



UNIVERSITAT OBERTA DE CATALUNYA (UOC)  
MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (*Data Science*)

## TRABAJO FINAL DE MÁSTER

ÁREA: 1

### **Análisis de las elecciones presidenciales de Estados Unidos de 2020 a través de Twitter**

---

Autor: Pablo Mas Cayuelas

Tutor: Josep Maria Grau Masot

Profesor: Albert Solé Ribalta

---

Barcelona, January 3, 2021



# Créditos/Copyright



Esta obra está sujeta a una licencia de Reconocimiento - NoComercial - SinObraDerivada.

# FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis de las elecciones presidenciales de Estados Unidos de 2020 a través de Twitter
Nombre del autor:	Pablo Mas Cayuelas
Nombre del colaborador/a docente:	Josep Maria Grau Masot
Nombre del PRA:	Albert Solé Ribalta
Fecha de entrega (mm/aaaa):	01/2021
Titulación o programa:	Máster en Ciencia de Datos
Área del Trabajo Final:	Análisis de datos de Twitter
Idioma del trabajo:	Español
Palabras clave	Twitter, campaña electoral, análisis de redes

*"Es lo que hay..."*

Michael Robinson

# Agradecimientos

A Luis, Jorge, José Manuel y Teresa por estar siempre para todo. A Germán y a Mario por dar siempre su punto de vista y que casi siempre sea el bueno. A Josep Maria por su atención y genial trato durante la elaboración de este Trabajo.

Gracias.

# Abstract

En este Trabajo Final se intenta seguir la campaña electoral correspondiente a las elecciones presidenciales del año 2020 en Estados Unidos, por medio de los mensajes generados en la red social Twitter durante los días previos al día de las elecciones. Así, se podrán extraer perfiles clave para el desarrollo de la misma, analizar las interacciones entre diferentes perfiles y extraer la relación que hay entre los mismos.

El primer paso del análisis consistirá en la extracción de todos los tuits por medio de palabras clave (los nombres de los partidos políticos o de sus líderes) o mediante *hashtags* que identifiquen a los mismos.

Además, para saber qué perfiles y palabras son los más importantes en este análisis se medirá la frecuencia de aparición de los mismos en los tuits extraídos, siendo las palabras y los perfiles con más frecuencia de aparición los más importantes para el presente Trabajo.

**Palabras clave:** Twitter, campaña electoral, análisis de redes, Estados Unidos, redes sociales.

# Índice

<b>Abstract</b>	<b>7</b>
<b>1. Introducción</b>	<b>13</b>
1.1. CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO	13
1.2. OBJETIVOS DEL TRABAJO	14
1.3. HIPÓTESIS DEL TRABAJO	15
1.4. ENFOQUE Y MÉTODO SEGUIDO	15
1.5. PLANIFICACIÓN DEL TRABAJO	16
1.6. BREVE SUMARIO DE LOS PRODUCTOS OBTENIDOS	16
1.7. BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA	17
<b>2. Estado del arte</b>	<b>18</b>
2.1. TWITTER COMO HERRAMIENTA COMUNICATIVA	18
2.2. TWITTER Y EL PROCESO ELECTORAL	19
2.3. TWITTER Y BIG DATA	20
<b>3. Diseño e implementación del trabajo</b>	<b>24</b>
3.1. EXTRACCIÓN DE LOS DATOS	24
3.2. PRESENTACIÓN DE LOS DATOS	28
3.3. ANÁLISIS DE LOS OBJETIVOS	29



3.3.1. SEGUIMIENTO DIARIO DE MENSAJES EN TWITTER . .	29
3.3.2. DIFERENCIA DE INTERACCIONES ENTRE PARTIDOS Y LÍDERES . . . . .	33
3.3.3. INTERACCIONES DE LAS PALABRAS MÁS FRECUENTES	36
3.3.4. ANÁLISIS DE LOS TUI TS CON CONTENIDO MULTIME- DIA Y HASHTAGS . . . . .	39
3.3.5. DISPOSITIVO DESDE EL QUE HAN SIDO ENVIADOS LOS TUI TS . . . . .	40
3.3.6. ANÁLISIS DE SENTIMIENTOS DE LOS USUARIOS DE TWITTER . . . . .	41
<b>4. Análisis de los resultados . . . . .</b>	<b>45</b>
4.1. SEGUIMIENTO DIARIO DE LOS MENSAJES EN TWITTER	45
4.2 DIFERENCIA DE INTERACCIONES ENTRE PARTIDOS Y LÍDERES . . . . .	46
4.3. INTERACCIONES DE LAS PALABRAS MÁS FRECUENTES .	47
4.4. ANÁLISIS DE LOS TUI TS CON CONTENIDO MULTIMEDIA Y HASHTAGS . . . . .	47
4.5. DISPOSITIVOS DESDE EL QUE HAN SIDO ENVIADOS LOS TUI TS . . . . .	48
4.6. ANÁLISIS DE SENTIMIENTOS DE LOS USUARIOS DE TWITTER . . . . .	48
<b>5. Conclusiones . . . . .</b>	<b>50</b>
<b>6. Líneas de trabajo futuras . . . . .</b>	<b>53</b>
<b>7. Bibliografía . . . . .</b>	<b>54</b>

# List of Figures

1	Tareas a realizar . . . . .	16
2	creación de una aplicación de Twitter para la obtención de las credenciales de OAuth . . . . .	21
3	Nubes de palabra y frecuencias de las mismas para el día 01/11/2020 con la librería <i>twitteR</i> . . . . .	30
4	Nubes de palabra y frecuencias de las mismas para el día 02/11/2020 con la librería <i>twitteR</i> . . . . .	30
5	Nubes de palabra y frecuencias de las mismas para el día 03/11/2020 con la librería <i>twitteR</i> . . . . .	31
6	Nubes de palabra y frecuencias de las mismas para el día 02/11/2020 con la librería <i>Rtweet</i> . . . . .	31
7	Nube de palabras de todos los tuits . . . . .	32
8	Frecuencia relativa de las palabras más repetidas . . . . .	32
9	Interacciones de los tuits que contienen los nombres de los candidatos	33
10	Interacciones de los tuits que mencionan los perfiles de los candidatos	34
11	Interacciones de los tuits de los perfiles de los candidatos . . . . .	34
12	Tuits y retuits por segundo que contienen <i>Trump</i> y <i>Biden</i> . . . . .	35

13	Tuits y retuits por segundo que contienen <i>@realDonaldTrump</i> y <i>@Joe-Biden</i> . . . . .	36
14	Tuits y retuits por segundo publicados por las cuentas <i>@realDonaldTrump</i> y <i>@JoeBiden</i> . . . . .	36
15	Interacciones de los tuits de las palabras más frecuentes . . . . .	38
16	Media de interacciones por tuit de las palabras más frecuentes . . . . .	38
17	Interacciones de tuits con contenido multimedia y sin él . . . . .	39
18	Interacciones de tuits con <i>hashtags</i> y sin ellos . . . . .	40
19	Interacciones de tuits dependiendo del dispositivo . . . . .	41
20	Sentimientos de los tuits por días del análisis . . . . .	42
21	Sentimientos de los tuits que mencionan el nombre de Trump y que nombran su perfil de Twitter . . . . .	43
22	Sentimientos de los tuits que mencionan el nombre de Biden y que nombran su perfil de Twitter . . . . .	43
23	Sentimientos de los tuits que mencionan los <i>hashtags</i> <i>#MAGA</i> y <i>#Trump2020</i> . . . . .	44

# List of Tables

1	Campos de los tuits descargados . . . . .	25
2	Distribución y cantidad de tuits extraídos por día con <i>Rtweet</i> . . . . .	26
3	Distribución y cantidad de tuits extraídos por día con <i>twitteR</i> . . . . .	27
4	Cantidad de tuits extraídos de los usuarios de Trump y Biden . . . . .	27
5	Campos de los tuits correspondientes a la librería <i>twitteR</i> . . . . .	28
6	Campos de los tuits correspondientes a la librería <i>Rtweet</i> . . . . .	29
7	Palabras más mencionadas por día . . . . .	45
8	Aparición de las palabras clave . . . . .	45
9	Aparición de las palabras clave . . . . .	46
10	Palabras con más interacciones de la campaña . . . . .	47

# 1. Introducción

## 1.1. Contexto y Justificación del trabajo

Es evidente la importancia de las redes sociales como instrumento comunicativo para los partidos políticos en las campañas electorales. Concretamente, la campaña presidencial de Estados Unidos en el año 2008 supuso un antes y un después en este aspecto (Abejón et, 2012), en la que esta influencia fue clave para que Barack Obama fuera elegido presidente. A partir de este momento, tanto los partidos políticos como los medios de comunicación con diferentes ideologías apostaron por un continuo uso de las redes sociales para comunicar sus intenciones e ideas políticas.

En esta línea, se tiene a la red social Twitter, que se ha convertido a nivel mundial en la principal herramienta comunicativa, pues posibilita la visualización, de forma compacta, de todo tipo de opiniones de los diversos partidos políticos, medios de comunicación, periodistas y otros perfiles de interés. Esta aparente diversidad es lo que la ha llevado al éxito, pues los usuarios pueden formar su opinión seleccionando perfiles de interés a los que seguir.

Por otro lado, también a modo de justificación del trabajo, resulta interesante ver cómo los partidos políticos y los medios de comunicación utilizan unas palabras u otras en función de la repercusión previa de las mismas: si la respuesta a un determinado tuit es abundante, se puede tener al mismo como referencia para profundizar en la línea tratada. Además, no solo es importante interaccionar con los máximos perfiles posibles, sino que estas cuentas, a su vez, tengan bastante repercusión (muchos seguidores, por ejemplo), pues podrán compartir la información inicial con sus seguidores y se llegará, indirectamente, a un número mayor de perfiles.

En este punto se pueden realizar incluso consideraciones éticas, pues semanalmente se publican noticias sobre la creación de cuentas falsas que tratan de publicitar y dar más interacciones a determinados tuits. De este modo, resulta interesante diferenciar entre la repercusión aparente de las publicaciones y la repercusión real en la sociedad.

En lo que respecta a la motivación personal, soy un usuario activo de Twitter y

lo utilizo en muchos casos como fuente de información política, pues tengo bastante interés en este tipo de noticias. Siempre me ha resultado curioso ver cómo, en función de las personas a las que sigo, me llegan tuits con una determinada ideología, y este Trabajo es ideal para comprender el porqué de esta cuestión.

Por tanto, tras evidenciar la importancia de Twitter para publicitar opiniones políticas, sería útil analizar los tuits publicados durante esta campaña electoral para poder comprender cómo funcionan las interacciones en esta red social, por medio de un seguimiento diario de cada mensaje generado en la misma. Distinguiéndolos por días, además, se podría ver una tendencia de los mismos y apreciar si, a medida que se acerca el día de las elecciones, hay determinados perfiles y palabras clave que tienen más interacciones.

El período de tiempo estudiado abarcan los tres días previos a las elecciones: 1, 2 y 3 de noviembre de 2020. Durante este tiempo, se han extraído los datos (tuits) a estudiar.

Para extraer y analizar los tuits, se ha utilizado el programa Rstudio, el cual cuenta con librerías concretas para la extracción de tuits (*twitteR* o *Rtweet*), especificando en el mismo las palabras clave o los perfiles concretos a los que se quiere acceder para el análisis.

Una vez extraídos los tuits es posible utilizar métricas de centralidad para ver si hay perfiles clave a los que vayan dirigidos la mayoría de tuits.

## **1.2. Objetivos del trabajo**

Los objetivos principales de este trabajo de investigación son:

- Hacer un seguimiento diario de los mensajes generados y difundidos en Twitter durante la campaña electoral estadounidense del año 2020.
- Una vez determinados los perfiles importantes, realizar un análisis cualitativo de sus publicaciones y así determinar palabras clave para la campaña.

Como objetivos secundarios:

- Analizar las diferencias en las interacciones entre los diferentes partidos y líderes políticos.
- Observar si los tuits con contenido multimedia tienen más interacciones que los que no cuentan con él.
- Analizar cuáles son las palabras cuyos tuits cuentan con un mayor número de interacciones.

Es conveniente describir como podemos conseguir alguno de los objetivos. Por ejemplo, para ver si los tuits con contenido multimedia tienen más interacciones que el resto, se podría separar la base de datos en dos; por un lado, aquellos que contienen este contenido (que serán aquellos que tengan un enlace a dicho contenido del tipo "<https://t.co/>", pues así es como se enlazan todas las imágenes y vídeos en esta red social); y por otro lado el resto, que no lo contendrán.

### 1.3. Hipótesis del trabajo

La hipótesis principal del trabajo es:

- Hay claras diferencias en las interacciones de los tuits que hacen referencia a Donald Trump y a Joe Biden.

Por tanto, la pregunta a responder en el Trabajo será:

- ¿Existe diferencia en el número de interacciones y en la frecuencia de publicación de tuits que hacen referencia a los candidatos a la presidencia de Estados Unidos?

### 1.4. Enfoque y método seguido

Para todo el proceso de obtención y análisis de datos, se ha utilizado el programa *Rstudio*. Además, para la extracción de los mismos, se ha recurrido principalmente a las librerías *Rtweet* y *twitteR*.

Como es lógico, es interesante obtener tuits que hablen sobre las elecciones (partidos políticos, candidatos): esto se puede conseguir principalmente por medio de *hashtags* que involucren a los participantes en las elecciones.

Además, pensando en la limpieza de los datos, se comprobará que no hay tuits

repetidos que puedan llevar a que el análisis no sea del todo riguroso: por ejemplo, si un mismo tuit tiene dos *hashtags* diferentes, es posible que quede guardado dos veces en nuestra base de datos por haberlo descargado por duplicado, una vez con cada *hashtag*.

## 1.5. Planificación del trabajo

Para una correcta elaboración del trabajo, podemos dividir éste en varias etapas, estimando fechas de inicio y fin de las mismas:

Tarea	Duración	Inicio	Fin
Definición y justificación del trabajo	11 días	16/09/2020	27/09/2020
Extracción de los datos	43 días	27/09/2020	03/11/2020
Análisis de los datos	19 días	03/11/2020	22/11/2020
Obtención de conclusiones	14 días	22/11/2020	06/12/2020
Redacción de la memoria	14 días	06/12/2020	20/12/2020
Presentación	7 días	20/12/2020	27/12/2020

Figure 1: Tareas a realizar

Fuente: elaboración propia

No es una planificación compleja, pues el inicio de cada tarea se corresponde con el fin de la tarea anterior.

## 1.6. Breve resumen de los productos obtenidos

En las páginas 12, 13 y 14 de este trabajo se tiene un índice-resumen de las figuras y tablas elaboradas para el mismo. Todas ellas se han elaborado para la consecución de los diferentes objetivos propuestos anteriormente. Para ello, básicamente se ha recurrido al programa *Rstudio*, como se ha mencionado con anterioridad, que permite la creación de los diagramas de barra y *wordclouds* (nubes de palabras) comentados en el resto de capítulos de la memoria.



## 1.7. Breve descripción de los otros capítulos de la memoria

El siguiente capítulo de esta memoria está dedicado al estado del arte referente a la línea de trabajo que se está desarrollando. Éste se dedica, en primer lugar, a la exposición de Twitter como herramienta comunicativa, pues es de vital importancia entender cómo funcionan las interacciones en esta red social para poder obtener conclusiones reales para el trabajo. En concreto, un subapartado de este capítulo es dedicado a la relación directa de Twitter con el proceso electoral, mencionando estudios y conclusiones referentes a otros procesos electorales. Por último, se expone cómo funciona la API de Twitter, a la que se recurre para el proceso de extracción de datos.

El tercer apartado de esta memoria es la parte principal del trabajo, pues hace referencia al análisis y a la implementación del mismo. En él, se explica el proceso de extracción de datos con las librerías *Rtweet* y *twitteR*, se presentan todos los datos obtenidos y se exponen los gráficos elaborados para la consecución de cada uno de los objetivos propuestos en este apartado.

El cuarto capítulo se dedica principalmente al análisis de los resultados obtenidos, extrayendo directamente conclusiones para los objetivos del subapartado 1.2. En el capítulo quinto se exponen las conclusiones generales del trabajo y en el sexto se comentan posibles mejoras del mismo y líneas de trabajo para ampliarlo.

## 2. Estado del arte

En este apartado se debe exponer cómo, hasta la fecha, se han abordado objetivos e hipótesis similares a las de este trabajo. Una gran parte del mismo estará relacionada con la forma en la que nuestro programa de extracción y análisis de datos trabaja (*Rstudio*), mencionando las librerías más importantes para este fin (*Rtweet* y *twitterR*), así como sus funcionalidades.

Actualmente, son numerosos los trabajos realizados en lo que respecta a la extracción y análisis de información de Twitter. Si bien otros programas como *Python* ofrecen también esta posibilidad, *Rstudio* proporciona herramientas más sencillas para este propósito.

### 2.1. Twitter como herramienta comunicativa

Las redes sociales son sistemas de comunicación básicos. Comparándolas con los medios de comunicación tradicionales, tienen todos los elementos de estos, pero incorporan la posibilidad de interacción entre los medios y los usuarios, así como entre los distintos usuarios entre sí (Campos Freire, 2008).

Los retuits, *likes* y *hashtags* hacen que todo esto sea posible: la información se propaga por la red social de forma exponencial, pues cada usuario hace, por medio de un retuit, que cierta información llegue a todos sus seguidores, y así sucesivamente. De esta forma, a diferencia de los medios de comunicación tradicionales, los propios usuarios deciden qué información es la que va a llegar al resto, por lo que es también clave saber qué mensajes van a tener éxitos y cuáles no.

Por lo tanto, si un usuario quiere hacer llegar al resto un mensaje concreto, es importante también la manera de hacerlo. En este sentido, varios autores (Alvídrez, 2016) han analizado el efecto de diferentes estilos lingüísticos en Twitter, sobre todo en lo que respecta a la credibilidad de los tuits, analizando incluso por géneros la misma. En esta misma línea, otros autores (Lee, 2012) escriben sobre la personalización de los mensajes: cuando estos se encuentran personalizados hay una mayor probabilidad de atraer al público, pero éste es generalmente un público muy específico. Esto es directamente aplicable a los mensajes políticos.

Una vez que llegan los mensajes a su público, que estos sean influyentes también es importante. Si existe influencia social, las ideas pueden llegar a difundirse a través de las redes sociales de una forma extremadamente rápida. Hay autores (Anagnostopoulos et, 2008) que afirman la existencia de esta influencia en redes sociales, pero solo llegan a medirla cualitativamente.

Se debe subrayar que el perfil de usuario de Twitter es limitado; las opiniones más frecuentes que se realizan en esta red social están asociadas a un perfil concreto, por lo que esto no debe hacer pensar que las opiniones de Twitter sea el pensamiento general de la sociedad. Ahora bien, podríamos preguntarnos si ambos pensamientos están relacionados entre sí.

La capacidad de interaccionar con otros usuarios es clave en este punto. Los *likes*, *rtweets*, menciones y seguidores permiten que se vinculen los usuarios, pero esto se convierte en un arma de doble filo: el hecho de que un usuario pueda seleccionar a quién seguir hace que el contenido que llegue a él sea, generalmente, afín a su ideología. Además, Twitter, con sus últimas actualizaciones, ordena los tuits que llegan a los usuarios según sus intereses, por lo que las restricciones aumentan en este sentido.

## **2.2. Twitter y el proceso electoral**

En Twitter las opiniones sobre los procesos electorales están a la orden del día. Los candidatos utilizan esta red social prácticamente como su principal recurso para llegar a la mayor parte de la población posible.

Como se mencionó en la introducción, la campaña presidencial estadounidense del año 2008 supuso un antes y un después en el uso de Twitter para la comunicación política (Abejón et, 2012). Más todavía es reconocida la importancia de esta red social en los últimos años, con el ejemplo de uno de los candidatos a las elecciones de 2020, Donald Trump: en el proceso electoral de 2016 ganó las elecciones con muchos de los medios de comunicación en su contra, por lo que su publicidad por redes fue clave para este propósito.

Las últimas investigaciones en la relación de Twitter con las campañas electorales

analizan desde la credibilidad de los candidatos en sus discursos en redes, llegando a la conclusión de que las opiniones de estos son creíbles para el caso del proceso electoral de 2016 (Pressgrove et, 2018), hasta el tono de los candidatos (Enil, 2017). Para analizar este tono, por ejemplo, los autores presentan un análisis cuantitativo y cualitativo de los tuits con más interacciones de las elecciones de los Estados Unidos en el año 2016, etiquetándolos con el candidato al que apoyan el tuit o, en caso de no apoyar a ninguno, como neutrales, pudiendo concluir diferencias entre las estrategias de ambos candidatos (en este caso Hillary Clinton y Donald Trump).

En el punto anterior se ha evidenciado que la muestra más representativa del pensamiento actual de la sociedad lo podemos encontrar en Twitter, pero no debemos caer en extrapolar estas opiniones como el pensamiento general de la misma. En esta línea, hay autores (Gayo-Avello, 2012) (Gayo-Avello, 2013) que hablan de la posibilidad de predecir el resultado electoral en base a las opiniones en Twitter. Aparadamente es posible pensar que las opiniones reflejadas en Twitter tienen una relación lineal con las opiniones reales de la sociedad pero, cuando se realiza un análisis profundo de la situación, hay muchas consideraciones que pueden pasar por alto en el análisis, algo más sutiles. En estos artículos se concluye que el poder electoral de las redes sociales se ha exagerado y que actualmente la investigación no ha proporcionado pruebas sólidas que hagan que un análisis de Twitter sea mejor, por ejemplo, que una encuesta.

### **2.3. Twitter y Big Data**

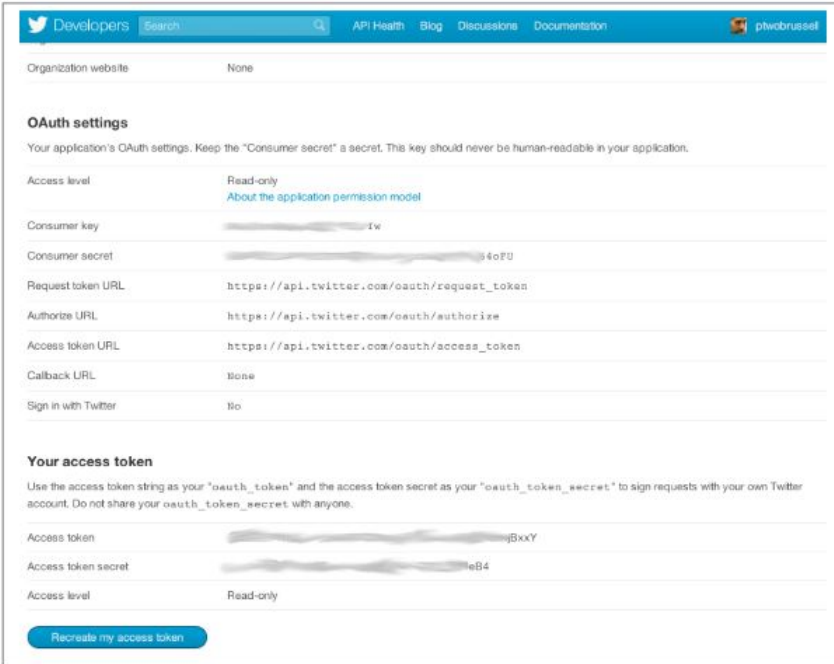
Como es de esperar, Twitter es una fuente inmensa de información que se encuentra abierta al consumo propio por medio de una API limpia y bastante documentada. Los datos de Twitter tienen la particularidad de que podemos obtenerlos prácticamente a tiempo real, lo cual hace que sea la muestra más representativa del pensamiento actual de la sociedad (Matthew et, 2013).

La API de Twitter permite la extracción de tuits y la información de los mismos y de perfiles concretos. En el código de *Rstudio* se debe hacer referencia a nuestro nombre de usuario y contraseña: como esta información no es conveniente que la escribamos en el código, se recurre generalmente a *OAuth*, que nos permite identi-

ficarnos con nuestra cuenta sin necesidad de escribir la contraseña. Para ello, por medio de una librería PHP, *OAuth* hace de intermediario entre los otros usuarios y nuestra cuenta; es decir, estamos recurriendo a una aplicación que va a autorizar el acceso a los datos de nuestra cuenta. Para la identificación utilizaremos los siguientes parámetros:

- *consumer\_key*: contraseña de la API de Twitter.
- *consumer\_secret*: Usuario de la API de Twitter.
- *access\_token*: *token* de acceso a la API: solo de manera temporal.
- *access\_secret*: clave del *token* de acceso a la API.

Estas credenciales permiten publicar tuits en la cuenta a la que están ligados, por lo que se debe tratar con la misma confidencialidad que la contraseña.



The screenshot shows the Twitter Developers OAuth settings page. The header includes the Twitter logo, 'Developers' text, a search bar, and links for 'API Health', 'Blog', 'Discussions', and 'Documentation'. The user 'ptwobruasel' is logged in. The 'Organization website' is set to 'None'. The 'OAuth settings' section includes a warning about the 'Consumer secret' and a table of settings: 'Access level' (Read-only), 'Consumer key' (a long alphanumeric string), 'Consumer secret' (a long alphanumeric string), 'Request token URL' (https://api.twitter.com/oauth/request\_token), 'Authorize URL' (https://api.twitter.com/oauth/authorize), 'Access token URL' (https://api.twitter.com/oauth/access\_token), 'Callback URL' (None), and 'Sign in with Twitter' (No). The 'Your access token' section includes instructions and a table with 'Access token' (a long alphanumeric string), 'Access token secret' (a long alphanumeric string), and 'Access level' (Read-only). A 'Recreate my access token' button is at the bottom.

Figure 2: creación de una aplicación de Twitter para la obtención de las credenciales de OAuth

Fuente: elaboración propia

Como se ha mencionado con anterioridad, la extracción de la información pre-

sente en Twitter ha sido abordada desde diferentes perspectivas y programas. Para este trabajo, se recurre a *Rstudio*, por lo que también es necesario teorizar sobre sus herramientas y en especial sobre las librerías *Rtweet* y *twitteR*, a las que se recurren frecuentemente para el propósito del mismo.

Con estas librerías es posible la extracción de los tuits de interés. Por ejemplo, es posible utilizar la función *userTimeline()* para descargar tuits de un solo usuario especificando cuántos tuits queremos del mismo. De esta manera, en numerosos artículos se han extraído los  $n$  primeros tuits de los principales candidatos a unas elecciones (en este caso se extraen los  $n$  primeros tuits de Donald Trump y Joe Biden). El inconveniente de la API de Twitter es que no se pueden solicitar tuits de hace más de siete días, lo cual limitará nuestro estudio.

Una vez extraídos los tuits de interés, no resulta demasiado difícil la representación de las palabras más utilizadas por cada perfil o incluso en toda nuestra base de datos de manera genérica. Así, por ejemplo, si se quisiera saber qué palabras utilizan más los candidatos o en general por los usuarios que tuitean en base a las elecciones, se podría realizar una representación con ayuda de la librería *ggplot2* de cada string en función de la frecuencia en la que aparecen en cada perfil.

Es posible también la extracción de los tuits que contengan un mismo *hashtag* por medio de la función *search\_tweets()* de *Rtweet* y de su correspondiente de la librería *twitteR* (*searchtwitter()*), introduciendo el *hashtag* (o incluso palabras clave) en esta función. De esta forma, se podrían utilizar *hashtags* referentes a las elecciones de Estados Unidos para filtrarlos y poder ver cuáles son las palabras más utilizadas, en general, por los usuarios de Twitter.

Así, por medio de estos buscadores, es posible llevar a cabo uno de los objetivos del trabajo, que es comprobar la presencia de polarización de los usuarios de Twitter entre los dos principales candidatos: se puede ver cuántos tuits hablan de un partido u otro, o de un candidato u otro.

Un punto clave dentro del trabajo y de estas librerías hace referencia a las interacciones de los tuits. Por medio de estas es posible la consecución de varios de los objetivos iniciales de este trabajo, como el análisis de las diferencias en las interac-

ciones entre los diferentes partidos y líderes políticos; ver si los tuits con contenido multimedia tienen más interacciones que los que no cuentan con él; o el análisis de la cantidad de interacciones que tiene, de media, los tuits con una determinada palabra.

Por medio de los campos *retweet\_count* y *favourite\_count* de *twitteR* y *retweet-Count* y *favouriteCount* de *Rtweet* es posible ver las palabras que tengan más o menos repercusión, pues permitirán etiquetar cada tuit con su número de interacciones (retuits o likes) y comparar el número de retuits de un tuit concreto con el de otros. En este sentido, será posible filtrar los tuits por candidato y ver cuál de los dos tienen más interacciones.

El segundo de los objetivos secundarios (comparación del contenido multimedia) también podrá llevarse a cabo por este método: por medio de la separación de los tuits que tienen este contenido de los que no y, apoyándonos también en las funciones *retweet\_count* y *favourite\_count*, será posible ver cuál de las dos clases de tuits (con contenido multimedia y sin él) tienen más interacciones.

### 3. Diseño e implementación del trabajo

Durante los días previos a las elecciones estadounidenses (3 de noviembre de 2020) se han extraído por medio de la API de Twitter los ficheros utilizados para el análisis del presente trabajo. Los días utilizados para el análisis han sido el 1 de noviembre, 2 de noviembre y 3 de noviembre de 2020, que se corresponden con los tres días previos a las elecciones (los datos extraídos el 3 de noviembre fueron antes de la salida de los primeros resultados). Se han elegido estos días debido a que era el momento en el que se podían obtener una mayor cantidad de tuits.

Con la ayuda de *Rstudio* y de sus librerías *twitterR* y *Rtweet*, se han exportado un total 53 archivos en formato *.csv* para el análisis de los mismos.

#### 3.1. Extracción de los datos

La forma a proceder para esta extracción ha sido la siguiente: en primer lugar, se ha realizado un análisis cualitativo de las palabras y hashtags más importantes para la campaña estadounidense. Se ha buscado, en general, que estas palabras sean lo más imparciales posible para que los tuits extraídos no estén muy polarizados, aunque también se han introducido otras que, por el interés de las mismas, son propias de usuarios que son partidarios de un determinado partido.

En segundo lugar, por medio de las librerías anteriormente mencionadas, ha sido posible la descarga de entre unos 15000 y 35000 tuits que contengan una determinada palabra, acotando además las fechas a nuestro parecer. El código utilizado para la extracción se adjunta a este trabajo.

En la siguiente tabla se muestran las palabras y hashtags utilizados para extraer los tuits. También aparecen los perfiles de Twitter de los dos candidatos a las elecciones, Donald Trump y Joe Biden, que se corresponden con tuits que mencionan estos perfiles.



PALABRAS Y HASHTAGS	DESCRIPCIÓN
<i>Trump</i>	Nombre de uno de lo candidatos
<i>Biden</i>	Nombre de uno de los candidatos
<i>vote</i>	Verbo votar en inglés
<i>@realDonaldTrump</i>	Usuario de Twitter de Donald Trump
<i>@JoeBiden</i>	Usuario de Twitter de Joe Biden
<i>#ElectionDay</i>	Hashtag utilizado para hablar de las elecciones estadounidense
<i>#MAGA</i>	Hashtag utilizado especialmente por los seguidores de Donald Trump que hace referencia a <i>Make America Great Again</i>
<i>#MAGA2020</i>	Similar al anterior
<i>#Trump2020</i>	Hashtag utilizado por seguidores y no seguidores de Trump para realizar comentarios sobre él
<i>#Election2020</i>	Hashtag utilizado por los usuarios de Twitter para hablar de las elecciones

Table 1: Campos de los tuits descargados

Como se ha mencionado con anterioridad, se han utilizado dos librerías para la extracción de los tuits. La primera de ellas es *Rtweet*: por medio de ella es posible extraer cualquier cantidad de tuits que contengan una determinada palabra, hashtag o usuario, con la función *search\_tweets*. De esta manera, proporciona como *output* un *dataframe* con un total de 90 atributos para cada tuit descargado, los cuales contienen el texto del propio tuit, el nombre de usuario, la fecha en la que se ha creado, la cantidad de retuits, de favoritos, archivos multimedia, hashtags, etcétera. Muchos de ellos serán utilizados para el posterior análisis. Esta librería, además, nos permite extraer tuits populares: es decir, aquellos que aparecen como destacados cuando se busca una palabra en Twitter. También se han extraído estos por separado para analizar determinados objetivos del trabajo.

Para mayor comodidad, se han exportado los tuits en diferentes archivos *.csv* con el código que se adjunta a este trabajo.

La distribución y cantidad de tuits extraídos por día, *hashtags* y menciones con la librería *Rtweet* se muestra a continuación. Para cada palabra en la que se ha basado la extracción, se tienen tuits correspondientes con el día 2 de noviembre de 2020 y tuits populares que abarcan diferentes días, desde el propio 2 de noviembre

a 15 días previos.

PALABRAS Y HASHTAGS	02/11/2020	Tuits populares (diferentes días)
<i>Trump</i>	10000	160
<i>Biden</i>	10000	118
<i>vote</i>	4975	138
<i>@realDonaldTrump</i>	4405	141
<i>@JoeBiden</i>	4250	118
<i>#ElectionDay</i>	7975	98
<i>#MAGA</i>	3015	62
<i>#MAGA2020</i>	3278	54
<i>#Trump2020</i>	3446	50
<i>#Election2020</i>	6884	100

Table 2: Distribución y cantidad de tuits extraídos por día con *Rtweet*

Otra de las librerías utilizadas ha sido *twitteR*. En comparación con la anterior, ésta proporciona bastante menos atributos que la primera (16 en total), pero se han extraído también varios archivos filtrando los tuits por palabras, como se ha comentado con anterioridad. La mayoría de los atributos importantes continúan presenten con esta librería: favoritos, retuits, hashtags, fecha de creación, texto del tuit, etcétera. La función usada para este propósito es muy parecida a la anterior: *searchTwitter*.

Para poder utilizarla era necesario acceder a la API de Twitter de la forma en la que se expresa en el Estado del Arte de este trabajo: con las credenciales *consumer\_key*, *consumer\_secret*, *access\_token* y *access\_secret* se puede utilizar directamente *searchTwitter*.

De esta librería se ha utilizado la función *userTimeline*, consiguiendo así los últimos tuits de Joe Biden y de Donald Trump, lo cual es importante para el análisis posterior.

Además, con ella se han obtenido los archivos en formato *.json*, pero se han transformado a un *dataframe* y, posteriormente, también se han exportado como archivos *.csv* para trabajar con ellos con mayor facilidad.

La distribución y cantidad de tuits extraídos por día, *hashtags* y menciones con esta librería se muestran a continuación, abarcando los días 1, 2 y 3 de noviembre como se ha mencionado con anterioridad.

PALABRAS Y HASHTAGS	01/11/2020	02/11/2020	03/11/2020
<i>Trump</i>	10000	10000	10000
<i>Biden</i>	10000	10000	10000
<i>vote</i>	5000	5000	5000
@realDonaldTrump	5000	5000	5000
@JoeBiden	5000	5000	5000
#ElectionDay	8000	8000	8000
#MAGA	5000	5000	5000
#MAGA2020	5000	5000	5000
#Trump2020	5000	5000	5000
#Election2020	8000	8000	8000

Table 3: Distribución y cantidad de tuits extraídos por día con *twitteR*

La cantidad de tuits de los usuarios de Twitter de Donald Trump y Joe Biden también se exponen en la siguiente tabla.

USUARIO	TUITS DEL <i>TIMELINE</i>
@realDonaldTrump	299
@JoeBiden	422

Table 4: Cantidad de tuits extraídos de los usuarios de Trump y Biden

Se ha obtenido un total de 54 archivos *.csv* porque cada vez que se procedía a la exportación de tuits esto solo se podía realizar en cantidades menores de 18000. Así, por ejemplo, para la extracción de los tuits que contienen la palabra *trump* se han obtenido un total de 5 archivos: uno de ellos con los tuits destacados que contenían esta palabra, otro con los últimos tuits desde el momento de la extracción que la contenían (para estos dos primeros conjuntos de datos se ha utilizado *Rtweet*) y tres archivos más con la librería *twitteR* filtrados por días diferentes: los días 1 de

noviembre, 2 de noviembre y 3 de noviembre del año 2020.

Cabe destacar que en los diferentes archivos de los tuits extraídos en base a un mismo término no existen elementos duplicados, pues se han extraído tuis de diferentes días y, aunque hay dos archivos correspondientes al 2 de noviembre, no coinciden las horas de extracción de los mismos. De todos modos, para estar seguros, se ha realizado en el código de *Rstudio* la función *unique()*.

De esta forma, como se tiene un total de diez palabras y hashtags en los que se basa este trabajo, se obtienen 50 archivos (5 por término). Además, se han extraído con *Rtweet* cuatro archivos adicionales con los *timeline* de Donald Trump y Joe Biden, con la función descrita anteriormente.

### 3.2. Presentación de los datos

En esta parte del trabajo se presentan los diferentes archivos utilizados para el mismo junto a los atributos más importantes de ellos. Cabe diferenciar, principalmente entre dos tipos de archivos: aquellos que han sido extraídos con la librería *twitter* y aquellos extraídos con *Rtweet*. Hay claras diferencias, como se ha comentado anteriormente, entre estos datos: la primera librería proporciona *dataframes* que cuentan con 90 atributos, mientras que los de la segunda tienen 16. Si bien los atributos más importantes continúan estando, algunos de ellos cambian de nombre, y es importante distinguir este hecho. Veamos dos tablas resumen de los atributos de cada una de estas librerías.

CAMPO	DESCRIPCIÓN
created_at	Fecha y hora en la que ha sido creada el tuit
text	Texto del tuit
source	Dispositivo desde el que ha sido enviado el tuit
favourite_count	Número de favoritos que tiene el tuit
retweet_count	Número de retuits que tiene el tuit
hashtags	Hashtags que contiene el tuit (si no tiene, el campo es nulo)
media_type	Tipo de archivo multimedia que tiene el tuit (si no tiene, el campo es nulo)

Table 5: Campos de los tuits correspondientes a la librería *twitter*

CAMPO	DESCRIPCIÓN
created	Fecha y hora en la que ha sido creada el tuit
text	Texto del tuit
source	Dispositivo desde el que ha sido enviado el tuit
favouriteCount	Número de favoritos que tiene el tuit
retweetCount	Número de retuits que tiene el tuit

Table 6: Campos de los tuits correspondientes a la librería *Rtweet*

### 3.3. Análisis de los objetivos

Esta parte se estructura en función de los objetivos del presente trabajo. Cada apartado está dedicado a un determinado análisis de datos correspondiente con el cumplimiento de los objetivos descritos en el punto uno.

#### 3.3.1. Seguimiento diario de mensajes en Twitter

Se han cargado los datos obtenidos con las librerías anteriores. Se ha trabajado con los diferentes días de la campaña por separado, con todos los archivos de ese día juntos, y se ha analizado cuáles son las palabras más mencionadas de cada día, creando una nube de palabras para mostrar de forma visual cuáles son estas, así como los perfiles a los que más referencia hacen.

Para este propósito, aparte de las librerías anteriormente nombradas, se ha utilizado *wordcloud* para las representaciones en forma de nubes de palabra.

Tras la carga efectiva de cada uno de los tuits utilizados para este análisis, se ha utilizado únicamente el atributo *text* de los mismos para conseguir este objetivo, pues con el texto del tuit es posible ver qué palabras son las más frecuentes para los tuits que contienen una palabra en concreto.

Una vez almacenados en un *array* el cuerpo de cada tuit, se han eliminado todos los signos de puntuación, números, *links* y retuits para obtener únicamente las palabras individualmente de cada dato. Además, por medio de la librería anteriormente mencionada se han conseguido eliminar aquellas palabras que carecían de significado (*stopwords*) tanto en inglés como en castellano. También se han eliminado duplica-

dos, pues en los diferentes archivos extraídos por día ha podido aparecer el mismo tuit en más de una ocasión, evitando así redundancias. Además, en determinados casos, se ha tenido que codificar en UTF-8 determinadas palabras que no aparecía correctamente.

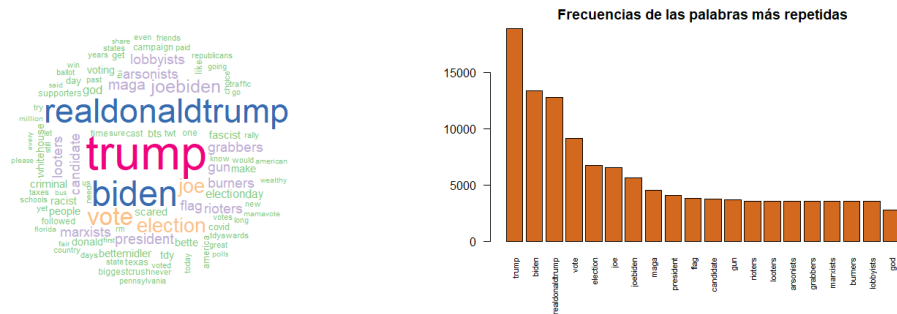


Figure 3: Nubes de palabra y frecuencias de las mismas para el día 01/11/2020 con la librería *twitteR*

Fuente: elaboración propia

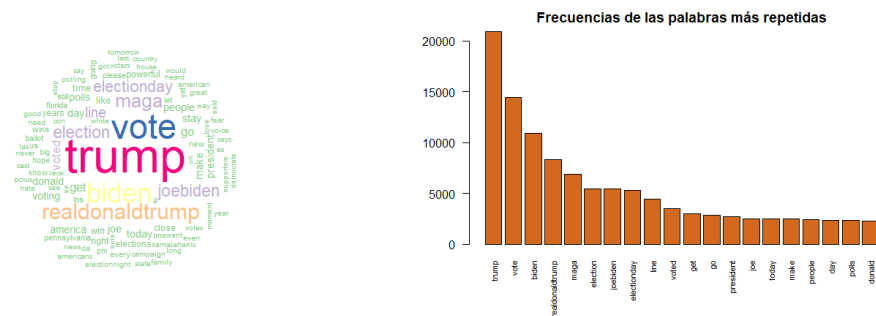


Figure 4: Nubes de palabra y frecuencias de las mismas para el día 02/11/2020 con la librería *twitteR*

Fuente: elaboración propia

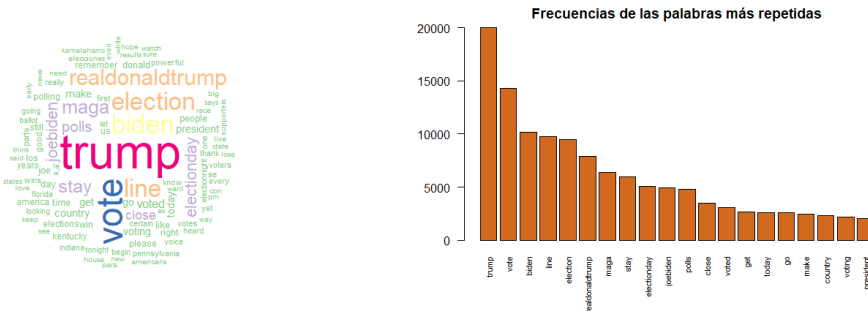


Figure 5: Nubes de palabra y frecuencias de las mismas para el día 03/11/2020 con la librería *twitteR*

Fuente: elaboración propia

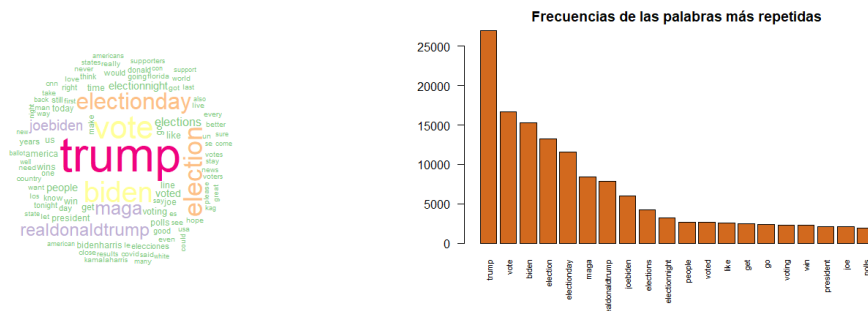


Figure 6: Nubes de palabra y frecuencias de las mismas para el día 02/11/2020 con la librería *Rtweet*

Fuente: elaboración propia

Como se aprecia, las palabras más repetidas en todos los días son *trump*, *biden* y *vote*. Este resultado es totalmente lógico, pues estas palabras son precisamente las que se han utilizado para el proceso de la extracción. Por tanto, el hecho de que estas sean las palabras más repetidas no es del todo relevante, aunque sí es importante el hecho de que unas de ellas aparezcan en mayor proporción que otras.

Finalmente, se ha juntado la información de todos los archivos para ver, en general, qué palabras son las más repetidas a lo largo de todos los días. El resultado del mismo se ve a continuación.





### 3.3.2. Diferencia de interacciones entre partidos y líderes

Una vez que se ha obtenido, como era lógico, que los perfiles más importantes son los de Donald Trump y Joe Biden en esta campaña electoral, así como que el *top 5* de las palabras más mencionadas en general son *trump*, *vote*, *biden*, *@realDonaldTrump* y *election*, esta parte del trabajo se centra en ver cuántas interacciones, de media, tienen los tuits que mencionan a uno y otro candidato, y también las interacciones de los perfiles de Twitter de los mismos. Se considerará como interacción cualquier retuit o favorito que tenga el tuit. Para conseguir este objetivo, se tienen en cuenta los atributos descritos en las tablas 5 y 6 de los puntos anteriores: *favourite\_count* y *retweet\_count* de la librería *twitteR*; y *favouriteCount* y *retweetCount* de *Rtweet*.

Se adjuntan tres gráficos para conseguir este propósito: en el primero de ellos se aprecia la media de interacciones de los tuits que contienen la palabra *Trump* junto a la media de las interacciones que contienen *Biden*.

Para realizar esta media se han sumado los atributos *favourite\_count*, *retweet\_count*, *favouriteCount* y *retweetCount* de los ficheros extraídos en base a las palabras "Trump" y "Biden" y se ha dividido entre cada una de las longitudes de los ficheros.

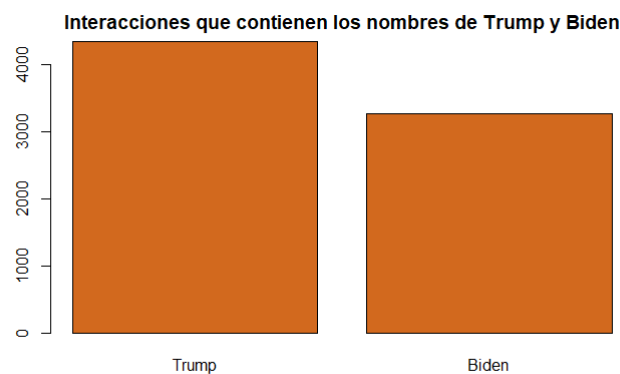


Figure 9: Interacciones de los tuits que contienen los nombres de los candidatos

Fuente: elaboración propia

En el segundo, se representan la media de interacciones de los tuits que contienen

las menciones a los perfiles de los dos candidatos. Para ello, se han sumado los atributos asociados a los retuits y favoritos de estos ficheros y se ha dividido entre la longitud de los mismos.

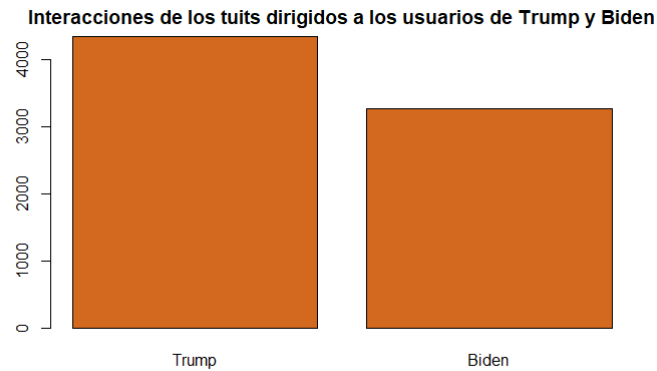


Figure 10: Interacciones de los tuits que mencionan los perfiles de los candidatos

Fuente: elaboración propia

En el tercero, se representan las interacciones, de media, que tienen los tuits de los perfiles de ambos candidatos. Para ello, se ha sumado los atributos asociados los retuits y favoritos de los ficheros que contienen los tuits de los *Timelines* de los mismos y se ha dividido posteriormente entre su longitud.

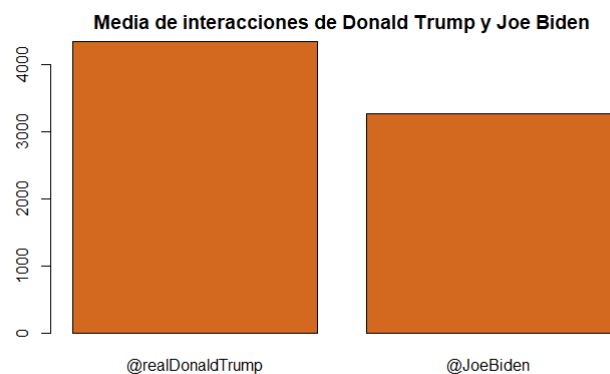


Figure 11: Interacciones de los tuits de los perfiles de los candidatos

Fuente: elaboración propia

Resulta también interesante ver con qué frecuencia se menciona a cada uno de los candidatos en Twitter, tanto su nombre como su perfiles (es decir, ver cuántos tuits por segundo se publican en la red social con los nombres de Donald Trump y Joe Biden, más allá de la cantidad de interacciones que tienen estos). Para exponer estos datos, se ha recurrido al atributo correspondiente a la fecha y hora a la que ha sido creado cada tuit: *created\_at* de la librería *twitteR* y *created* de *Rtweet*. En este punto se han tenido en cuenta tanto los tuits como los retuits obtenidos.

La forma a proceder para conseguir los tuits y retuits por segundos ha sido la siguiente: para cada archivo descargado durante la extracción, se ha obtenido el tiempo que pasa entre el primer y el último dato (primer y último tuit) del mismo. Para ello, se ha tenido que modificar con el código adjunto a este trabajo el atributo correspondiente a la fecha y hora para obtener los segundos que pasan entre el primero y el último tuit de la lista, pasando tanto los días, como las horas y minutos a segundos, para así restar el tiempo asociado al último tuit con el tiempo asociado al primero. Por tanto, dividiendo entre el número de datos del fichero, se consiguen los tuits y retuits por segundo correspondientes a este archivo. Como se tienen diferentes archivos con tuits que mencionan a ambos candidatos, se ha calculado la media de tuits y retuits por segundo de todos ellos, obteniendo los resultados que se exponen a continuación.

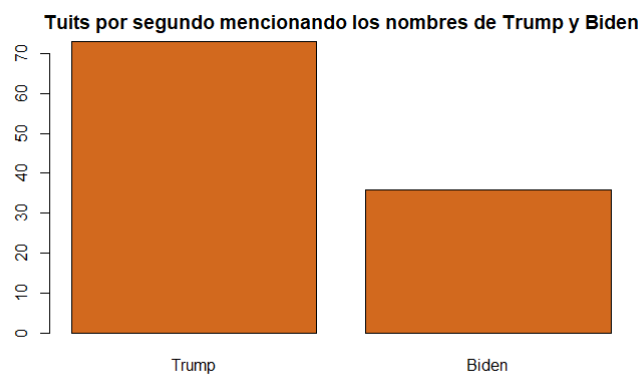


Figure 12: Tuits y retuits por segundo que contienen *Trump* y *Biden*

Fuente: elaboración propia

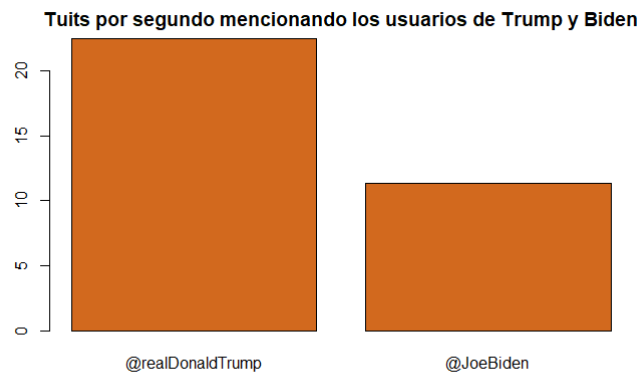


Figure 13: Tuits y retuits por segundo que contienen *@realDonaldTrump* y *@JoeBiden*

Fuente: elaboración propia

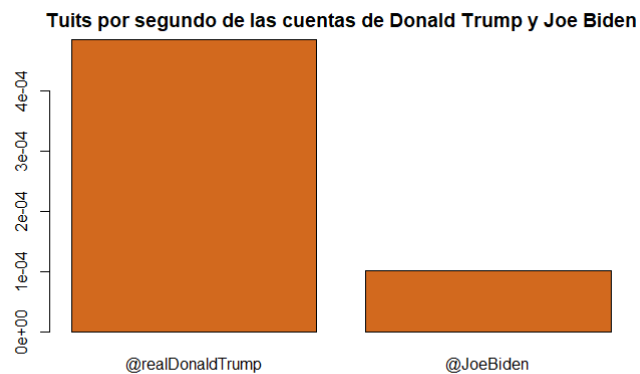


Figure 14: Tuits y retuits por segundo publicados por las cuentas *@realDonaldTrump* y *@JoeBiden*

Fuente: elaboración propia

### 3.3.3. Interacciones de las palabras más frecuentes

En esta parte del trabajo se intenta ver qué tuits tienen un mayor número de interacciones, dependiendo de las palabras que contengan. Por tanto, se observará, por ejemplo, si los tuits que mencionan a un candidato u otro tienen más o menos repercusión en Twitter. Cabe dejar claro que el hecho de publicar un tuit con

una determinada palabra no hará directamente que el tuit vaya a tener una mayor repercusión; en este apartado se intenta ver cuántas interacciones tienen los tuits dependiendo de las palabras que contengan.

Para el análisis de este objetivo, se ha tenido en cuenta tan solo los tuits populares obtenidos en la extracción con la librería *Rtweet*, trabajando así con un total de diez ficheros (uno por cada palabra y *hashtag*). Los tuits destacados son generalmente de perfiles de Twitter influyentes, como los propios candidatos (Trump y Biden), Barack Obama, etcétera, quienes son, por lo general, los que más interacciones tienen y, por tanto, los más útiles para analizar esta parte del trabajo.

No importa ya tanto distinguir el análisis por ficheros, por lo que se ha procedido a realizar el análisis con los tuits populares de los diez ficheros mencionados. De la misma forma que en el apartado anterior, se han considerado interacciones tanto los retuits como los favoritos, creando de esta manera una nueva variable llamada *interactions* para cada fichero que resulte de la suma de ambas cantidades.

Una vez unidos todos los ficheros de la forma en la que aparece en el código adjunto a este trabajo, se ha contado el número de interacciones por palabra. Se ha partido de las palabras con un mayor número de frecuencia de aparición obtenidas en el primer apartado de este análisis con las nubes de palabras, guardando las veinte palabras más populares de cada fichero para luego utilizarlas para este apartado.

Con la función *grep* del paquete *base* de *Rstudio* es posible localizar si una palabra está presente en un dataframe o no. Por medio de esta función e iterando con determinados bucles se ha conseguido obtener un dataframe de palabras populares con la frecuencia con la que aparecen en el total de ficheros y sus interacciones totales. La cantidad de interacciones por palabra se observa en el gráfico adjunto.

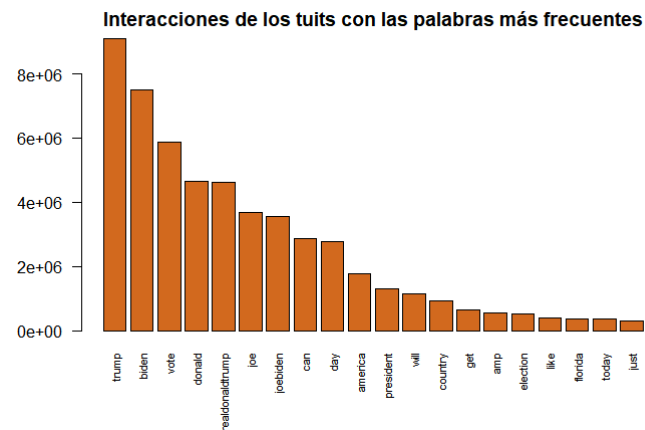


Figure 15: Interacciones de los tuits de las palabras más frecuentes

Fuente: elaboración propia

Igual o más importante que el número de interacciones es el número de interacciones por tuit de cada palabra, pues es todavía más significativo para analizar el presente objetivo. Por tanto, dividiendo la información del gráfico anterior entre el número de tuit en los que aparece cada palabra es posible obtener la frecuencia relativa de las mismas; es decir, el número de interacciones por tuit, el cual se muestra en la siguiente imagen.

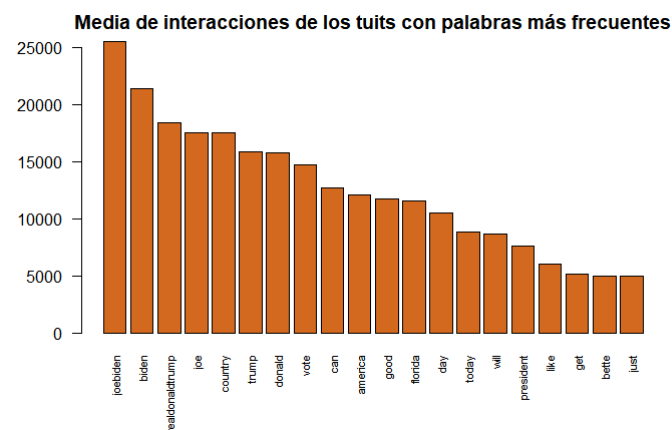


Figure 16: Media de interacciones por tuit de las palabras más frecuentes

Fuente: elaboración propia

### 3.3.4. Análisis de los tuits con contenido multimedia y hashtags

Uno de los objetivos de este trabajo es ver si los tuits con contenido multimedia tienen más o menos interacciones de media. Para el análisis de este objetivo se han utilizado, como en el apartado anterior, todos los tuits populares en conjunto. La librería *twitteR*, al extraer los tuits, nos proporciona el atributo *media\_type* descrito en la Tabla 5. En el caso en el que un tuit no tenga contenido multimedia, este atributo permanece vacío, por lo que para conseguir este objetivo ha bastado con separar la base de datos en dos: tuits que tienen contenido multimedia y tuits que no lo tienen, y ver la media de interacciones de cada una.

Para obtener el número de interacciones se han sumado las columnas asociadas a los retuits y favoritos de cada uno de los dos *dataframes* obtenidos con el atributo *media\_type* y se ha dividido entre las longitudes de los mismos. El resultado se muestra a continuación.

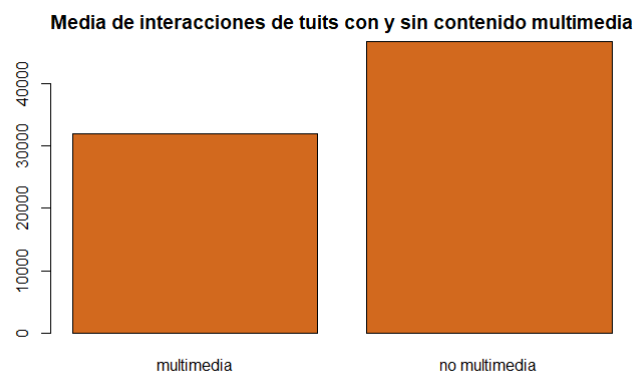


Figure 17: Interacciones de tuits con contenido multimedia y sin él

Fuente: elaboración propia

Del mismo modo, existe la pregunta de si un tuit que contiene *hashtags* tiene más interacciones, en general, que un tuit que no contenga. Para resolver esta cuestión se recurre al atributo *hashtags* de la base de datos, que permanece vacío mientras que un tuit no tenga este contenido. Procediendo de forma similar que con el contenido multimedia, se obtiene la siguiente imagen.

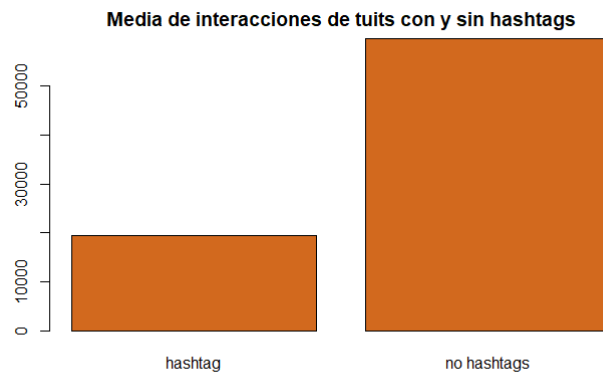


Figure 18: Interacciones de tuits con *hashtags* y sin ellos

Fuente: elaboración propia

Las conclusiones relativas a este subapartado se realizará en el punto 4 (Análisis de los resultados).

### 3.3.5. Dispositivo desde el que han sido enviados los tuits

Los archivos obtenidos en la extracción cuentan con el atributo *source* que, como se explica en las Tablas 5 y 6, hacen referencia al dispositivo desde el que ha sido enviado un determinado tuit. Este atributo puede tener, básicamente, cuatro valores posibles: *Twitter for iPhone*, *Twitter Media Studio*, *Twitter Web App* y *Twitter for Android* que hacen referencia a si el tuit ha sido enviado desde un iPhone, desde la Web, desde la app del ordenador o desde un dispositivo *Android*.

De forma similar a los puntos anteriores, se ha utilizado el fichero correspondiente a los tuits populares, dividiéndose éste en cuatro *dataframes* diferentes; uno para cada tipo de *source* que tengamos. Creando una variable en los mismos correspondiente a la cantidad total de interacciones de cada tuit, teniendo en cuenta tanto los favoritos como los retuits, y sacando la media de cada archivo de la misma forma que en los puntos anteriores, se ha obtenido el gráfico siguiente.



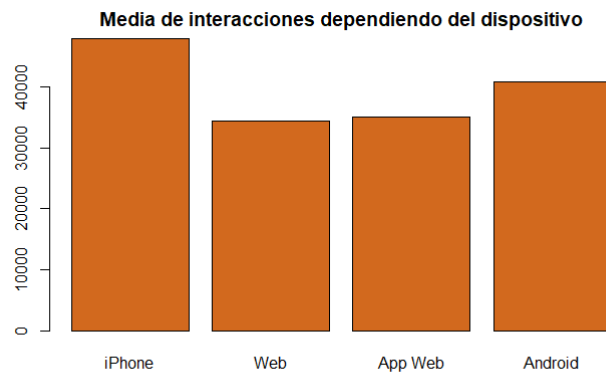


Figure 19: Interacciones de tuits dependiendo del dispositivo

Fuente: elaboración propia

### 3.3.6. Análisis de sentimientos de los usuarios de Twitter

Este último apartado del análisis de los datos está dedicado a observar si los usuarios de Twitter que opinaban sobre la campaña electoral lo hacían desde un punto de vista neutral o si las opiniones estaban polarizadas. Para ver esto, se recurre al análisis de sentimientos de los tuits. El paquete *tidytext* de *Rstudio* proporciona diccionarios de palabras que evalúan el nivel de sentimiento de las mismas, dando valores a estas palabras entre -5 y 5, siendo -5 el máximo de negatividad y +5 el máximo de positividad. Por tanto, una forma de analizar los sentimientos de un tuit es viendo el sentimiento de cada una de las palabras del mismo y realizando una media.

En primer lugar se han dispuesto los datos en formato *tidy*, con una palabra por fila de los datos (separando cada tuit en tantas palabras como tenga). Posteriormente, con la librería *tidytext* se ha evaluado el sentimiento de cada palabra de los archivos y mediante la función *inner join* se obtienen únicamente las palabras de los tuits que están en el diccionario utilizado.

Se han analizado, de la misma forma que en apartados anteriores, los tuits por días de publicación, para intentar ver, de esta manera si, conforme avanzan los días en la campaña electoral, los sentimientos de los usuarios de Twitter cambian. La

representación de los tuits por sentimiento y días se adjunta en las siguientes figuras.

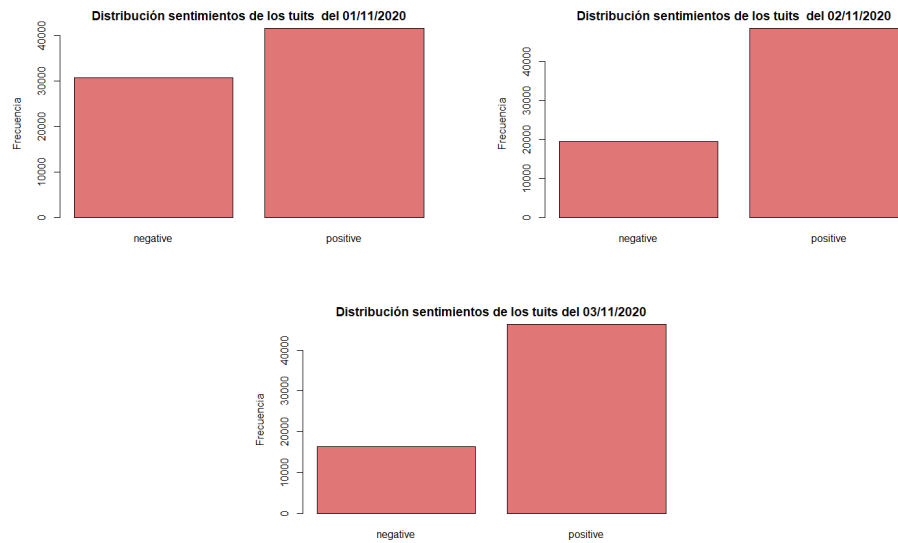


Figure 20: Sentimientos de los tuits por días del análisis

Fuente: elaboración propia

Se observa cómo, conforme pasan los días, los comentarios negativos disminuyen en general. De esta forma, el día 01/11/2020 los comentarios positivos se corresponden con un 57.49 %, mientras que en el 02/11/2020 se corresponden con el 71.45 % y el 03/11/2020 con el 73.91 %. No hay tanta diferencia si se calcula el tanto por ciento de los tuits totales que tienen una etiqueta de sentimiento para cada uno de los días, obteniéndose los siguientes porcentajes: 6.95 %, 6.86 % y 6.33 %, respectivamente.

Es posible ahora ver la distribución de los tuits anteriores dependiendo del término escogido para la extracción de los tuits, sobre todo para ver si los *hashtags* que se asocian a uno u otro candidato, como *#MAGA* o *#Trump2020* hacen que los tuits contengan más palabras de sentimiento que los que no los contienen.

Para los tuits que contienen la palabra *trump* y que mencionan su perfil de Twitter, si comparamos la cantidad de palabras que tienen sentimiento con el total de las mismas, se obtiene que el 11.22 % de las mismas tienen una etiqueta con sentimiento. Para los tuits que nombran el usuario de Trump, el 7.33 %. La proporción

de palabras positivas y negativas es la siguiente.

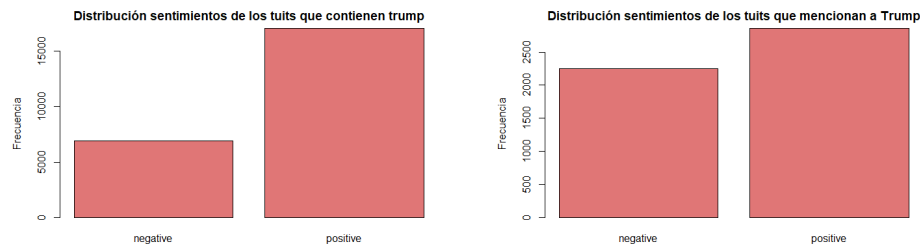


Figure 21: Sentimientos de los tuits que mencionan el nombre de Trump y que nombran su perfil de Twitter

Fuente: elaboración propia

Para los tuits que contienen la palabra *Biden*, se obtiene que el 7.73 % de las mismas tienen una etiqueta con sentimiento. Además, para los tuits que nombran el usuario de Biden, el 6.91 % la contienen. La distribución es la siguiente.

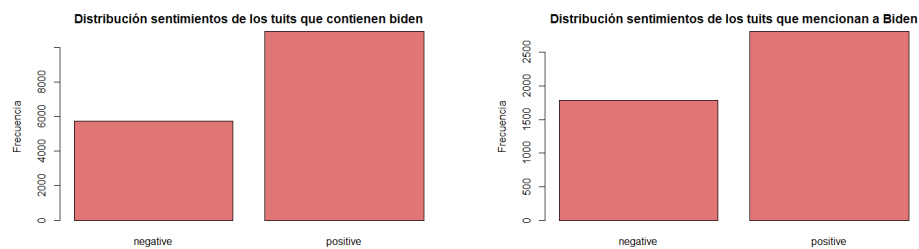


Figure 22: Sentimientos de los tuits que mencionan el nombre de Biden y que nombran su perfil de Twitter

Fuente: elaboración propia

En las anteriores figuras se aprecia que, en general, hay un mayor número de palabras con sentimientos positivos que negativos. Esta diferencia se acentúa si analizamos los tuits que contienen, por ejemplo, el *hashtag* *#MAGA*, el cual es utilizado principalmente por los seguidores de Donald Trump. Todavía son más las palabras positivas que contienen *#Trump2020*, característico también de los partidarios de este candidato.

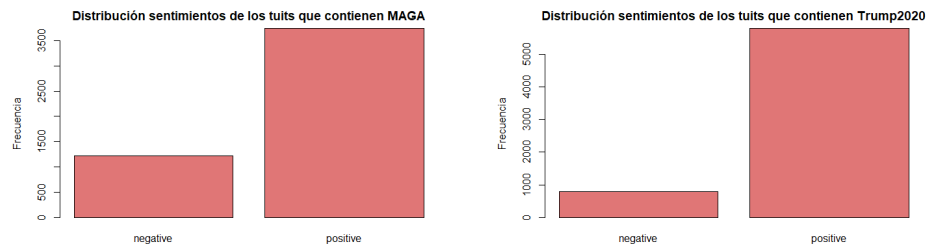


Figure 23: Sentimientos de los tuits que mencionan los *hashtags* *#MAGA* y *#Trump2020*

Fuente: elaboración propia

La cantidad de palabras con sentimiento también crece en estas últimas representaciones, pues el 14.48 % de las palabras totales de los tuits con el *hashtag* *#Trump2020* cuentan con etiqueta de sentimiento, así como el 8.78 % las que contienen *#MAGA*.

## 4. Análisis de los resultados

Este apartado se dedica a comentar los resultados de las figuras expuestas en el análisis de los objetivos. De la misma forma que en las páginas anteriores se comentarán conclusiones individuales para cada uno de los objetivos propuestos inicialmente en este Trabajo Final.

### 4.1. Seguimiento diario de los mensajes en Twitter

Por medio del atributo *text* anteriormente citado, se ha visto en este apartado cuáles son las palabras más mencionadas para cada día del estudio por separado. En la siguiente tabla se muestran las cinco palabras más frecuentes para cada día y librería.

DÍA Y LIBRERÍA	PALABRAS MÁS MENCIONADAS				
01/11/2020 - <i>twitteR</i>	Trump	Biden	@realDonaldTrump	<i>vote</i>	<i>election</i>
02/11/2020 - <i>twitteR</i>	Trump	vote	Biden	@realDonaldTrump	#MAGA
03/11/2020 - <i>twitteR</i>	Trump	<i>vote</i>	Biden	<i>line</i>	<i>election</i>
02/11/2020 - <i>Rtweet</i>	Trump	<i>vote</i>	Biden	<i>election</i>	#electionday

Table 7: Palabras más mencionadas por día

En general, se ve que la palabra más utilizada los días previos a las elecciones es el nombre apellido de uno de los candidatos: Donald Trump. Se ha de recordar que es esperable este resultado, pues ha sido una de las palabras en base a las cuales se han obtenido los datos. La representación más significativa es la correspondiente a la Figura 8, que muestra por medio de un diagrama de barras la frecuencia de las palabras más repetidas en general, obteniendo el siguiente Top 5 de las mismas, con el porcentaje de aparición respecto al total de palabras.

PALABRA	PORCENTAJE (%)
Trump	3.52
<i>vote</i>	2.21
Biden	2.02
@realDonaldTrump	1.49
<i>election</i>	1.42

Table 8: Aparición de las palabras clave

Todas eran bastante predecibles, como se ha comentado con anterioridad, siendo "Trump" la más abundante de todos los tuits, casi duplicando, como mínimo, a todas las demás respecto a apariciones. "Trump" aparece en siete de cada doscientas palabras.

Es también importante ver cuáles son las palabras más repetidas por los usuarios en general entre aquellas que no han sido utilizadas para la extracción. De esta forma, se tienen las siguientes cinco palabras más frecuentes:

PALABRA	PORCENTAJE (%)
<i>line</i>	0.65
Joe	0.53
<i>president</i>	0.44
<i>voted</i>	0.42
<i>polls</i>	0.40

Table 9: Aparición de las palabras clave

La palabra *line* es la más mencionada por los usuarios de Twitter en los tuits correspondientes a la extracción. En el contexto de las elecciones, este término va ligado al verbo *vote* y hace referencia a la disciplina de voto de los partidos políticos. La palabra *Joe* es la segunda más mencionada de este subgrupo, correspondiente al nombre de Biden. Del mismo modo, se tienen en menor proporción los términos *president* (presidente), *voted* (participio del verbo votar) y *polls* (encuestas).

#### 4.2 Diferencia de interacciones entre partidos y líderes

En este subapartado se ha estudiado la diferencia entre las interacciones de Donald Trump y Joe Biden, tanto en lo que respecta a la mención de sus nombres en la red social, como la mención de sus perfiles de Twitter, así como las interacciones que tenían los tuits publicados en sus cuentas personales.

En general, se observa que los tuits que mencionan o nombran a Trump cuentan con más interacciones que los que mencionan o nombran a Biden. El nombre de Biden tiene un 24.85 % menos de interacciones que el de Donald Trump, y los tuits que mencionan su usuario de Twitter un 40.02 % menos que los que mencionan el

perfil @realDonaldTrump.

En lo que respecta a las interacciones de cada uno de los perfiles de Twitter, los tuits que publica Joe Biden tienen un 13.24 % menos de interacciones que los que publica Donald Trump.

Del mismo modo, se ha analizado cuántos tuits por segundo se publican mencionando a uno u otro candidato, llegando a las cantidades de 73 tuits por segundo mencionando el nombre de Trump y 36 tuits por segundo mencionando el de Biden. Es decir, se menciona el doble de veces a Donald Trump que a Joe Biden.

La proporción se mantiene intacta en lo que respecta a la mención de sus perfiles de Twitter: se nombra a @realDonaldTrump 22.5 veces por segundo y a @JoeBiden 11.32 veces.

También hay claras diferencias entre las cuentas de Twitter de ambos candidatos, pues durante la campaña Donald Trump ha publicado una media de 41.878 tuits por día, mientras que Joe Biden apenas 8.703 tuits de media.

#### 4.3. Interacciones de las palabras más frecuentes

Uno de los datos más significativos el trabajo es que, a pesar de que se mencione en menor cantidad que a Donald Trump, los tuits que mencionaban el perfil de Joe Biden eran los que más interacciones tenían de media, seguidos por los que mencionaban la palabra *Biden*, tal y como se muestra en la Figura 15. En tercera posición se tiene al perfil de Twitter de Trump. En la siguiente tabla se resume la media de interacciones por tuit de cada palabra.

PALABRA	INTERACCIONES
@JoeBiden	25500
Biden	21397
@realDonaldTrump	18395
Joe	17507
<i>Country</i>	17482

Table 10: Palabras con más interacciones de la campaña

#### 4.4. Análisis de los tuits con contenido multimedia y hashtags

Por lo general, se observa que los tuits que no tienen contenido multimedia y los que no tienen *hashtags* tienen más interacciones que los que sí lo tienen. Esto no debe interpretarse directamente como que el hecho de poner un *hashtag* o una foto hará que el tuit publicado vaya a tener un menor número de interacciones, pues debe tenerse también en cuenta que los perfiles más importantes, que son los que más interacciones tienen, quizá no utilizan tanto este tipo de recurso.

De todos los tuits populares, la media de interacciones de los tuits con contenido multimedia se sitúa en 32019 interacciones, mientras que la de los tuits que no tienen este contenido en 46813.

La media de interacciones de los tuits con *hashtags* es de 19532 interacciones, mientras que la de los tuits que no tienen *hashtags* es de 59795. Cabe recordar que para la obtención de estos datos se han utilizado únicamente los archivos de los tuits destacados.

#### **4.5. Dispositivos desde el que han sido enviados los tuits**

Como parte complementaria al trabajo, se ha estudiado cuántas interacciones de media tenían los tuits populares dependiendo del dispositivo desde el que estos hayan sido enviado. La gráfica de la Figura 18 responde a esta pregunta, obteniendo que los tuits enviados desde *iPhone* cuentan con una media de 47996 interacciones, mientras que los tuits enviados desde la web con 34492 interacciones de media, los enviados desde la App Web con 35130 y los enviados desde Android con 40932.

Del mismo modo que se ha realizado la consideración en el apartado anterior, estos datos de interacciones no deben asociarse directamente a que el tipo de dispositivo desde el que se envía un tuit hace que éste tenga mayor repercusión, sino a que las personas más influyentes en Twitter publican con más frecuencia desde *iPhone*.

#### **4.6. Análisis de sentimientos de los usuarios de Twitter**

En este último apartado se ha intentado analizar los sentimientos de los usuarios de Twitter que hablaban de la campaña electoral estadounidense conforme ésta avanzaba. De todos los días estudiados, se aprecia que, conforme se acercaba el día



de las elecciones, la cantidad de palabras negativas decrecía respecto a las positivas, teniendo un 42.51 % de comentarios negativos el día 01/11/2020, un 28.55 % el día 02/11/2020 y un 26.09 % el día 03/11/2020, justo antes de comenzar las elecciones. En cambio, no se aprecian diferencias en estos días en los porcentajes que representan los tuits con sentimientos de los tuits totales, con cantidades entre el 6 % y el 7 % en todos estos días.

En lo que respecta a la distribución de sentimientos, separando por diferentes palabras clave o *hashtags*, los usuarios cuyos sentimientos han sido más destacados son los correspondientes al *hashtag* *#Trump2020*, pues el 14.48 % de las palabras de estos archivos tenían sentimientos, bien positivos, bien negativos. Los archivos con un menor número de sentimientos de sus tuits son los que mencionaban el perfil de Joe Biden: el 6.91 % de las palabras de estos tenían un sentimiento definido.

Además, el *hashtag* *#Trump2020* era la palabra que mayor proporción de sentimientos positivos respecto a los negativos cuenta, pues los negativos representan el 11.91 % del total, mientras que los positivos el 88.09 %.

De esta forma, podemos objetar que los tuits con *hashtags* y palabras asociados a un candidato concreto (como *#Trump2020* y *#MAGA*) hacen que haya una mayor diferencia entre tuits de carácter positivo y negativo, y también hacen que el porcentaje de tuits con sentimientos sea mayor. Este último porcentaje es una medida de la polarización de los usuarios, pues se entiende que, cuanto menos palabras neutras tengan los tuits, menos polarizados estarán estos.

## 5. Conclusiones

En este Trabajo Final se han intentado seguir las opiniones en Twitter de la campaña electoral estadounidense de 2020, con la mirada puesta dar respuestas a los objetivos iniciales del mismo. Una vez revisados trabajos previos, así como las relaciones de la red social con los procesos electorales, descritos en el estado del arte, se ha procedido a la familiarización con la API de Twitter, la cual es la primera herramienta a la que se recurrió para el proceso de extracción de los tuits.

La extracción de los datos ha sido el primer proceso de la implementación del trabajo, el cual se ha llevado a cabo de una forma satisfactoria, pues los diferentes términos utilizados para filtrar los tuits correspondientes a la campaña electoral han servido para llevar a cabo las demás partes del trabajo. Además, la extracción de los tuits de los dos principales aspirantes a la presidencia de EEUU (Donald Trump y Joe Biden) también ha sido útil, pues posteriormente se ha observado que, evidentemente, estos eran los perfiles a los que más iban dirigidos los tuits.

El hecho de que se hayan utilizado dos librerías diferentes para este primer apartado (*twitteR* y *Rtweet*) también ha sido positivo para el trabajo, pues con ambas se han alcanzado los objetivos iniciales del mismo, algo que de forma individual no se habría logrado.

Los *hashtags*, palabras y perfiles de Twitter utilizados para filtrar los tuits de toda la red social han estado activos durante todo el proceso de extracción, haciendo que los datos descargados durante los días previos a las elecciones fueran variados y se actualizaran rápidamente para descargar una mayor cantidad de los mismos y poder tener una amplia base de datos desde la que trabajar.

Una vez concluido este primer proceso, se han explorado los atributos que las librerías *twitteR* y *Rtweet* proporcionan, presentados en las Tablas 5 y 6. Todos ellos han sido útiles para alcanzar cada uno de los objetivos del trabajo, siendo el atributo *text* de los mismos el principal con el que se ha trabajado, pues corresponde al texto de los tuits y a partir de él se han obtenido las palabras y los perfiles clave de la campaña. Otros atributos, como *source*, *hashtags* o *media.type* han tenido un papel más complementario, pero también han servido para tener una idea general

de cómo funcionan las interacciones en Twitter.

La parte principal del trabajo ha sido el análisis de cada uno de los objetivos fijados en la primera parte del mismo. Ésta ha sido dividida en tantas partes como objetivos iniciales se fijaron al inicio, y en ella se han adjuntado las representaciones adecuadas con ayuda *Rstudio* que daban respuesta a dichos objetivos.

Las conclusiones directas que se pueden extraer del análisis se enumeran a continuación.

- Las palabras Trump y *vote* son las que aparecen con mayor frecuencia en los datos extraídos. Aunque estas hayan sido utilizadas para el proceso de extracción y sea lógico que aparezcan frecuentemente, sí es importante el hecho de que sean la primera y la segunda más mencionada, así como que la palabra Trump se mencione casi el doble de veces que otras.
- Entre las palabras que no han sido utilizadas para el proceso de extracción, la palabra *line* es la más mencionada por los usuarios de Twitter. El segundo lugar lo ocupa "Joe", nombre de uno de los candidatos a la presidencia.
- Los tuits que mencionan a Trump tienen un mayor número de interacciones que los que mencionan a Biden. En cambio, los tuits publicados por Joe Biden tienen una mayor repercusión que los publicados por Donald Trump.
- En el momento de la extracción, se mencionó el doble de veces más a Donald Trump que a Joe Biden (se publicaron el doble de tuits por segundo). Además, Trump publicó cinco veces más que Biden durante la campaña.
- Los tuits que cuentan con contenido multimedia tienen un menor número de interacciones que los que no cuentan con él. Del mismo modo, los tuits que contienen *hashtags* tienen menos interacciones que los que no los contienen.
- Los tuits que fueron enviados desde *iPhone* cuentan con un mayor número de interacciones que los enviados con otros dispositivos.

- Los tuits con *hashtags* y palabras asociadas a un candidato concreto hacen que el porcentaje de tuits con sentimientos sea mayor que el de los tuits con *hashtags* y palabras neutrales.

## 6. Líneas de trabajo futuras

Para la ampliación de este trabajo es posible recurrir a otras funciones de las librerías utilizadas para la extracción. Por ejemplo, con la función *location()* es posible la obtención de las coordenadas geográficas desde la que han sido enviados los tuits. De esta forma, se podrá ver si en determinadas zonas de Estados Unidos se publican tuits en favor de un candidato u otro, y ver si estas zonas están asociadas a los feudos demócratas y republicanos (estados que tradicionalmente han votado a favor de un determinado candidato).

Además, es posible recurrir al análisis de grafos para la detección de diferentes comunidades que publiquen tuits sobre el proceso electoral. De esta forma, es posible desarrollar un estudio sobre las posiciones de los usuarios de Twitter a favor de uno u otro candidato. El objetivo de este análisis podría ser la identificación dichas comunidades con la ideología o partido político que representan.

Esto es posible realizarlo principalmenete de dos formas: con el programa *Gephi* es posible extraer "en directo" los usuarios que interaccionan con los tuits que contienen un determinado *hashtag* o palabra, de la misma forma que se ha hecho en este trabajo. Los nodos serían los diferentes usuarios de Twitter que han publicado un determinado tuit, y dos nodos estarán conectados si, por ejemplo, han retuiteado un mismo tuit.

La segunda de las formas es realizar la extracción con *Rstudio* o *python* con las librerías *twitteR*, *Rtweet* o *Tweepy* y posteriormente se deberán transformar los datos para que *Gephi* pueda leerlos correctamente.

## 7. Bibliografía

- [1] Abejón, P., Sastre A. y V. Linares, (2012) "Facebook y Twitter en campañas electorales en España". Anuario Electrónico de Estudios en Comunicación Social, volumen 5, número 1. Universidad Complutense de Madrid, Universidad de los Andes. Disponible en: <http://erevistas.saber.ula.ve/index.php/Disertaciones/>
- [2] Alvírez, S.; Franco-Rodríguez, O. (2016). "Estilo comunicativo súbito en Twitter: efectos sobre la credibilidad y la participación cívica". Comunicar, v. 24, N 47, 89-97. <https://doi.org/10.3916/C47-2016-09>
- [3] Anagnostopoulos A.; Jumar R.; Mahdian M. (2008) "*Influence and Correlation in Social Networks*", Yahoo! Research.
- [4] Bernard J. Jansen, Mimi Zhang, Kate Sobel and Abdur Chowdury, (2009) "Twitter power: Tweets as electronic word of mouth". *Journal of the American society for information science and technology*, 60(11):2169-2188.
- [5] Campos Freire, F. (2008): "Las redes sociales trastocan los modelos de los medios de comunicación tradicionales". Revista Latina de Comunicación Social, N 63, 287-293.
- [6] Enli, G. (2017). *Twitter as arena for the authentic outsider: exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election*. European Journal of Communication Vol 32, Issue 1, p. 50 – 61. Prepublicacion.
- [7] Gayo-Avello D. (2013) *A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data*. Social Science Computer, 31(6) 649-679.
- [8] Gayo-Avello D. (2012) *No, you cannot predict election with Twitter*. IEE Computer Society, November/December.
- [9] Lee E., Youn Oh S. (2012) "*To Personalize or Depersonalize? When and How Politicians' Personalized Tweets Affect the Public's Reactions*", Journal of Communication, N 62, 932-949. doi:10.1111/j.1460-2466.2012.01681.x
- [10] Matthew A. Russel, (2013) "*Mining the Social Web. Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub*", 2 Edición. O'Reilly Media Inc.

[11] Pressgrove, G., Carolyn, K. (2018) *Stewardship, credibility and political communications: A content analysis of the 2016 election*. Public Relations Review. MorganTown, Elsevier.

[12] Stephenson A. (2019) "*Australian Election Analysis (Textual Analysis of Tweets)*", artículo de Rpubs: [https://rpubs.com/alex\\_stephenson/510788](https://rpubs.com/alex_stephenson/510788)