020: https://colab.research.google.com/drive/1SbHBVWW3Qr8LliIm3INsUFhKIlHgbRdl?usp=sharing
6. 30 de abril de 2020: https://colab.research.google.com/drive/1QtpFjxKV829A1S5OEJIfM_tuNYU2yv8q?usp=sharing

Reino Unido

1. 29 de marzo de 2020: https://colab.research.google.com/drive/1ulfADzzIgO09qX1XEKQtWjKlu71QwEvv?usp=sharing
2. 5 de abril de 2020: https://colab.research.google.com/drive/1b7IVwLaHhZvHnqLoeltwORf9GHG_Koms?usp=sharing
3. 12 de abril de 2020: https://colab.research.google.com/drive/1YH3oG_PcuWg8oikRjDSJ9sJPy-DIORy5?usp=sharing
4. 19 de abril de 2020: https://colab.research.google.com/drive/12OHJqxDU7R2zMRidhT6LRyQVg6KMK9_j?usp=sharing
5. 26 de abril de 2020: https://colab.research.google.com/drive/14f-_8uMMEOzL1SdBS7xQA1Lji0d2L8lh?usp=sharing
6. 30 de abril de 2020: https://colab.research.google.com/drive/1Ltq3pvtVi9Q6Yi_RvfhZi5j3lv4V_lRF?usp=sharing

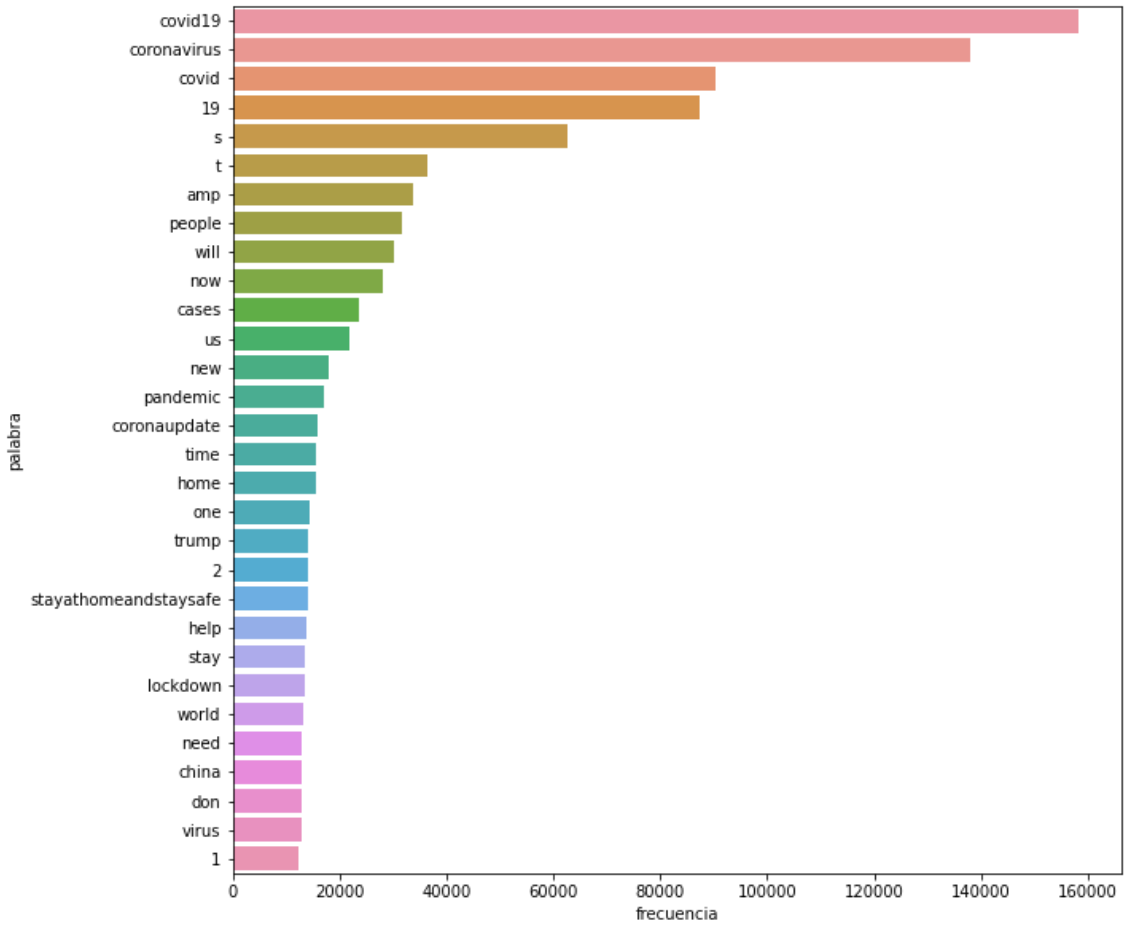
India

1. 29 de marzo de 2020: https://colab.research.google.com/drive/1CRVBRa50jvGMHERGe-HQqbm8cda-KBL5?usp=sharing
2. 5 de abril de 2020: https://colab.research.google.com/drive/1quEwiqmxrYDSgtrtKCGy2q8Eb3Lw3ylB?usp=sharing
3. 12 de abril de 2020: https://colab.research.google.com/drive/1YnqsV1jg6yxOmMo-6x4SM9_mx_9Yq4xd?usp=sharing
4. 19 de abril de 2020: https://colab.research.google.com/drive/1NA1bqOqYnlkXo0ycaL0uq2Dfkx7nRxsZ?usp=sharing

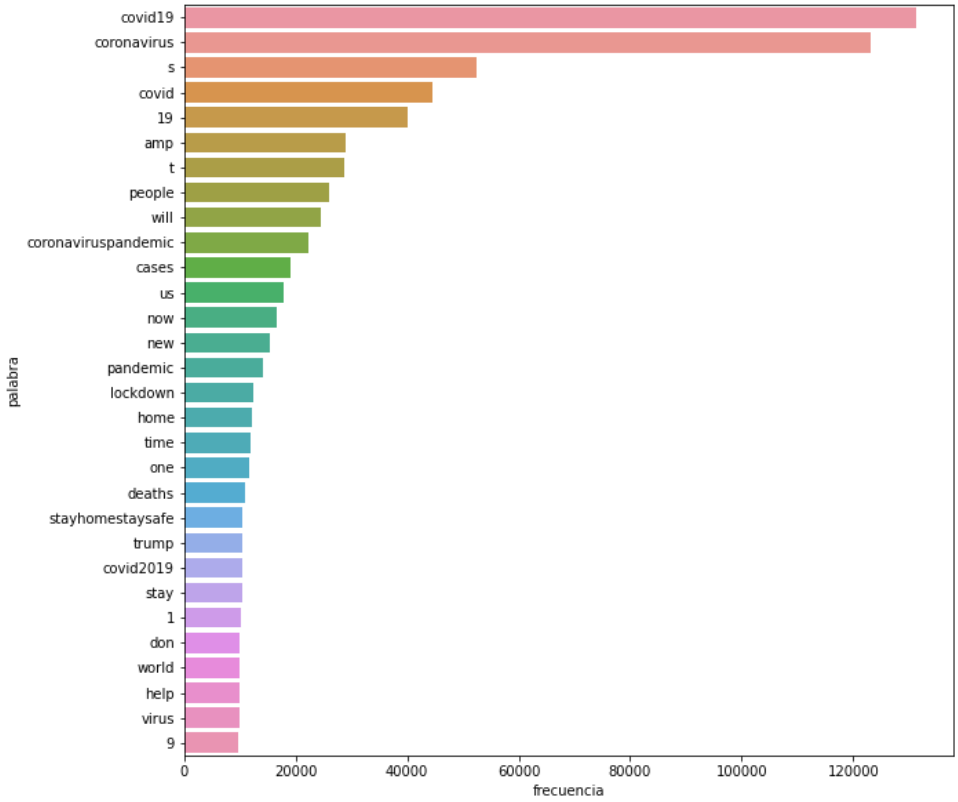
5. 26 de abril de 2020: https://colab.research.google.com/drive/1fvHjvHqxRRJdDrj_GijCr5YxqO4Vtsb2?usp=sharing
6. 30 de abril de 2020: https://colab.research.google.com/drive/1yGBWOpf-0X235EOJoXa4EgrW1OvnDccF?usp=sharing

Resultados del análisis general

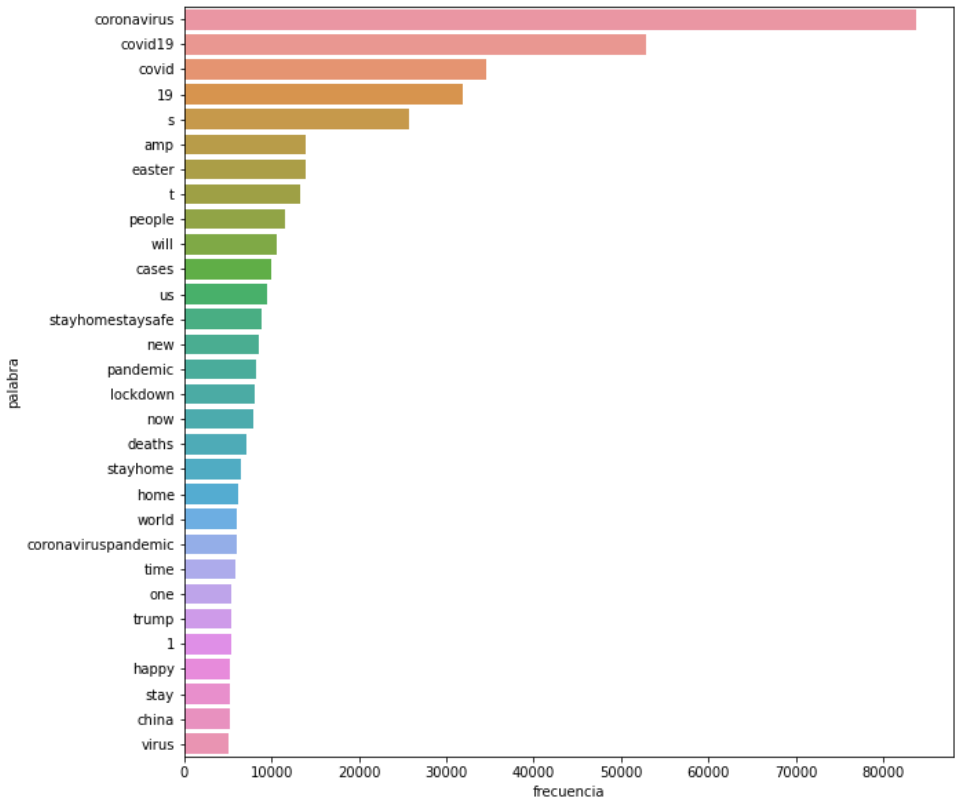
1. Palabras más usadas en los tuits con menciones



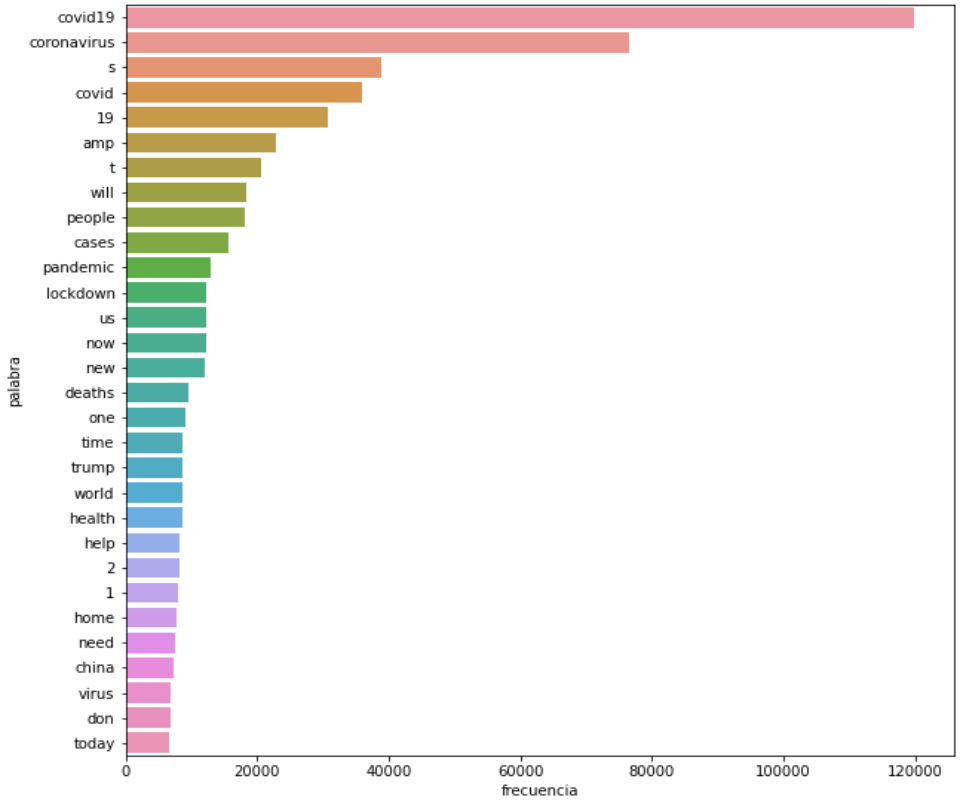
29 de marzo



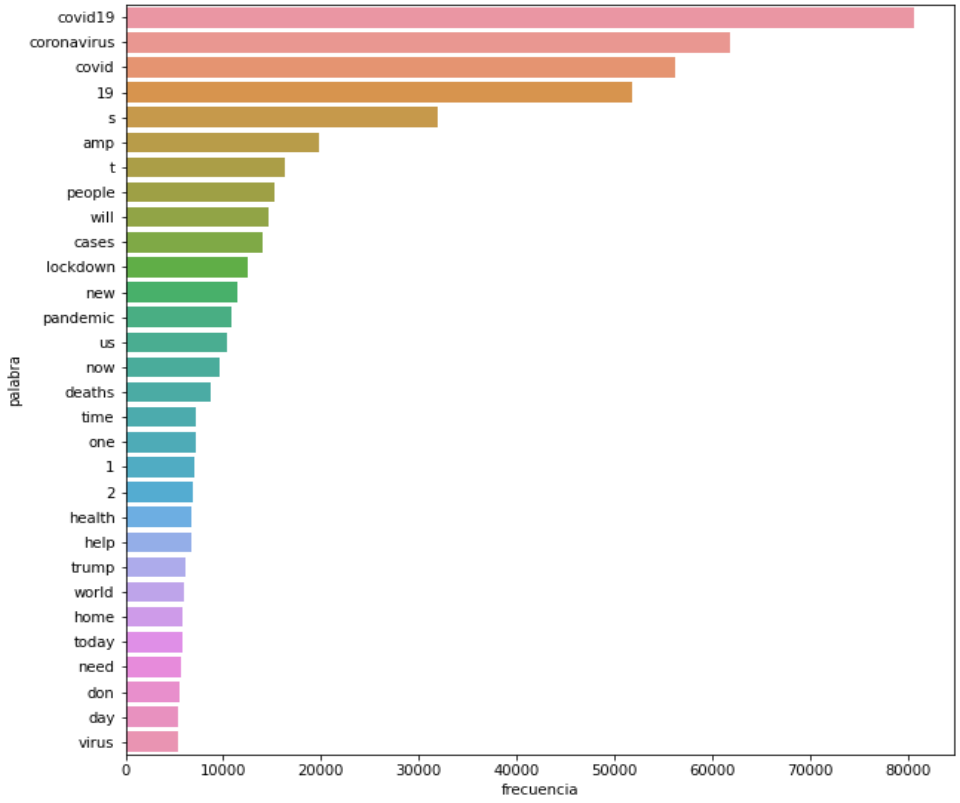
5 de abril



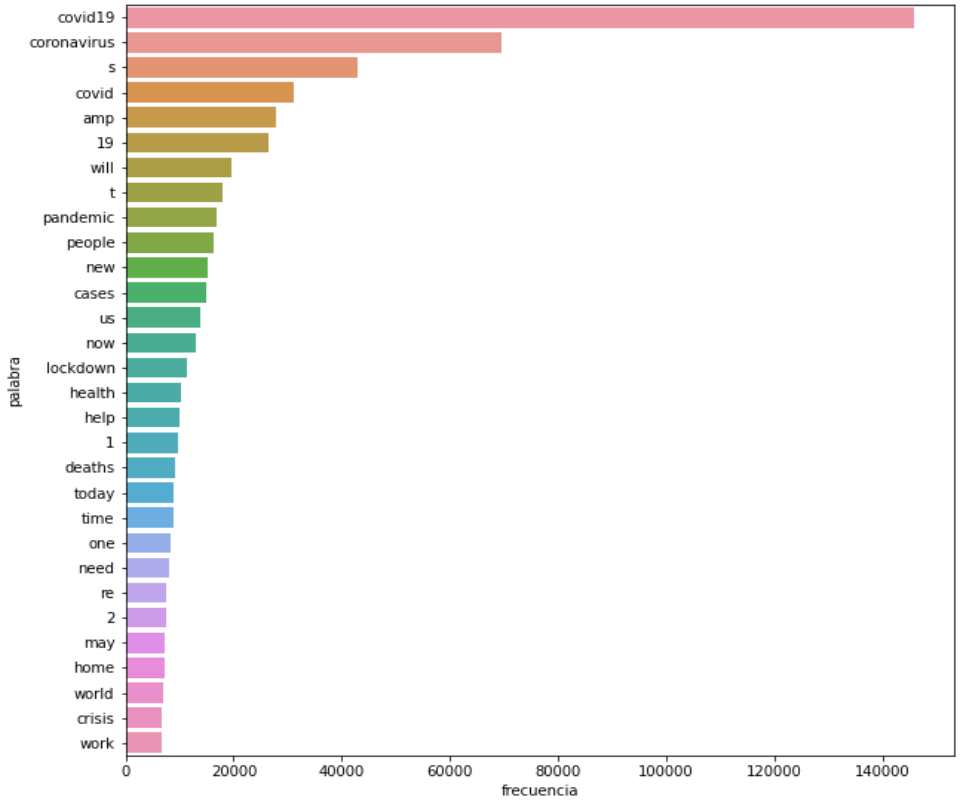
12 de abril



19 de abril

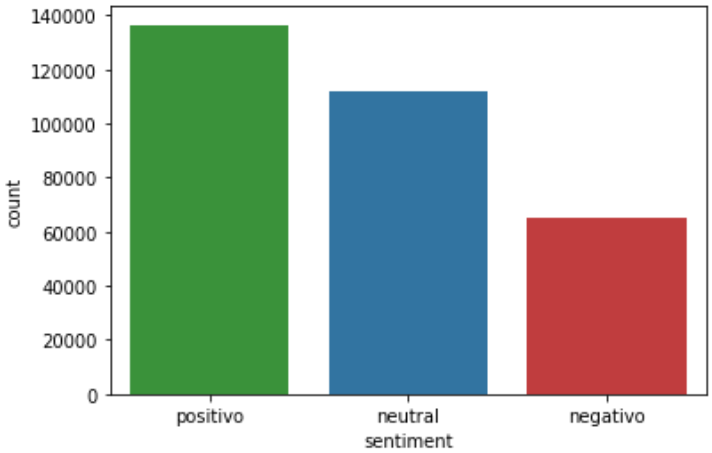


26 de abril

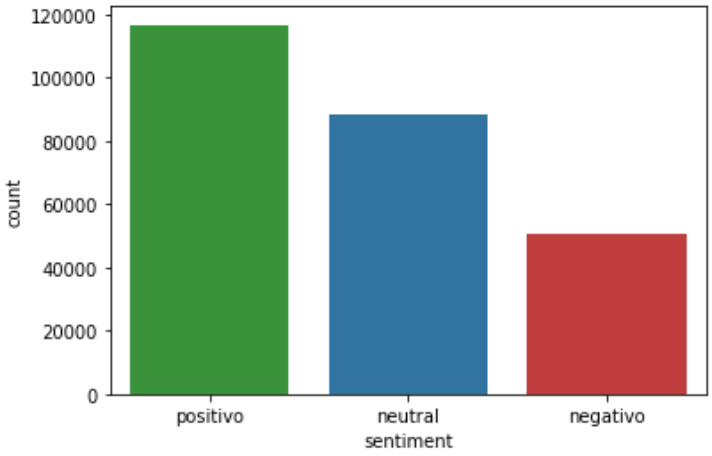


30 de abril

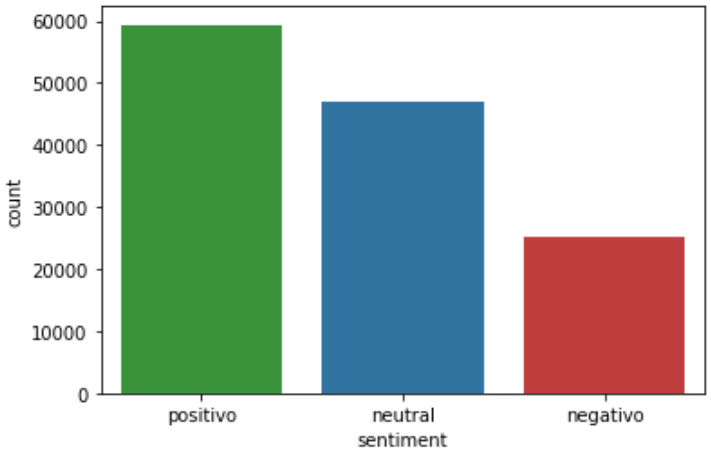
2. Clasificación de sentimiento



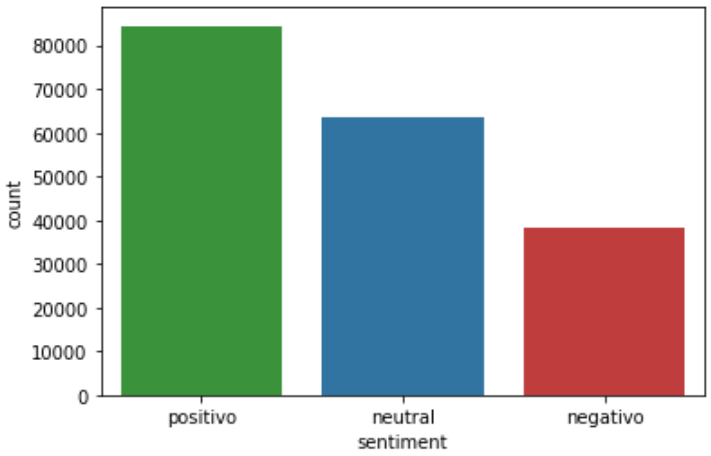
29 de marzo



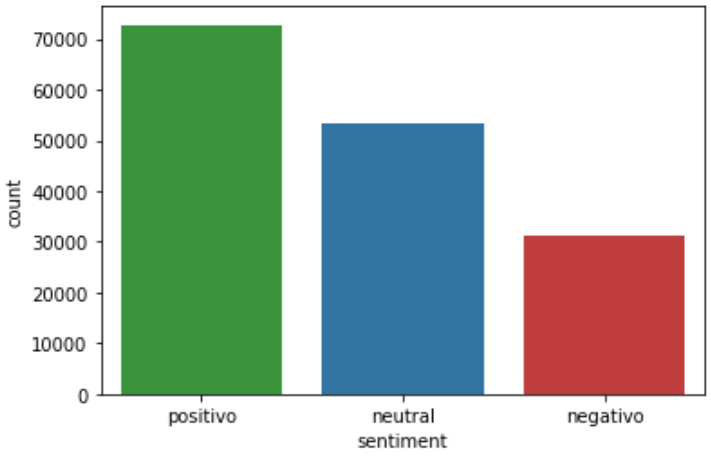
5 de abril



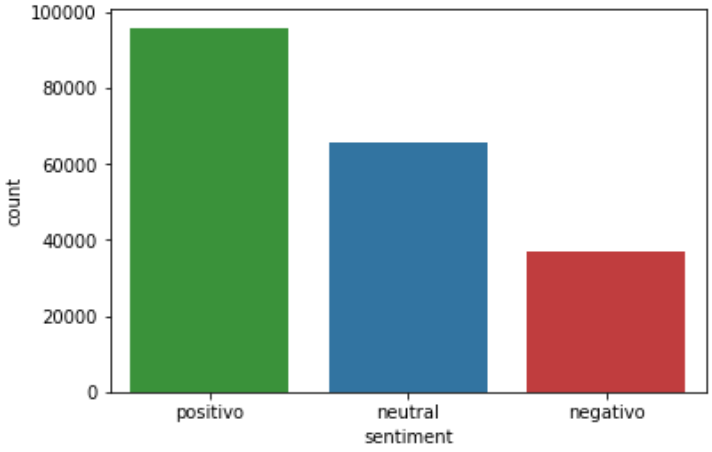
12 de abril



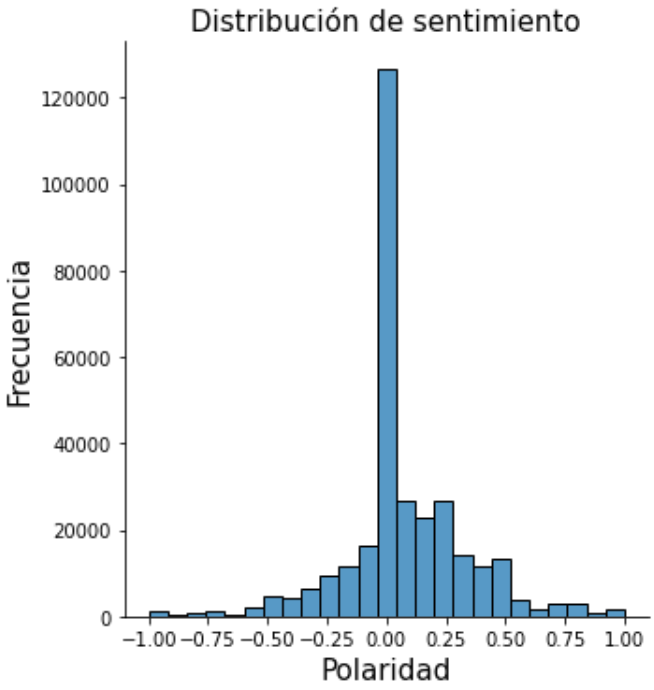
19 de abril



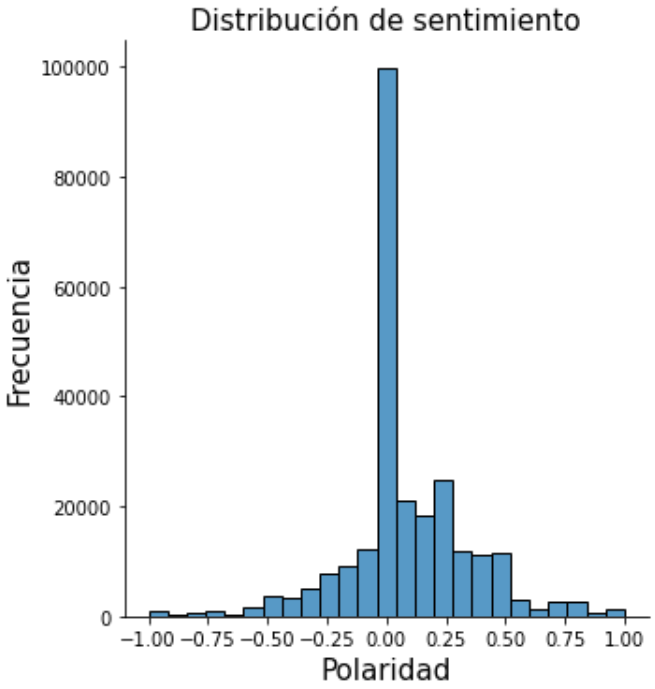
26 de abril



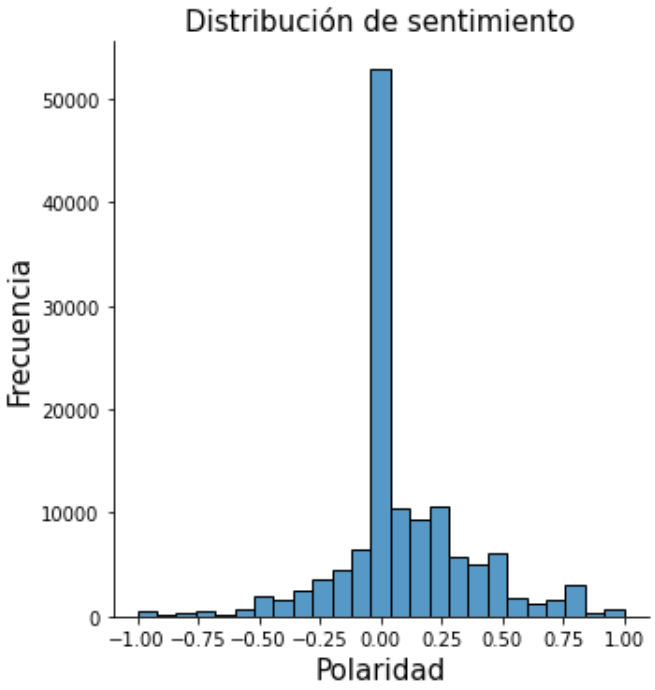
30 de abril



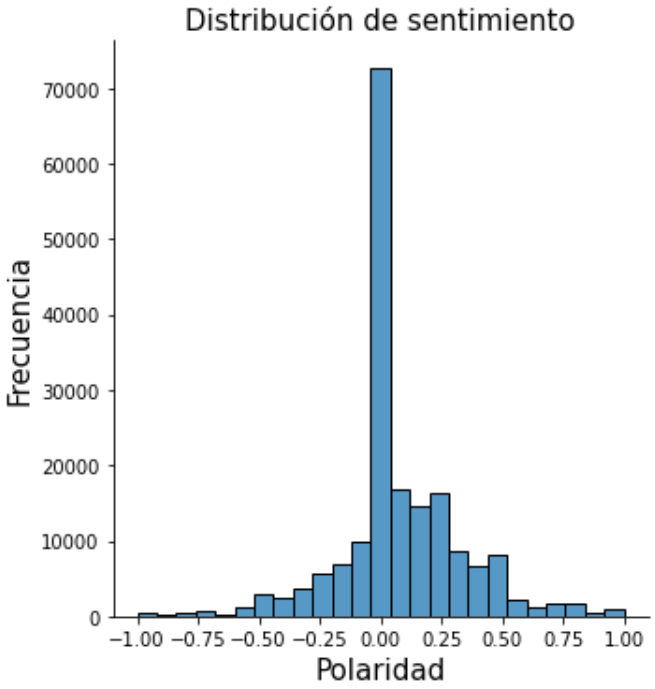
29 de marzo



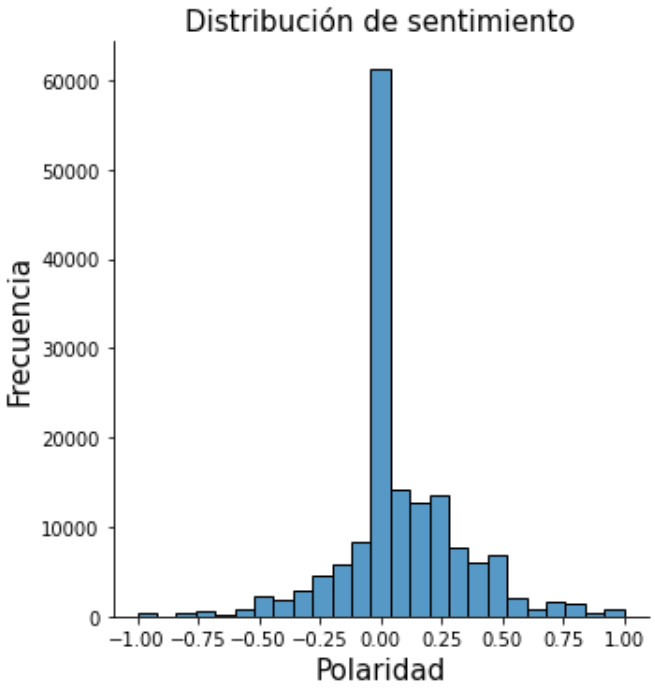
5 de abril



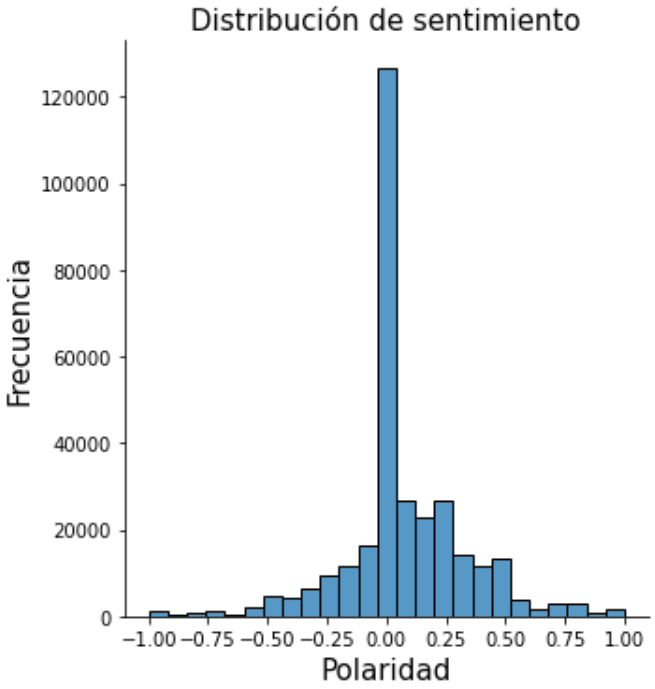
12 de abril



19 de abril

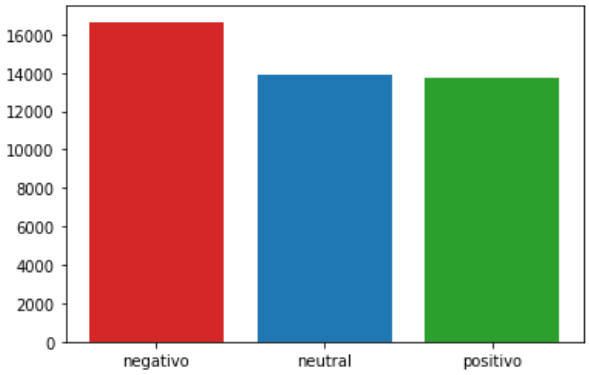


26 de abril

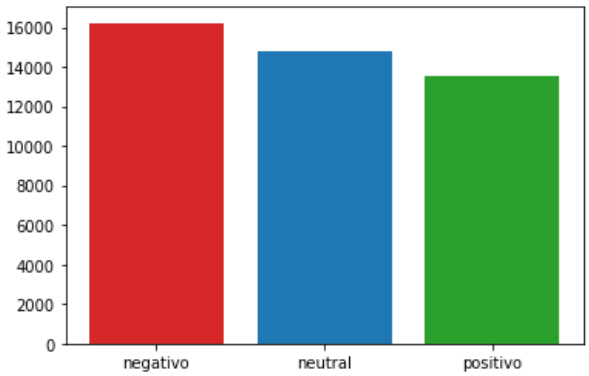


30 de abril

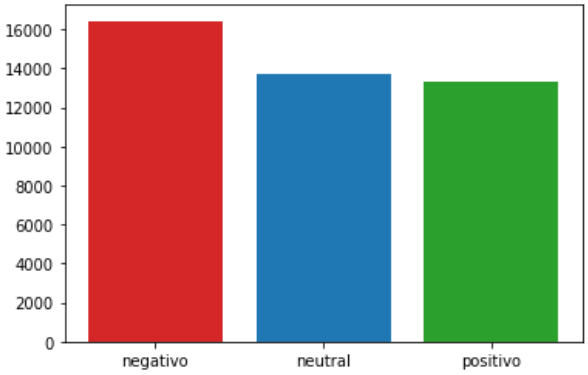
3. Ratio de favoritos



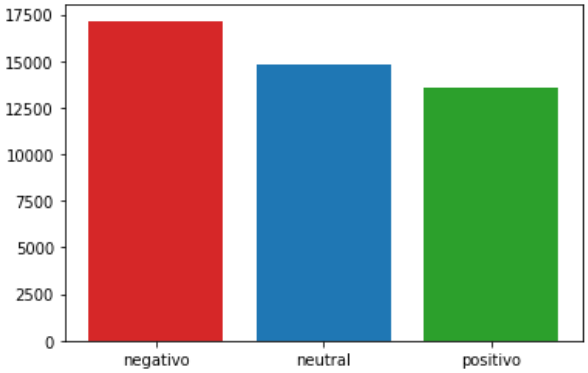
29 de marzo



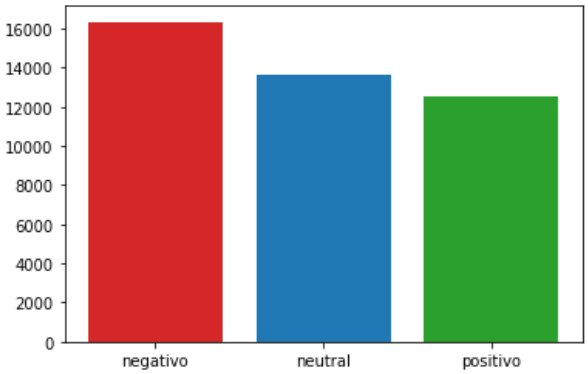
5 de abril



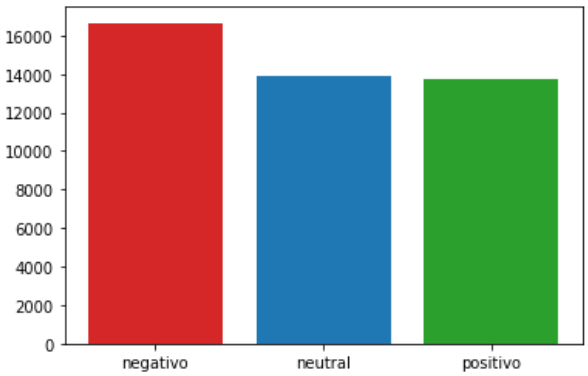
12 de abril



19 de abril



26 de abril



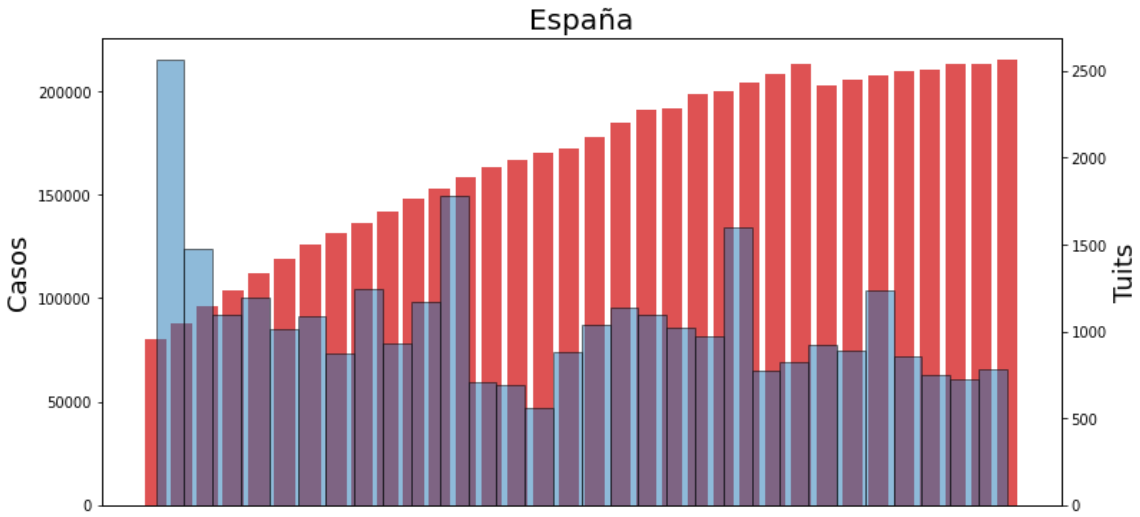
30 de abril

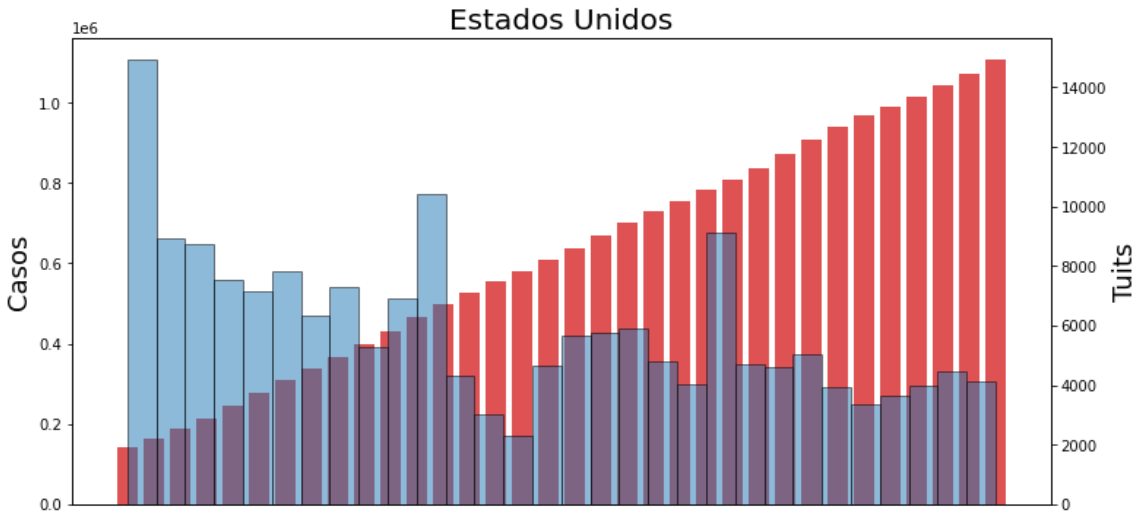
Anexo B - Análisis comparativo

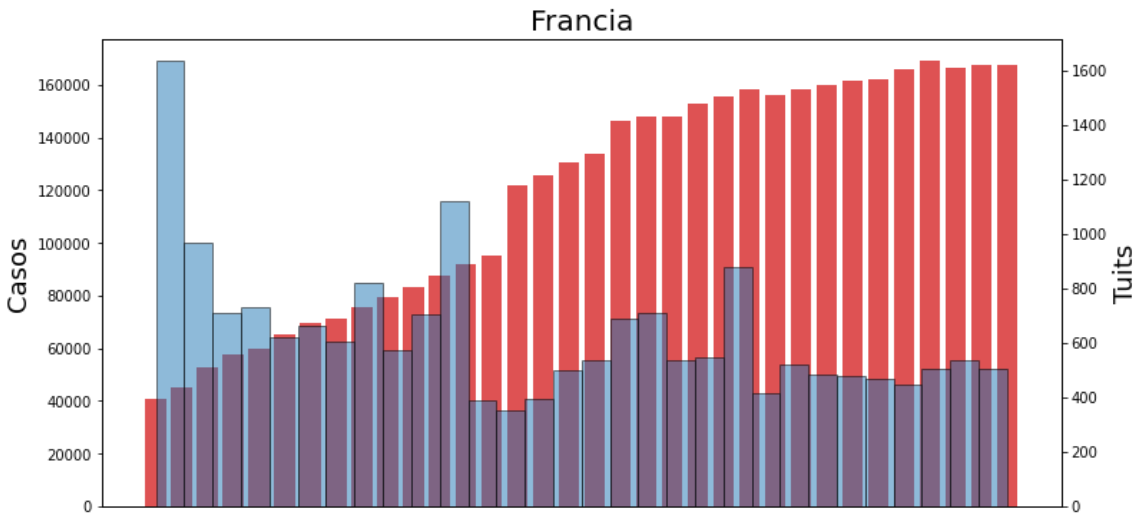
Cuadernos

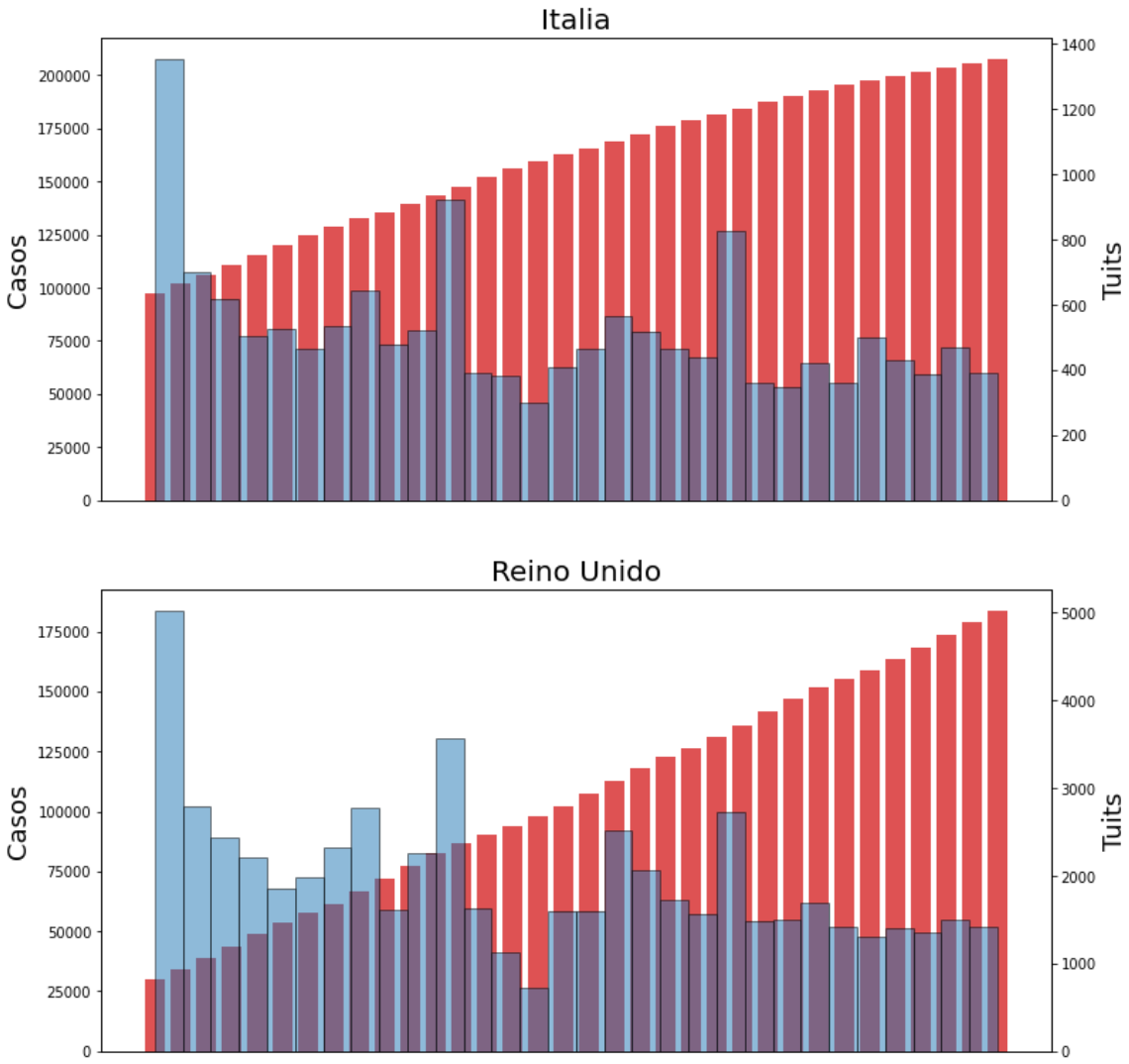
1. España: https://colab.research.google.com/drive/1h4rSFSxFhG_4opGDJVWfj7YLZ7mXCsLX?usp=sharing
2. Estados Unidos: https://colab.research.google.com/drive/1svaIfABFv9auXujzF3RYwgUSkhGbhOj3?usp=sharing
3. Italia: https://colab.research.google.com/drive/1hcPVpx5jyf-G77kk2UvWxBAiFXeWppBz?usp=sharing
4. Francia: https://colab.research.google.com/drive/1dQ4hIpQEifFKoIltkGaUMqvQy6KEZwAU?usp=sharing
5. Reino Unido: https://colab.research.google.com/drive/1fEFUm1B5c0yVGrdOlBUdfptCv6FMDB60?usp=sharing
6. Turquía: https://colab.research.google.com/drive/1c8F-FbLnqVwd2VwDf--09un_x1jqeiBg?usp=sharing
7. India: https://colab.research.google.com/drive/1NUSe6MhqsmdKnNLZlvgZZZv8z6NZXnUu?usp=sharing
8. Japón: https://colab.research.google.com/drive/1IIMdalPpaM6Zft-NYxDfwdIn9VA4K2-3?usp=sharing
9. México: https://colab.research.google.com/drive/1qAzAcAUBx25GbEhhGD9DDEJIJpGuezcn?usp=sharing
10. Brasil: https://colab.research.google.com/drive/1bANJ5yr0LH3CFfn4VmlJwznWUOJ0QLbc?usp=sharing
11. Comparativa de los contagios de todos los países: https://colab.research.google.com/drive/1yF32mRpxcjedY9ir2pr778rmzBB9gVcm?usp=sharing
12. Comparativa de los tuits de todos los países: https://colab.research.google.com/drive/1zLRBaxbDwkI6C1zp7ubuSzcdbgz0FE-r?usp=sharing

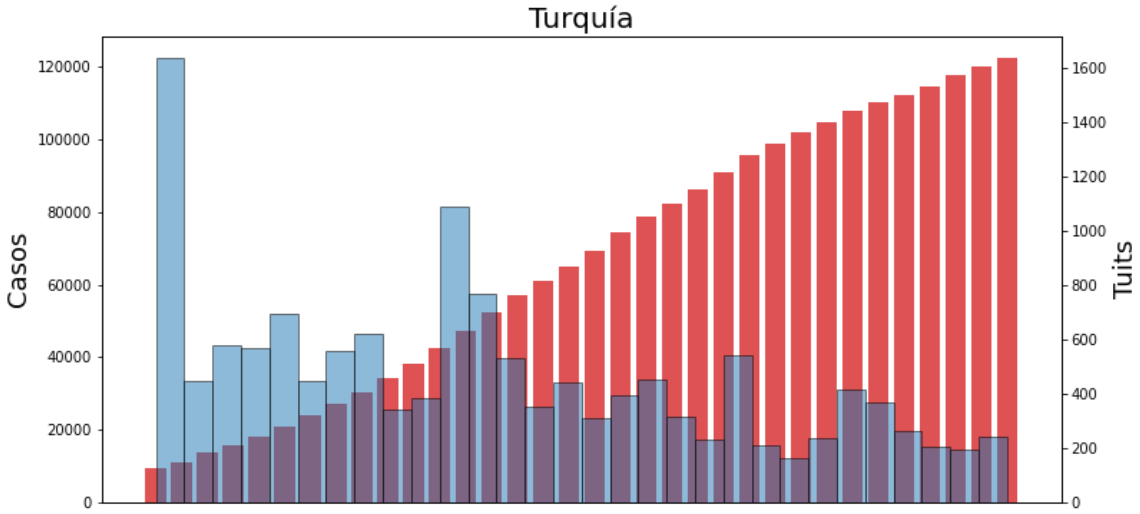
Resultados

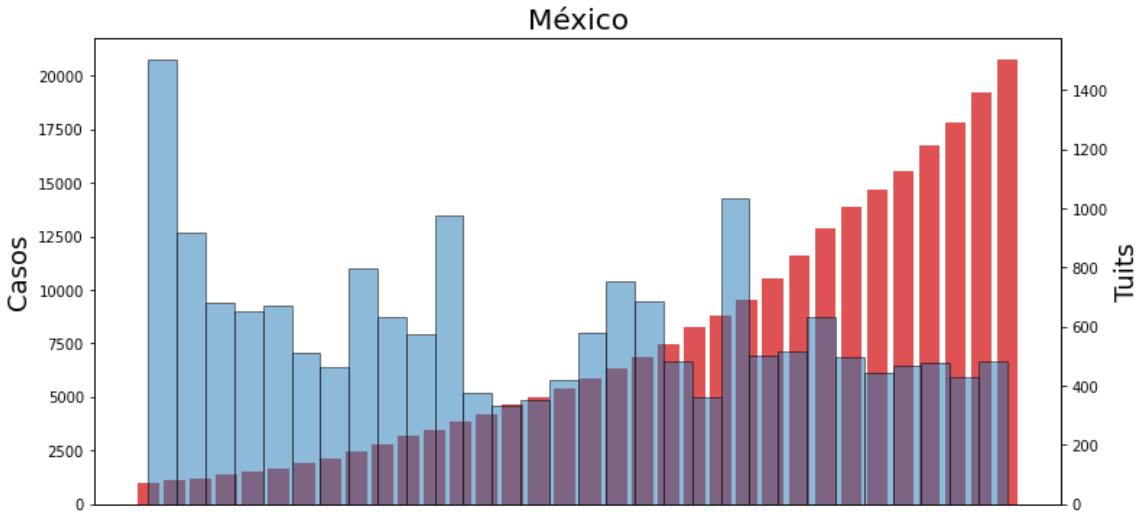

España
Casos
Tuits
0
50000
100000
150000
200000
0
500
1000
1500
2000
2500

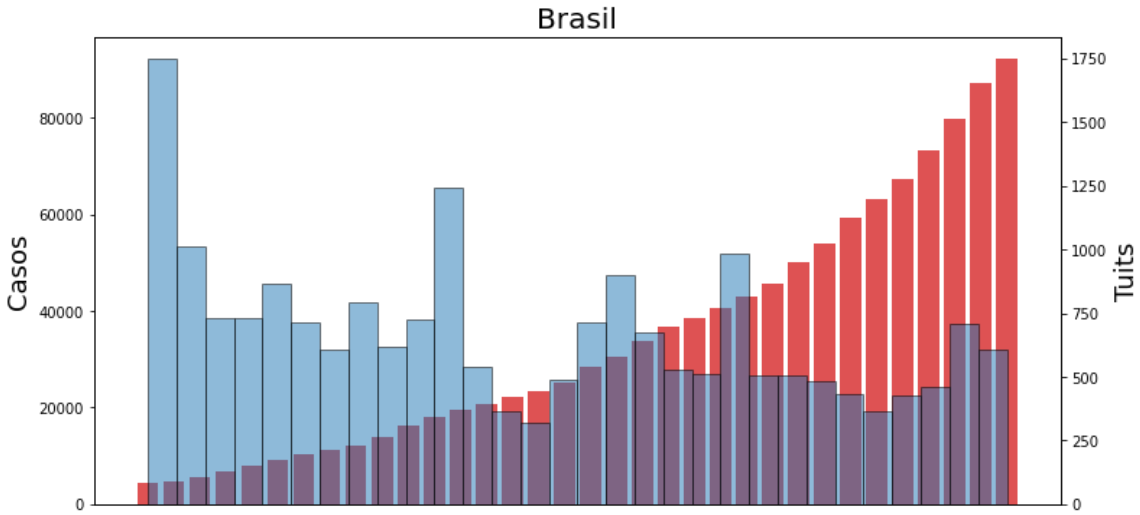

Estados Unidos
1e6
Casos
Tuits
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000
14000

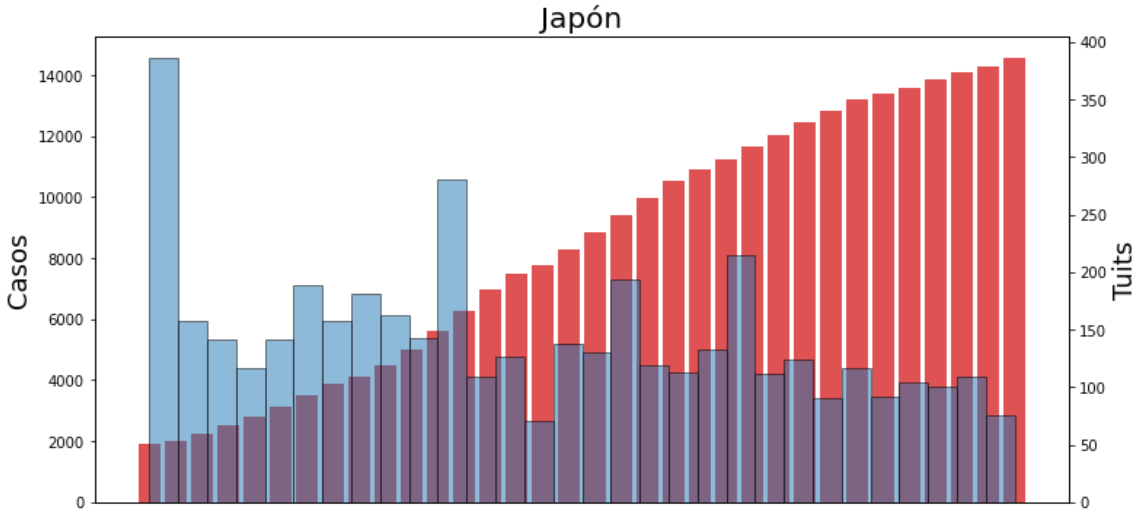

Francia
Casos
Tuits
0
20000
40000
60000
80000
100000
120000
140000
160000
0
200
400
600
800
1000
1200
1400
1600

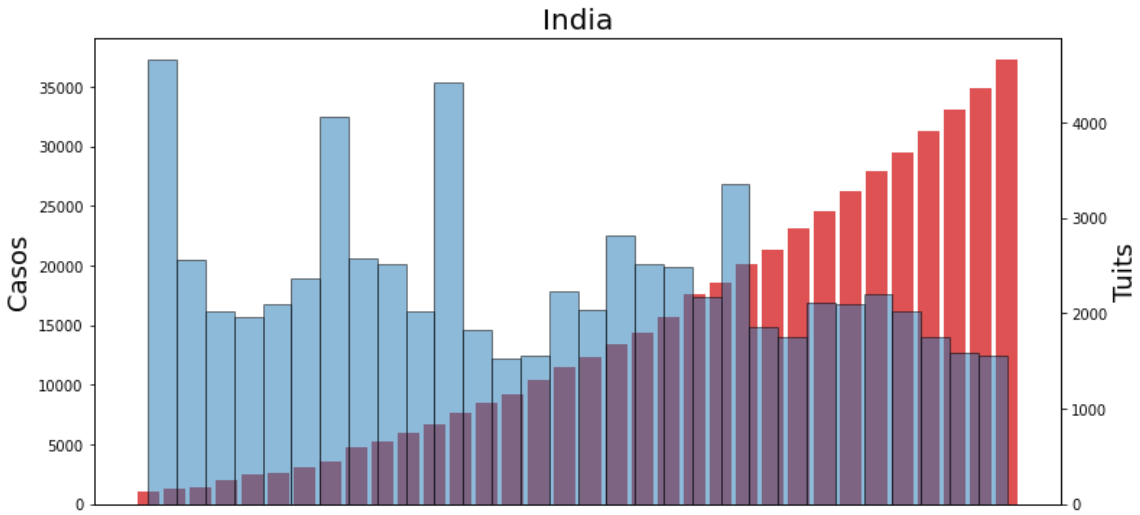

Italia
Casos
Tuits
0
25000
50000
75000
100000
125000
150000
175000
200000
0
200
400
600
800
1000
1200
1400
Reino Unido
Casos
Tuits
0
25000
50000
75000
100000
125000
150000
175000
0
1000
2000
3000
4000
5000


Turquía
Casos
Tuits
0
20000
40000
60000
80000
100000
120000
0
200
400
600
800
1000
1200
1400
1600


México
Casos
Tuits
0
2500
5000
7500
10000
12500
15000
17500
20000
0
200
400
600
800
1000
1200
1400


Brasil
Casos
Tuits
0
20000
40000
60000
80000
0
250
500
750
1000
1250
1500
1750


Japón
Casos
Tuits
0
2000
4000
6000
8000
10000
12000
14000
0
50
100
150
200
250
300
350
400


India
Casos
Tuits
0
5000
10000
15000
20000
25000
30000
35000
0
1000
2000
3000
4000