Grado en Estadística

TRABAJO FIN DE GRADO

Análisis de Sentimientos con Twitter: Turismo y Política Electoral

Realizado por: Rafael Sánchez del Hoyo

Dirigido por: Inmaculada Barranco Chamorro

Sevilla, Junio de 2019

Índice general

Prólogo

El trabajo de fin de grado presentado a continuación, tiene el nombre de "Análisis de Sentimientos con Twitter: Turísmo y Política Electoral". El período de investigación y redacción de este trabajo de fin de grado ha durado desde octubre de 2018 hasta junio de 2019. El trabajo lo he llevado a cabo debido a la gran utilidad que tiene explotar los datos que se generan gracias a las redes sociales, concretamente Twitter. A pesar de haber tenido infinidad de problemas a la hora de cargar y manejar los tweets, finalmente he sabido solucionarlo.

Me gustaría dar las gracias a mi tutora por su buena orientación y paciencia durante el proceso de realización del proyecto. También me gustaría agradecer a Drimay Consultores S.L. en concreto a Mónica Rivera y a Manuel Martínez, por su gran aportación a la hora del aprendizaje del manejo de grandes masas de datos. Y por último, agradecer a Luis Valencia, profesor de Inteligencia Artifical por su aportación en dicha asignatura que ha sido de gran utilidad para la realización del trabajo, y a Pedro Luque por poner a disposición del alumnado una herramienta para la facilitación de la creación del TFG.

Rafael Sánchez del Hoyo

Resumen

Hoy en día, gran parte de la población hace uso de al menos una red social para interactuar con otras personas. Realmente estas plataformas no sólo las forman individuos sino que también las integran pequeñas y grandes empresas que encuentran en ellas una forma de promocionarse.

En este trabajo nos vamos a centrar en la red social "Twitter" que tiene aproximadamente 500 millones de usuarios generando 65 millones de tweets al día.

Twitter nos puede proporcionar valiosa información sobre temas muy diversos de la sociedad actual y, sobre todo, la opinión de sus usuarios respecto de esos temas, por ello consideramos de gran interés realizar un análisis de sentimientos con esta red social.

El trabajo consta de tres capítulos:

Capítulo I: Explicamos por qué se utiliza Twitter, las nociones básicas para su entendimiento y conceptos básicos para facilitar la comprensión de este trabajo.

Capítulo II: Nos centramos en el turismo por ser el gran motor de nuestra economía. Realizamos un análisis estadístico y de sentimientos en el que se comparan distintas ciudades del país en relación con el turismo. Nos enfocamos en la ciudad de Sevilla como base del estudio y la comparamos con tres de las ciudades andaluzas más turísticas, Córdoba, Málaga, y Granada, y con las tres ciudades más importantes del resto del país entre las que se encuentran Madrid, Barcelona y Valencia. Explicamos la metodología utilizada.

Capítulo III: Ponemos nuestra atención en otro punto de la actualidad: la política en clave electoral. Analizamos las elecciones del 28 de abril de 2019 contemplando lo que opinan los usuarios de Twitter de los cinco principales candidatos a la presidencia del Gobierno de España. Se desarrolla temporalmente en los días 22 y 23 de abril, días de los principales debates y el propio 28, día de las elecciones. En dicho estudio no se hacen valoraciones políticas.

El trabajo se completa con la bibliografía utilizada.

Abstract

Nowadays, most of people in our society use at least one social media platform in order to interact with other individuals. But those platforms are not only used by individuals, they are also used by small and big businesses and enterprises in order to promote themselves to a wider audience.

In this project we are going to cover the social media called "Twitter" which has nearly 500 millions of users who generate around 65 millions tweets per day.

Twitter can provide us with valuable information about different topics of interest in our society. Especially, the opinion of the users regarding these topics. That is the reason why we consider of major interest to carry out an analysis of sentiments about this social media.

This project has three chapters :

Chapter I: We explain why Twitter is used, basic knowledge for its understanding, and basic concepts to ease the understanding of this project.

Chapter II: We focus on the tourism sector as it is the main source of income in our economy. We perform both a statistical analysis of sentiments, in which we compare different cities of our country regarding the tourism sector. We focused on the city of Seville, as base of our study, and we compare it with the other three most touristic Andalusian cities, which are Córdoba, Málaga and Granada. At the same time, we also compare Seville with the three most important cities of the country, which are Madrid, Barcelona and Valencia. We explained the methodology used.

Chapter III: We focus our attention on another important current topic: sentiments obout politics during previous days and the day of general elections in Spain. We analyze the elections that happened on the 28th of April of 2019, by analysing the opinions of the Twitter users about the five candidates to run the Spanish Government. Our study takes place during the days 22nd and 23rd of April, which are the days in which relevant debates took place, and the day 28th April itself, day of the elections. Political assessments are not expressed in this research.

Relevant bibliography has also been included.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción y metodología

1.1. Introducción a twitter

Twitter es una red social de microblogging que permite escribir y leer mensajes en Internet que no superen los 280 caracteres. El microblogging es un tipo de blog, que se caracteriza la brevedad de sus mensajes y su facilidad de publicación. Esta nueva forma de comunicación, permite a sus usuarios estar en contacto en tiempo real con otras personas a través de mensajes breves de texto por medio de una sencilla pregunta: ¿Qué estás haciendo?.

Twitter fue creada originalmente en California en el año 2006, dicha red social empezó a tomar más y más popularidad e importancia hasta el punto de posicionarse en una de las más importantes. Se estima que actualmente tiene 500 millones de usuarios generando 65 millones de tweets al día.

Por un lado, twitter es una red social en la que se suele opinar, ya que la mayor parte de sus usuarios son personas individuales que no sólo hablan de qué están haciendo, cómo, con quién y por qué, sino que además, también se comenta lo que ocurre en la actualidad. Al ser instantáneo es una buena forma de observar la opinión de la sociedad actual sobre un tema. Por otro lado, también hay usuarios pertenecientes a empresas o compañías, en este caso serían de interés los periódicos, cuentas oficiales de establecimientos, etc.

A continuación, se definen una serie de terminos que facilitarán la comprensión del estudio:

- Tweet: mensaje publicado en twitter. Es de menos de 280 caracteres, originalmente la dimensión de estos textos era de menos de 140, pero debido a su excesiva brevedad fue modificado.
- Cuenta de Twitter: también conocida como usuario de twitter. Es la herramienta que necesitan los usuarios de twitter para poder utilizar twitter. No es necesario para poder observar tweets, pero sí para poder crearlos. Siempre está precedida de "@", por lo que si se quiere mencionar a una persona se tiene que poner lo siguiente: @ nombredeusuario
- Hashtag: son palabras o grupos de palabras (unidas, es decir, sin espacios) que son precedidas por la almohadilla "#", que se conoce como el símbolo hash en inglés. Además, se sabe que tag significa etiqueta en inglés. El hashtag no es más que una

manera de etiquetar o clasificar los tweets, por lo que se pueden agrupar en torno a un mismo tema.

- Mención: ocurre cuando se crea un tweet hablando de alguien o con alguien. Podría decirse que es un tipo de tweet.
- Retweet: ocurre en el momento que se comparte el tweet de una persona.

1.2. Lenguaje de programación R

El lenguaje de programación utilizado para la realización del análisis de este trabajo ha sido R. R es un lenguaje de programación orientado al análisis estadístico y representación gráfica de los datos obtenidos. Además, es un lenguaje con licencia GNU, es decir, es libre, gratuito y abierto. El Lenguaje R se creó en 1993, en la universidad de Auckland. Viene derivado de otros dos lenguajes, que son S y Scheme. Sus creadores son Ross Ihaka y Robert Gentleman.

Características generales del lenguaje R:

- Posibilidad de crear gráficos, basado en LaTeX. Además, existe la posibilidad de imprimir directamente sobre la pantalla.
- Gran cantidad de herramientas estadísticas: modelos lineales y no lineales, tests estadísticos y algoritmos de clasificación y agrupamiento, entre otras.
- Posibilidad de crear tus propias funciones.
- Lenguaje de programación sencillo.
- Almacenamiento y manejo eficaz de los datos.

Algunas de las técnicas estadísticas que proporciona R están implementadas en el propio R al instalarlo, y otras se pueden obtener descargando una serie de paquetes almacenados en la nube. Para la realización de este trabajo se ha utilizado RStudio, que es un editor de R de gran utilidad ya que, si hay algún error en el código, informa de cuál es el error y en qué parte está. Además, gracias al diseño es de fácil comprensión. En este caso, se ha utilizado Rmarkdown, que es una parte de RStudio que puede generar documentos pdf.

Los paquetes utilizados son los siguientes:

- TwitteR: es el paquete utilizado para la descarga de tweets.
- tidyverse: es un paquete de análisis de datos que es utilizado para el manejo de datos y su posterior visualización.
- RColorBrewer: es un paquete que contiene paletas de colores útiles para la visualización de gráficos.
- wordcloud: paquete utilizado para la creación de las nubes de palabras.
- tm: paquete específico para la minería de textos.
- xtable: paquete utilizado para la realización de tablas en R Markdown.
- tidytext: paquete utilizado para el manejo de textos.
- scales: paquete utilizado para cambiar la escala de los datos, en el caso de este trabajo se utiliza para cambiar a porcentaje.

1.3. Obtención de datos

Existen diversas formas de obtener datos a raíz de twitter, pero todas comienzan de la misma forma. Primero hay que tener una cuenta de twitter para poder crear una API de twitter para poder descargar tweets. Una API de twitter es una aplicación predeterminada que permite realizar la descarga de tweets públicos sin necesidad de consentimiento o de tweets privados con el consentimiento del usuario, en este caso se han descargado tweets públicos. Una vez creada la API, se obtienen una serie de contraseñas y códigos que serviran para poner en contacto a RStudio con la API y poder realizarse la descarga. Después de esto, se inserta en RStudio el siguiente código.

```
# consumerKey = "at7F*************************"
# consumerSecret = "tBgr*************************"
# accessToken = "1095*************************"
# accessSecret = "k2KM*************************"
# options(httr_oauth_cache=TRUE)
# setup_twitter_oauth(consumer_key = consumerKey,
#                     consumer_secret = consumerSecret,
#                     access_token = accessToken,
#                     access_secret = accessSecret)
```

A continuación, se recogen distintas formas de descargar tweets. Por un lado, se pueden descargar todos los tweets que contengan una palabra en concreto (con la función "searchTwitter") y por otro, los tweets que ha realizado una persona en concreto (con la función "userTimeline"), ambas funciones de un paquete llamado *twitteR*. En el caso de este estudio, se ha utilizado la primera opción, ya que de esta forma se puede observar de qué se habla sobre un usuario en concreto. El código utilizado para la primera descarga de tweets de turismo es el siguiente, se ha realizado el mismo código para todas y cada una de las cuentas del estudio.

```
# sevilla_tweets1=searchTwitter("@Sevilla_turismo",n=6000,
#                                 since="aaaa-mm-dd")
# tweets.df.1 <- twListToDF(sevilla_tweets1)
# write.csv(tweets.df.1, "datos_tweets_sevilla_turismo.csv")
```

Se puede observar que la función "searchTwitter" tiene los siguientes valores: palabra que se quiere buscar, número máximo de tweets que se quieren descargar y desde qué día se quiere que se produzca la descarga.

Esta forma de descargar los tweets tiene un problema, y es que solo puede descargar tweets de 10 días antes al día de la descarga. Este problema supone que finalmente se tienen varias bases de datos con los tweets, ya que en el caso del análisis turísticos se han analizado 21 días. La realización de este proceso ha consistido en descargar tweets cada 10 días y después unir dichas bases de datos en una. Por lo que para la segunda descarga se hace lo siguiente.

```
# sevilla_tweets1=searchTwitter("@Sevilla_turismo",n=6000,
#                                 since="aaaa-mm-dd")
# tweets.df.1 <- twListToDF(sevilla_tweets1)
# write.csv(tweets.df.1, "datos_tweets_sevilla_turismo2.csv")
#
```

```
# datos1 <- read.csv("datos_tweets_sevilla_turismo.csv")
# datos2 <- read.csv("datos_tweets_sevilla_turismo2.csv")
#
# datos_sevilla_turismo <- bind_rows(datos2, datos1)
#
# write.csv(datos_sevilla_turismo, "datos_tweets_sevilla_turismo.csv")
```

También existe la posibilidad de añadir el comando "exclude:retweets" a continuación de la palabra que se quiere buscar, como se ha hecho en el caso del análisis político. Esta función hace que no se incluyan los retweets en el momento de la descarga de los datos.

```
# iglesias_tweets <- searchTwitter("@Pablo_Iglesias_ exclude:retweets",
#                                  n=50000, since="aaa-mm-dd")
# tweets.df.iglesias <- twListToDF(iglesias_tweets)
# write.csv(tweets.df.iglesias, "datos_tweets_iglesias.csv")
```

Por cada tweet que se ha descargado se tiene la siguiente información:

- Contenido del tweet
- Fecha y hora de su creación
- Dispositivo desde el que se ha realizado el tweet
- Usuario
- Otros datos no relevantes para el estudio

Si se refiere a la función "UserTimeline", que como ya se ha dicho anteriormente se encarga de descargar los tweets realizados por el usuario que se indique, el código es el siguiente.

```
# sevilla_tweets1<- userTimeline("Sevilla_turismo", n = 6000)
# sevilla_tweets_text1 <- sapply(sevilla_tweets1, function(x) x$getText())
```

De esta forma los tweets descargados tienen un formato similar al explicado anteriormente.

1.4. ¿Qué es un dendrograma?

Dendrograma proviene del griego *dendrón*, que significa árbol y es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado. Un dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente). Este tipo de representación permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre grupos de ellos. Observando las sucesivas subdivisiones podemos hacernos una idea sobre los criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc. También podríamos referirnos al dendrograma como la ilustración de las agrupaciones derivadas de la aplicación de un algoritmo de clustering jerárquico, ya que, pueden emplearse para evaluar la cohesión de los clústeres

que se han formado y proporcionar información sobre el número adecuado de clústeres que deben conservarse.

Ejemplo de dendrograma:

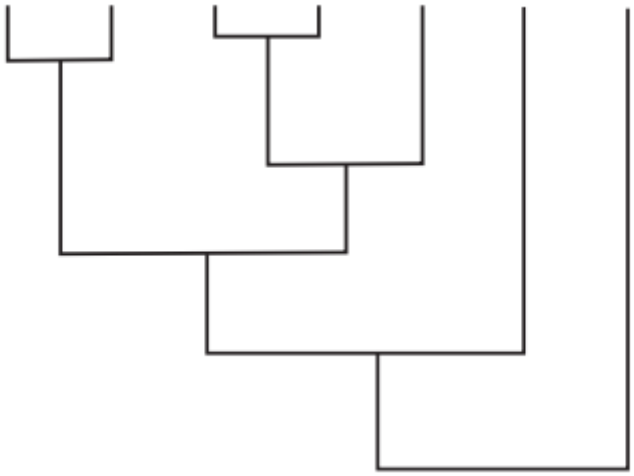


Figura 1.1: Ejemplo dendrograma

El clustering jerárquico tiene por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud. Dentro de los clustering jerárquicos, se ha seleccionado el método de Ward, que consiste en unir los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster. Una investigación llevada a cabo por Kuiper y Fisher probó que este método era capaz de acertar mejor con la clasificación óptima que otros métodos (mínimo, máximo, media y centroide). Por este motivo ha sido el seleccionado para realizar este trabajo.

1.5. Análisis de sentimientos

El análisis de sentimientos, también conocido como minería de opinión (opinion mining), se trata de una tarea de clasificación masiva de documentos de manera automática, que se centra, entre otras cosas, en catalogar los documentos en función de la connotación positiva o negativa del lenguaje utilizado en el mismo. Es importante mencionar que estos tratamientos generalmente se basan en relaciones estadísticas y de asociación, no se basan en análisis lingüístico.

En términos generales, el análisis de sentimientos intenta determinar la actitud de un interlocutor o usuario con respecto a algún tema o la polaridad contextual general de un documento. La actitud puede ser su juicio o evaluación, estado afectivo (el estado emocional del autor al momento de escribir), o la intención comunicativa emocional (el efecto emocional que el autor intenta causar en el lector). El análisis de sentimientos es extremadamente útil en la monitorización de las redes sociales, ya que permite hacer una idea de la opinión pública general sobre ciertos temas.

Los beneficios del análisis de sentimientos son numerosos e importantes. La habilidad de extraer información de datos de las redes sociales es una práctica que ya están adoptando organizaciones a nivel mundial.

Se ha probado que los cambios en el sentimiento de las redes sociales se corresponden a cambios en la bolsa. Por ejemplo, si por redes sociales se publica algo de carácter negativo respecto a una empresa y tiene mucha repercusión, eso puede acarrear que la empresa baje considerablemente sus ventas. Otro ejemplo en el que es de utilizado el análisis de sentimientos es en el sector de la política, ya que, la administración de Obama utilizó el análisis de sentimientos para sondear la opinión pública sobre sus políticas y mensajes de campaña antes de las elecciones presidenciales del 2012.

El análisis de sentimientos tiene dos enfoque principales:

- Enfoques semánticos.
- Enfoques basados en aprendizaje automatizado.

1.5.1. Enfoques semánticos

El enfoque semántico se caracterizan por el uso de diccionarios de términos con orientación semántica de polaridad u opinión. Existen gran cantidad de diccionarios como por ejemplo:

- afinn: (intensidad) se encarga de catalogar de -5 a 5 una serie de palabras, siendo -5 muy negativa y 5 muy positiva. Es el que se ha utilizado en este trabajo.
- elhpolar: (polaridad) se encarga de catalogar las palabras en que tengan connotación positiva o negativa.
- senticon: (emoción) clasifica las palabras en emociones como sorpresa, tristeza, enfado, etc.

La ventaja principal de los enfoques semánticos es que los errores son relativamente sencillos de corregir, añadiendo cuantos términos fuera necesario, y se podría obtener una precisión tan alta como se quisiera, simplemente invirtiendo más tiempo en la construcción del diccionario. Sin embargo, el esfuerzo para construir un diccionario para un cierto dominio, empezando de cero, es muy elevado, porque se basa en mucho trabajo manual, así que en general son menos adaptables.

1.5.2. Enfoques basados en aprendizaje automatizado

Los enfoques basados en aprendizaje computacional consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos anotados, donde cada texto habitualmente se representa con un vector de palabras (bag of words), n-gramas o skip-grams, en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la negación, la subjetividad o la ironía. Los sistemas utilizan diversas técnicas, aunque las más populares son los clasificadores basados en SVM (Support Vector Machines), Naive Bayes y KNN (K-Nearest Neighbor). En las investigaciones más recientes se han empezado a utilizar otras técnicas más avanzadas, como LSA (Latent Semantic Analisis) e incluso Deep Learning.

La ventaja principal de los enfoques basados en aprendizaje automático es que cuesta muy poco construir un analizador de sentimientos a partir de la colección de textos

etiquetados, ya que la tarea de modelado reside en el algoritmo. Por ello es relativamente fácil construir clasificadores adaptados a un dominio determinado. Sin embargo, suelen ser una caja negra en la que corregir errores o añadir nuevos conocimientos es más complicado, y muchas veces sólo es posible ampliando la colección de textos y volviendo a entrenar el modelo.

Capítulo 2

Turismo

Uno de los principales motivos por el que se ha realizado un análisis turístico es porque el turismo es una de las principales fuentes económicas de nuestro país. Además, si nos centramos en Sevilla, se observa como todo va entorno al turismo, por lo que me pareció de gran interes analizar qué opinan las personas en relación al turismo de Sevilla y después poder compararlo con la opinión que se tiene de las ciudades más turísticas en España (ciudades comparables con Sevilla, es decir, ciudades en las que el principal motivo de turismo sea turismo de ciudad, no de playa o rural).

El principal objetivo de este capítulo es analizar diferentes cuentas de algunas ciudades turísticas, para conocer qué se está escribiendo en Twitter sobre turismo en ellas, qué cuentas son más activas, en qué día de la semana se han realizado más tweets, desde qué dispositivo se hizo el tweet, etc. Además, se ha realizado un análisis de sentimientos para observar la positividad (o negatividad) que expresan los usuarios al twittear sobre turismo.

El estudio se ha centrado en la comparación de las cuentas de la ciudad de Sevilla con las de algunas de las ciudades turísticas importantes de España: Córdoba, Málaga, Granada, Madrid, Barcelona y Valencia. Las cuentas seleccionadas son las siguientes:

1. SEVILLA:

- @ Sevilla_Turismo: Consta de 13,8k de seguidores y 2.299 me gustas.
- @ sevillaciudad: Con 116k de seguidores y 4.375 me gustas.

2. CÓRDOBA:

- @ CordobaESP: Consta de 17,4k de seguidores, 5.675 me gustas.
- @ PatiosdeCordoba: Tiene 3.297 seguidores, 1.885 me gustas.

3. MÁLAGA:

- @ turismodemalaga: Cuenta oficial con 27,3k de seguidores y 2.932 me gustas.
- @ vivecostadelsol: Cuenta con 21,7k de seguidores y 5.517 me gustas.

4. GRANADA:

- @ turgranada: Consta de 22,3K seguidores y 26,7k me gustas.
- @ alhambracultura: Cuenta oficial del Patronato de la Alhmabra y Generalife. 60,4k de seguidores y 27,4k de me gustas.

5. MADRID:

- @ TurismoMadrid: cuenta oficial con 305k seguidores y 11,7k me gustas.
- @ Visita_Madrid: cuenta oficial con 60,3k de seguidores y 5.750 me gustas.

6. BARCELONA:

- @ turismoBCN: tiene 3.994 seguidores y 37 me gustas.
- @ sagradafamilia: Cuenta oficial con 16,9k de seguidores y 1.596 me gustas.

7. VALENCIA:

- @ c_valenciana: Cuenta oficial del Portal Oficial de Turismo de la Comunidad Valenciana. Consta de 105k de seguidores y 5.749 me gustas.
- @ Valenciaturismo: Cuenta con 33,1k de seguidores y 6.023 me gustas.

Se han seleccionado los tweets desde el 25 de febrero de 2019 hasta el 17 de marzo de 2019 de las cuentas citadas anteriormente, se dispone por tanto de un total de 3 semanas de información.

2.1. Análisis de tweets

En este apartado se ha realizado un análisis del número de tweets que se han recopilado, y de las menciones que se han realizado de las diferentes cuentas. Se ha analizado el número de tweets por ciudades, por día de la semana en el que fue escrito y el dispositivo de uso. Además, también, se ha analizado qué palabras/temas se han utilizado en las diferentes cuentas que se han analizado, si están relacionadas entre sí y el posible motivo de dicha relación.

2.1.1. Número de tweets

En la siguiente gráfica, se puede observar el número de tweets del total de las cuentas seleccionadas por cada ciudad.

Se puede observar, con las cuentas que se han estudiado, que la ciudad con mayor número de menciones es Madrid con cerca de 12.000 tweets, seguido de Valencia y Córdoba con cerca de 8.500 y 8.000, respectivamente. Sevilla se encuentra en cuarto lugar en cuanto el número de menciones, con cerca de 3.500 tweets. Las ciudades con menor número de menciones son Barcelona, Granada y Málaga, con poco más de 2.000 tweets.

A continuación, se puede ver el número de tweets por menciones de la cuenta y ciudad.

La anterior tabla es de utilidad para ver el número de tweets que se han escrito incluyendo el nombre de la cuenta. Se observa que la cuenta con mayor número de menciones es *@ visita_madrid* con 9.169, seguida de *@ cordobaESP* con 8.084 y *@ c_valenciana* con 5.995. Ademas, se destaca que las dos cuentas de Sevilla se han mencionado de una forma similar, entre las 1.400 y 2.000 menciones. Finalmente señalar que, las cuentas que han tenido menos menciones son *@ patiosdecordoba* con 87 tweets y *@ turismoBCN* con 82 tweets. Se observa que hay gran diferencia entre el número de tweets que se ha mencionado a una cuenta u otra, en algunos casos diferencia de cerca de 9.100 tweets.

En la siguiente tabla, se ha analizado qué días de la semana tienen una mayor o menor actividad en Twitter.

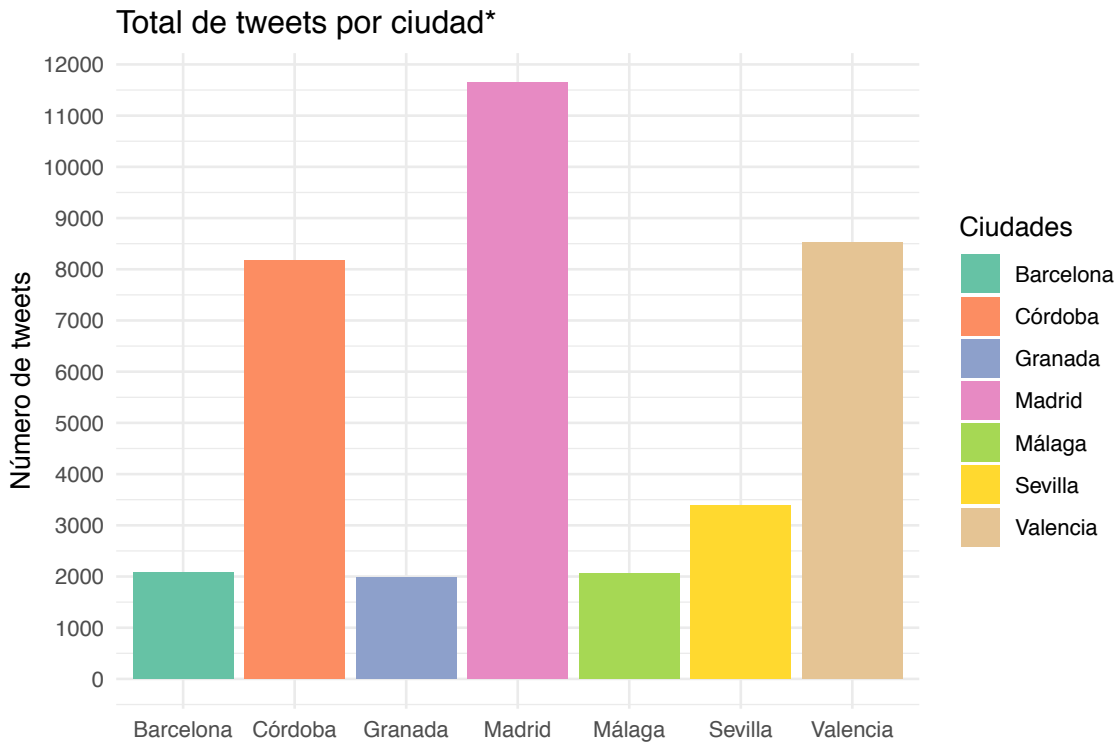


Figura 2.1: Total de menciones por ciudad. Nota: *Número de tweets de las menciones a las cuentas de cada ciudad

Se puede observar que los días con mayor número de tweets son el miércoles con 6.568 tweets, seguido del jueves con 6.112 y el martes con 6.010. Los días con menor número de tweets son el lunes con 5.158, seguido del sábado con 4.922 y el domingo con 3.633.

A continuación, se puede observar las menciones de las diferentes cuentas por día de la semana

Se observa que dependiendo de la cuenta, la actividad varía notablemente respecto a las menciones realizadas por días de la semana. De forma general, en la mayoría de las cuentas el máximo número de menciones se produce en días entre semana, excepto para el caso de *@ c_valenciana* en la que se produce en fin de semana. Si se analiza las cuentas en las que se han realizado un mayor número de menciones, para *@ visita_madrid* el día más activo son los miércoles, para *@ cordobaESP* son los viernes y *@ c_valenciana* con los sábados. Si atendemos a las cuentas que se han analizado de Sevilla, se observa que *@ sevilla_turismo* tiene un mayor número de menciones los jueves y en *@ sevillaciudad* son los martes.

A continuación, en la siguiente gráfica, se observa el número de tweets de las menciones a las cuentas de cada ciudad por día del mes, desde el 25 de febrero hasta el 17 de marzo.

Cuenta	Ciudad	Número de tweets
@sagradafamilia	Barcelona	2.001
@turismoBCN	Barcelona	82
@cordobaESP	Córdoba	8.084
@patiosdecordoba	Córdoba	87
@alhambracultura	Granada	1.361
@turgranada	Granada	618
@turismoMadrid	Madrid	2.470
@visita_madrid	Madrid	9.169
@turismodemalaga	Málaga	1.442
@vivecostadelsol	Málaga	610
@sevilla_turismo	Sevilla	1.977
@sevillaciudad	Sevilla	1.404
@c_valenciana	Valencia	5.995
@Valenciaturismo	Valencia	2.537

Cuadro 2.1: Número de tweets de las menciones por cuenta y ciudad

Día de la semana	Número de tweets
Lunes	5.158
Martes	6.010
Miércoles	6.568
Jueves	6.112
Viernes	5.434
Sábado	4.922
Domingo	3.633

Cuadro 2.2: Número de tweets de las menciones por día de la semana

En la gráfica anterior, se observa que Madrid es la ciudad con mayor número de menciones durante gran parte de los días a excepción de los días 2, 3 y 4 de marzo, que es superada por Valencia, esto puede ser debido a que son los primeros días de Las Fallas, y de los días 7, 8 y 15 de marzo fue superada por Córdoba. Además, se destaca que el 27 de febrero hay una alta actividad en la ciudad de Madrid, esto puede ser debido a que en esa fecha jugaron el Real Madrid y el F. C. Barcelona la vuelta de la semifinal de la copa del rey en Madrid. La ciudad de Sevilla alcanza su mayor número de menciones el día 5 de marzo y la mínima el 10 de marzo. Ademas, cabe destacar que ciudades como Barcelona, Granada y Málaga se encuentran durante todo el período de estudio con menos de 200 menciones por día.

En la siguiente tabla se puede observar el número de menciones de cada una de las cuentas por dispositivo en el que fue twitteada, es decir, si el tweet se ha realizado desde la Web, un dispositivo Android, iPhone, iPad u otros. Comentar que, la categoría otros engloba dispositivos como, por ejemplo, Blackberry.

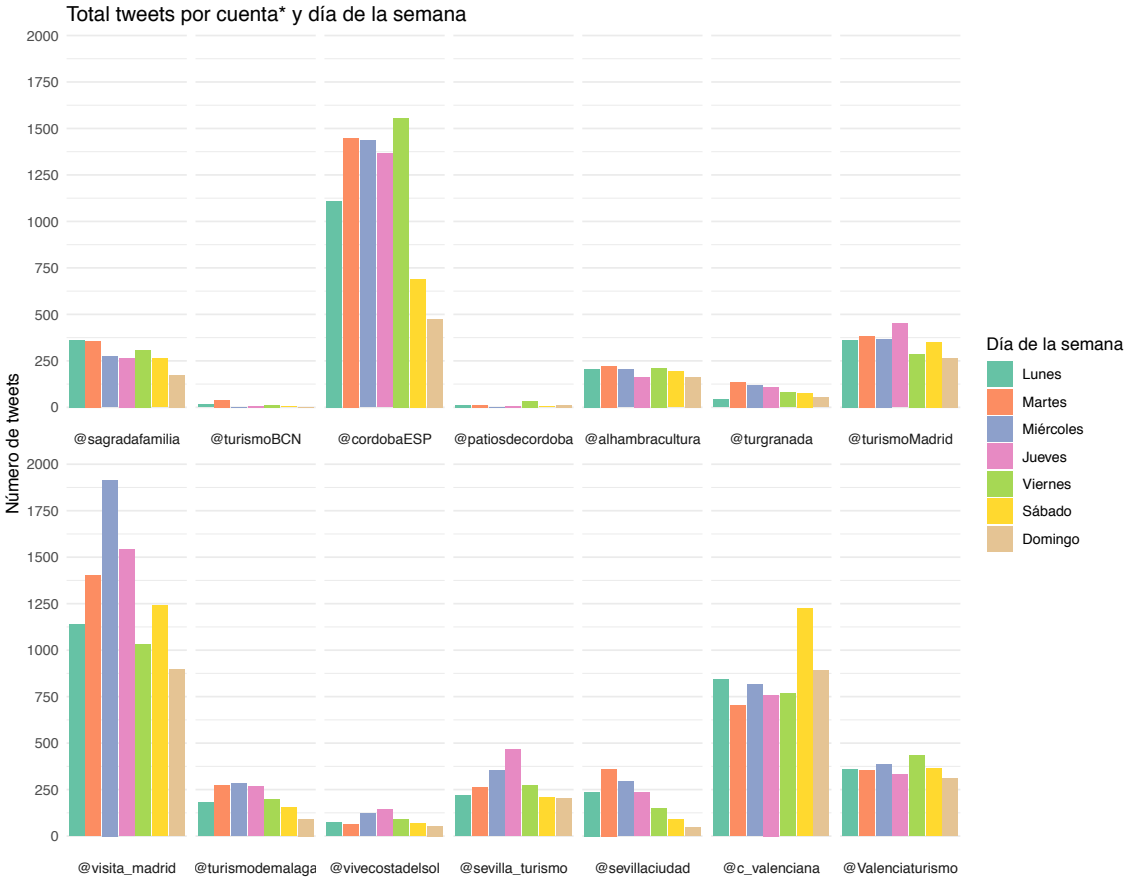


Figura 2.2: Total de menciones por cuenta y día de la semana. Nota: *Número de tweets de las menciones a las cuentas.

Se observa que a nivel general el dispositivo más utilizado ha sido Android, seguido de iPhone y Web, y el que menos ha sido iPad. Además, en las menciones de las cuentas como *@ sevillaciudad*, *@ vivecostadelsol* y *@ turgranada* se han utilizado iPhone y Android con una frecuencia similar.

2.1.2. Análisis por cuentas de Twitter

A continuación, se han analizado las diferentes cuentas de las que se han extraído tweets, sobre qué temas han hablado y qué palabras han utilizado durante el periodo donde se han recogido los tweets. Para ello, se dispondrá de una serie de tablas y gráficos de cada una de las ciudades.

2.1.2.1. Sevilla:

Como ya se ha comentado, en Sevilla, las cuentas que se han analizado son *@ sevilla_turismo* y *@ sevillaciudad*. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

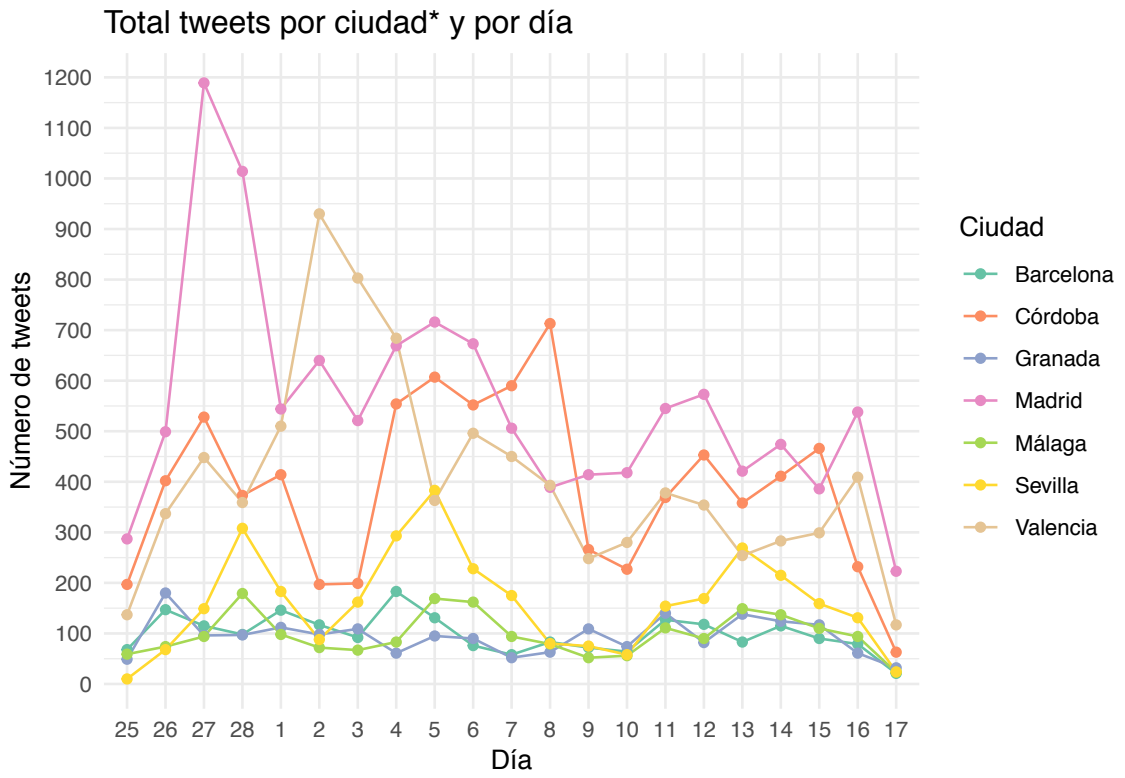


Figura 2.3: Total de tweets por ciudad y día del mes. Nota: *Número de tweets de las menciones a las cuentas de cada ciudad

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

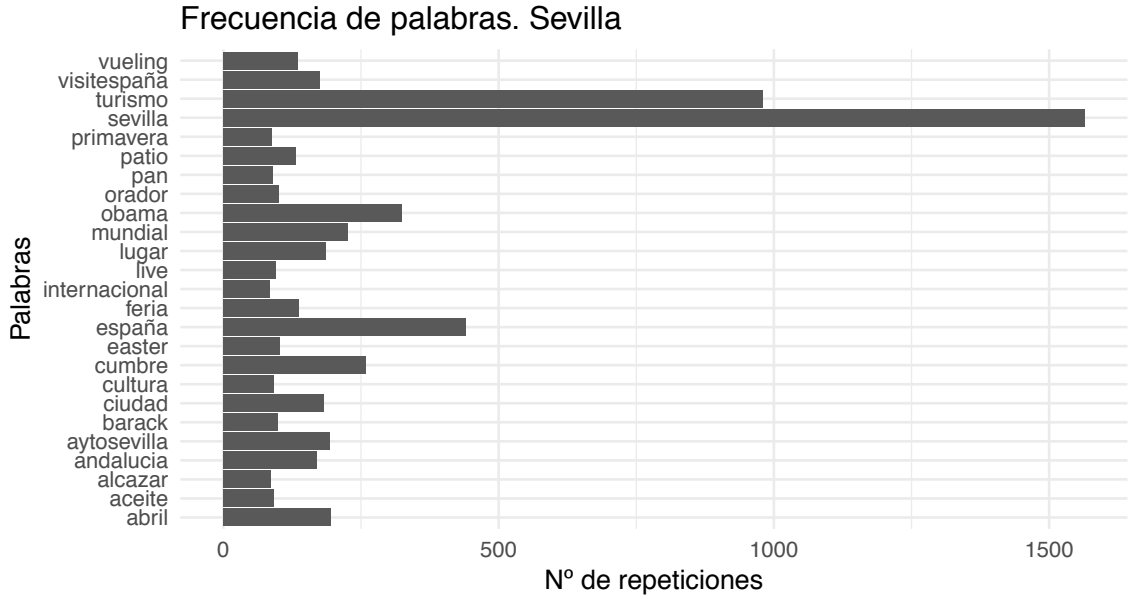


Figura 2.4: Frecuencia de palabras más utilizadas. Sevilla

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Ciudad	Cuenta	Android	iPhone	iPad	Web	Otros
Barcelona	@sagradafamilia	612	440	44	459	446
Barcelona	@turismoBCN	31	10	1	27	13
Córdoba	@cordobaESP	3.061	1.501	147	1.447	1.928
Córdoba	@patiosdecordoba	31	24	2	17	13
Granada	@alhambracultura	681	298	33	295	54
Granada	@turgranada	264	204	7	102	41
Madrid	@turismoMadrid	1.163	548	109	449	201
Madrid	@visita_madrid	3.886	1.905	165	1.794	1.418
Málaga	@turismodemalaga	548	318	41	313	222
Málaga	@vivecostadelsol	197	169	24	140	80
Sevilla	@sevilla_turismo	734	422	36	380	405
Sevilla	@sevillaciudad	514	456	34	243	157
Valencia	@c_valenciana	2.580	1.285	113	1.294	723
Valencia	@Valenciaturismo	1.169	664	43	412	249

Cuadro 2.3: Total tweets de las menciones por cuenta y dispositivo de uso

Palabra	Repeticiones	%	Palabra	Repeticiones	%
sevilla	1.564	25,11	feria	138	2,22
turismo	980	15,73	vueling	135	2,17
españa	440	7,06	patio	132	2,12
obama	324	5,20	easter	102	1,64
cumbre	258	4,14	orador	100	1,61
mundial	226	3,63	barack	99	1,59
abril	196	3,15	live	96	1,54
aytosevilla	193	3,10	cultura	92	1,48
lugar	186	2,99	aceite	92	1,48
ciudad	182	2,92	pan	90	1,44
visitespaña	175	2,81	primavera	89	1,43
andalucia	169	2,71	alcazar	86	1,38

Cuadro 2.4: Número de repeticiones y porcentaje de uso de cada palabra. Sevilla

Figura 2.5: Nube de palabras. Sevilla

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 84 veces. En la Tabla 2.4. se observa que las palabras más utilizadas en Sevilla respecto al turismo son *Sevilla*, *Turismo* y *España*, con 1.564, 980 y 440 repeticiones. Seguida de *Obama*, *cumbre* y *mundial* con 324, 258 y 226 repeticiones, respectivamente. También se podría destacar que *Barack* es una de las palabras utilizadas con 99 repeticiones y la palabra *feria* con 138 repeticiones. No obstante, dentro de las palabras más utilizadas, las menos utilizadas son *pan*, *primavera* y *alcázar*, con 90, 89 y 86 repeticiones, repsectivamente. El motivo por el que las palabras que han sido utilizadas de manera más frecuente, Sevilla, Turismo, España, Obama y Barack es debido a que en abril Barack Obama visitirá Sevilla, ya que será el invitado de honor de la décimo octava cumbre global del Consejo Mundial de Viajes y Turismo, el World Travel & Tourism. Por último, entre las palabras que se han mencionado de forma más frecuente están *pan*, *aceite* y *Andalucía*, esto puede ser debido a que el pan con aceite es un típico desayuno andaluz.

A continuación, se tiene un gráfico donde las palabras más utilizadas están presentes y forman grupos si están relacionadas entre ellas. Como ya se ha comentado, Obama visitará la ciudad de Sevilla, y las palabras *ciudad*, *obama*, *abril*, *mundial* y *cumbre* están formando un grupo. Otro grupo formado por las palabras *españa*, *andalucia*, *feria*, *aytosevilla*, *lugar* y *visitSpain*, puede ser que en la red se está hablando de la feria de abril como evento al que asistir en Sevilla. Hay otro grupo que está formado por las palabras *Sevilla* y *turismo*.

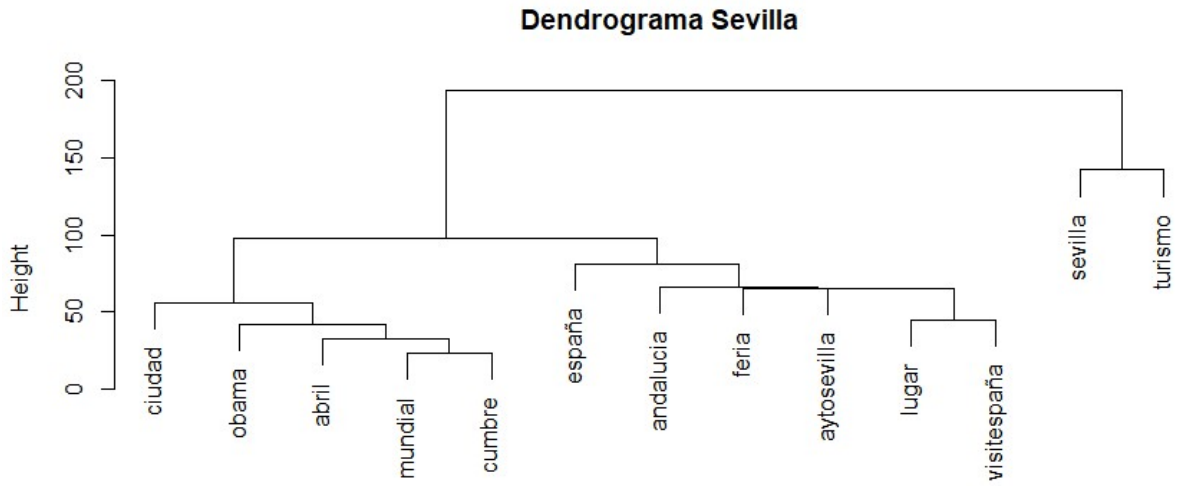


Figura 2.6: Dendrograma. Sevilla

2.1.2.2. Córdoba

Recordar que las cuentas que se han analizado son *@ cordobaESP* y *@ patiosdecordoba* de forma conjunta. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
cordoba	2.085	32,49	looks	175	2,73
moda	522	8,13	semana	166	2,59
ciudad	397	6,19	puente	162	2,52
spain	395	6,16	diadeandalucia	157	2,45
visitspain	360	5,61	carnaval	151	2,35
tendencias	292	4,55	calle	151	2,35
andalucia	226	3,52	felizmartes	145	2,26
instagram	216	3,37	vercordoba	143	2,23
ilusion	212	3,30	mañana	142	2,21
lugares	179	2,79	romano	141	2,20

Cuadro 2.5: Número de repeticiones y porcentaje de uso de cada palabra. Córdoba.

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

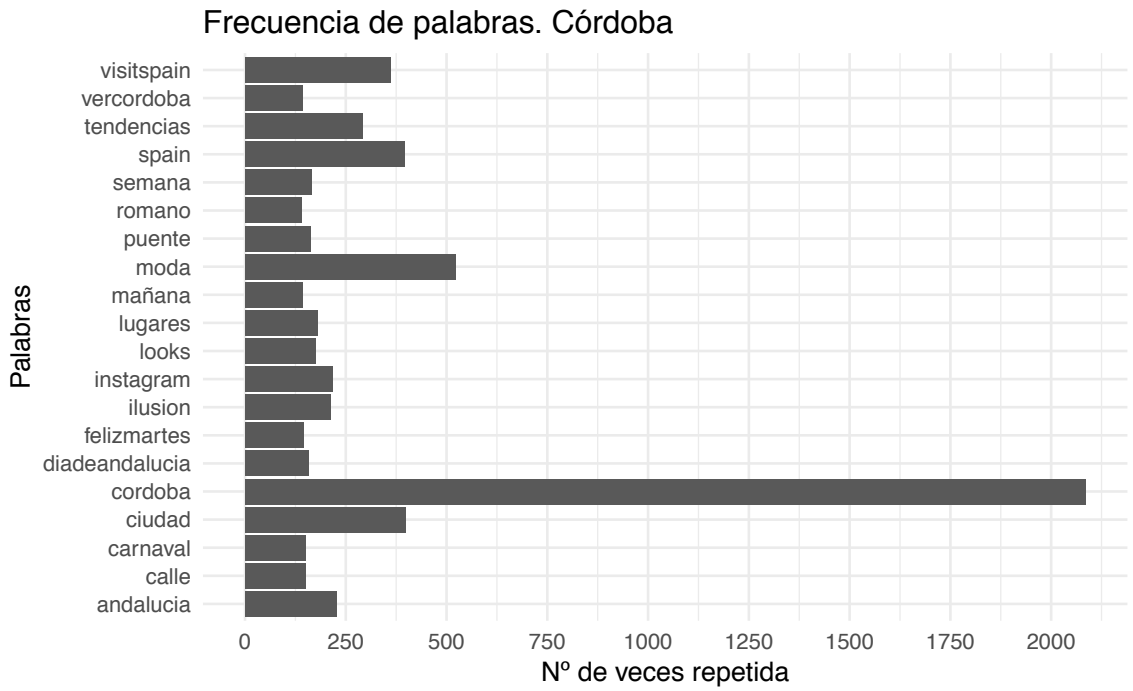


Figura 2.7: Frecuencia de palabras más utilizadas. Córdoba

A continuación, se tiene una nube de palabras.

Figura 2.8: Nube de plabaras. Córdoba

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 140 veces. Se observa que las palabras más utilizadas en Córdoba respecto al turismo son *Córdoba*, *moda* y *ciudad* con 2.085, 522 y 397 repeticiones, respectivamente. Como palabras menos frecuentes se tienen *vercordoba*, *mañana* y *romano* con 143, 142 y 141, respectivamente. También se puede destacar, palabras como *puente* y *romano*, ya que uno de los lugares más visitados en Córdoba es el Puente Romano. Además, se observa *diadeandalucia*, ya que como se ha dicho anteriormente, el día de Andalucía es el 28 de febrero.

A continuación, se tiene un gráfico en el que se establecen las relaciones entre palabras.

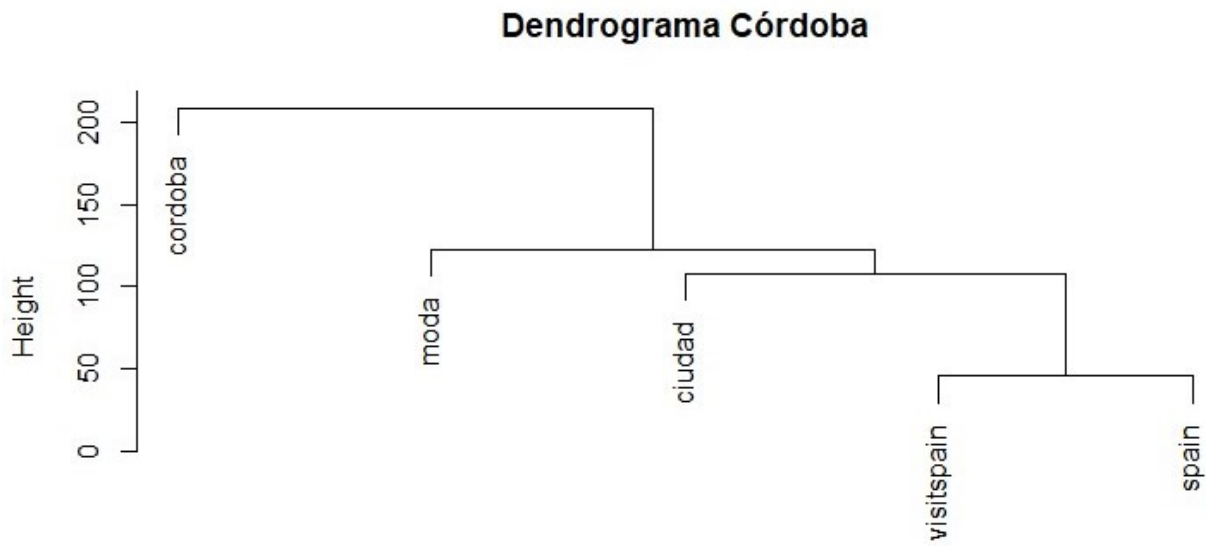


Figura 2.9: Dendrograma. Córdoba

Se puede observar que no hay gran número de palabras relacionadas en las cuentas seleccionadas de la ciudad de Córdoba. Se podría destacar que la palabra *moda* está relacionada con las palabras *ciudad*, *visitSpain* y *Spain*, ya que en Córdoba hay varias cuentas de moda muy activas en twitter.

2.1.2.3. Málaga

Como ya se ha comentado, en Málaga, las cuentas que se han analizado son *@ turismodemalaga* y *@ vivecostadelsol* de forma conjunta. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
malaga	1.133	28,24	perspective	99	2,47
spain	285	7,10	bird	98	2,44
visitspain	284	7,08	different	98	2,44
atardecer	214	5,33	tuhistoria	85	2,12
spainculturalheritage	200	4,99	aceitesdelmedit	79	1,97
ronda	169	4,21	elpimpimalaga	79	1,97
malagaciudadgenial	166	4,14	malagasanta	78	1,94
malagaturismo	144	3,59	spainexperience	68	1,69
visitmalaga	117	2,92	visitas	64	1,60
antequera	115	2,87	cine	57	1,42
informatico	114	2,84	turismo	57	1,42
fly	99	2,47	andalucia	56	1,40

Cuadro 2.6: Número de repeticiones y porcentaje de uso de cada palabra. Málaga.

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

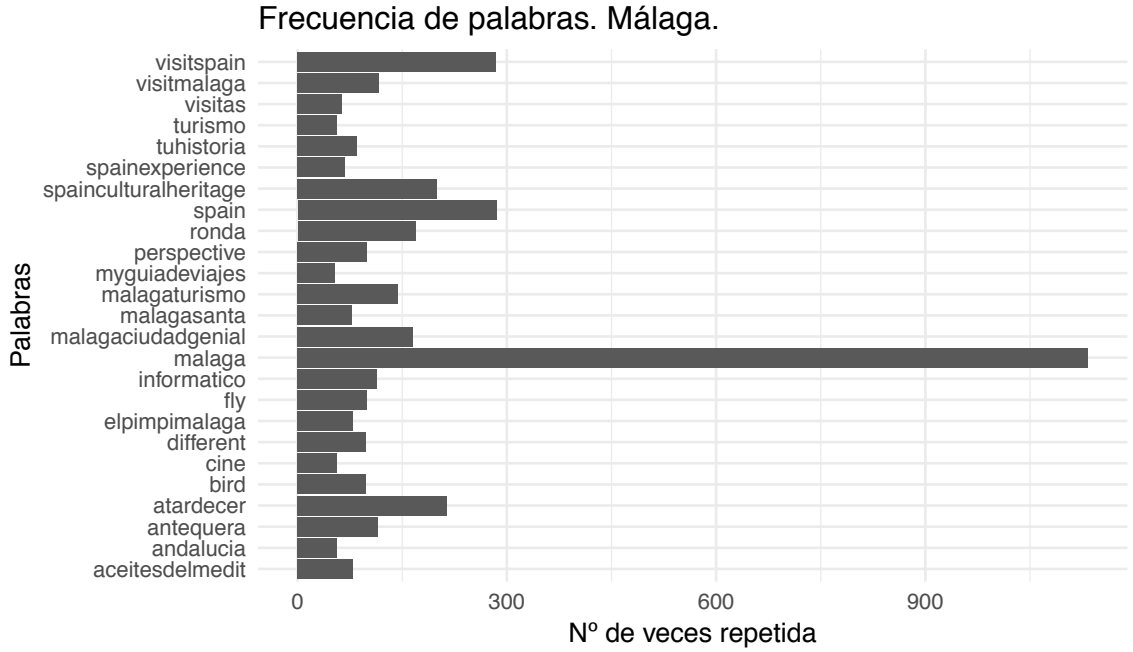


Figura 2.10: Frecuencia de palabras más utilizadas. Málaga

A continuación, se tiene una nube de palabras.

Figura 2.11: Nube de palabras. Málaga

Para hacer el análisis se han seleccionado las palabras que se repiten más de 51 veces. En la Tabla 2.6. se observa que las palabras que más se han utilizado en Málaga respecto al turismo son: *Málaga*, *Spain*, *visitSapain* y *atardecer*, con 1.133, 285, 284 y 214 repeticiones, respectivamente. Las palabras menos utilizadas son *turismo*, *cine* y *Andalucía* con 57, 57 y 56 repeticiones, respectivamente. También se puede destacar la palabra *elpimpimalaga* con 79 repeticiones, es una bodega-bar típica de Málaga que tiene mucha actividad en redes sociales.

Finalmente, al igual que en Córdoba y en Sevilla, observando el gráfico y la nube de puntos se ve que la palabra más utilizada con diferencia de las demás es el nombre de la propia ciudad, en este caso, Málaga.

A continuación, se tiene un gráfico en el que se establecen las relaciones entre las palabras formando diferentes grupos.

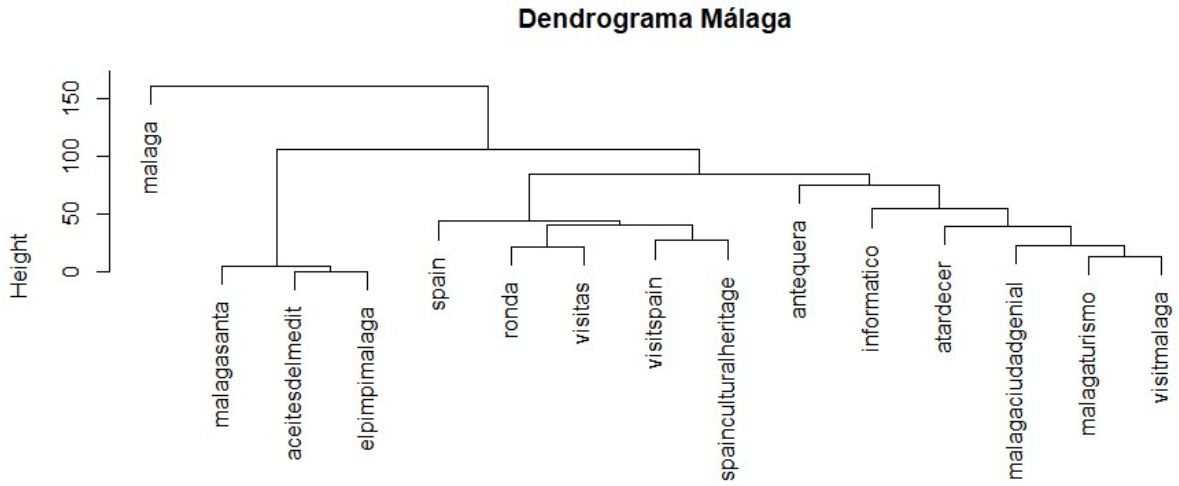


Figura 2.12: Dendrograma. Málaga

Se puede observar que por un lado, hay palabras como *Málagaturismo*, *visitMálaga* y *Málagaciudadgenial* que están relacionadas con *atardecer*, por lo que parece que los atardeceres en Málaga llaman la atención de los turistas. Hay otro grupo que está formado por *SpainCulturalHeritage*, *Ronda*, *VisitSpain* y *Spain*, esto es debido a que Ronda pertenece al patrimonio cultural de España, y es una de las ciudades más atractivas para visitar en dicha provincia.

2.1.2.4. Granada

Recordar que las cuentas que se han analizado son *@ turgranada* y *@ alhambracultura* de forma conjunta. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
alhambra	644	19,77	sonidos	74	2,27
granada	526	16,14	modernidad	72	2,21
torre	167	5,13	trono	71	2,18
vela	133	4,08	almuñecar	67	2,06
almeria	107	3,28	eagoficial	66	2,03
interior	103	3,16	visitas	64	1,96
spain	94	2,89	gratuitas	63	1,93
salon	91	2,79	españa	59	1,81
publico	86	2,64	generalife	59	1,81
monumento	84	2,58	exposicion	55	1,69
simbolos	81	2,49	estancia	54	1,66
palacio	79	2,42	lorca	52	1,60
andalucia	78	2,39	cultural	52	1,60
ciclo	76	2,33	dia	51	1,57

Cuadro 2.7: Número de repeticiones y porcentaje de uso de cada palabra. Granada.

En el siguiente gráfico, se pueden observar las palabras más utilizadas y su respectiva frecuencia.

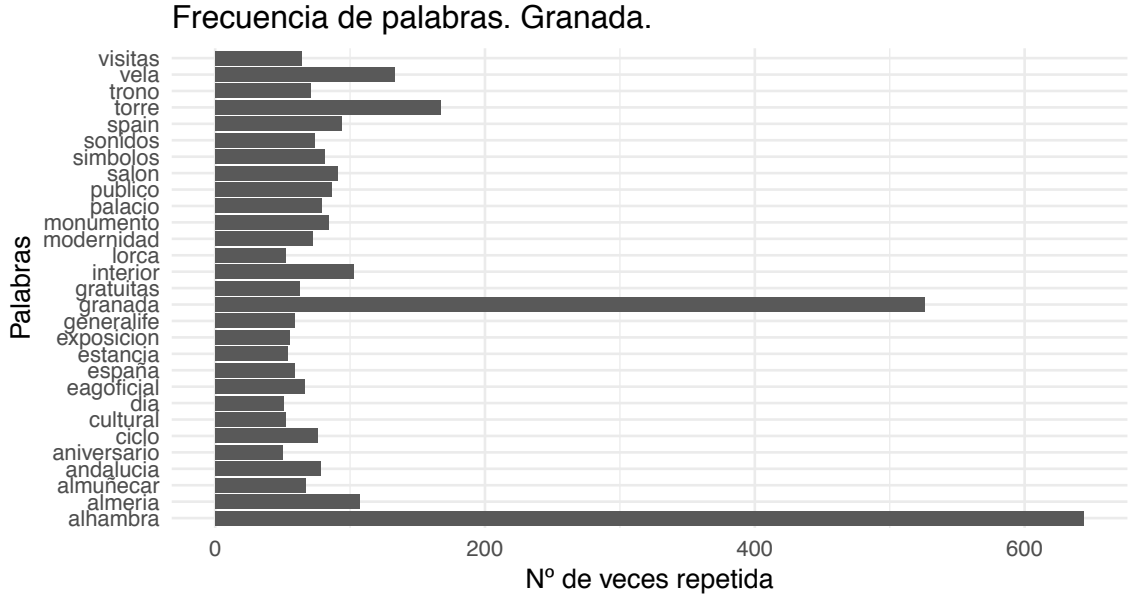


Figura 2.13: Frecuencia de palabras más utilizadas. Granada

A continuación, se tiene una nube de palabras.

Figura 2.14: Nube de palabras. Granada

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 49 veces. En la Tabla 2.7. se observa que las palabras más utilizadas en Granada respecto al turismo son *Alhambra*, *Granada* y *torre*, con 644, 526 y 167 respectivamente. Se puede destacar que en el caso de Granada la palabra más utilizada no es el nombre de la propia ciudad, se encuentra en la segunda posición. También cabe detacar la palabra *Almería* con 107 repeticiones, se posiciona en la quinta palabra más utilizada. Las palabras menos utilizadas son *cultural*, *Lorca* y *día* con 52, 52 y 51 repeticiones, respectivamente. El motivo por el que la palabra *Lorca* se encuentre entre las palabras más utilizadas en Granada puede ser debido a que desde el 11 de octubre hasta el 31 de marzo hay una exposición en el Centro Federico García Lorca llamada Lorca y Granada, que trata la relación que tuvo Lorca con Granada.

A continuación, se tiene un gráfico en el que se establecen las relaciones entre las palabras formando diferentes grupos.

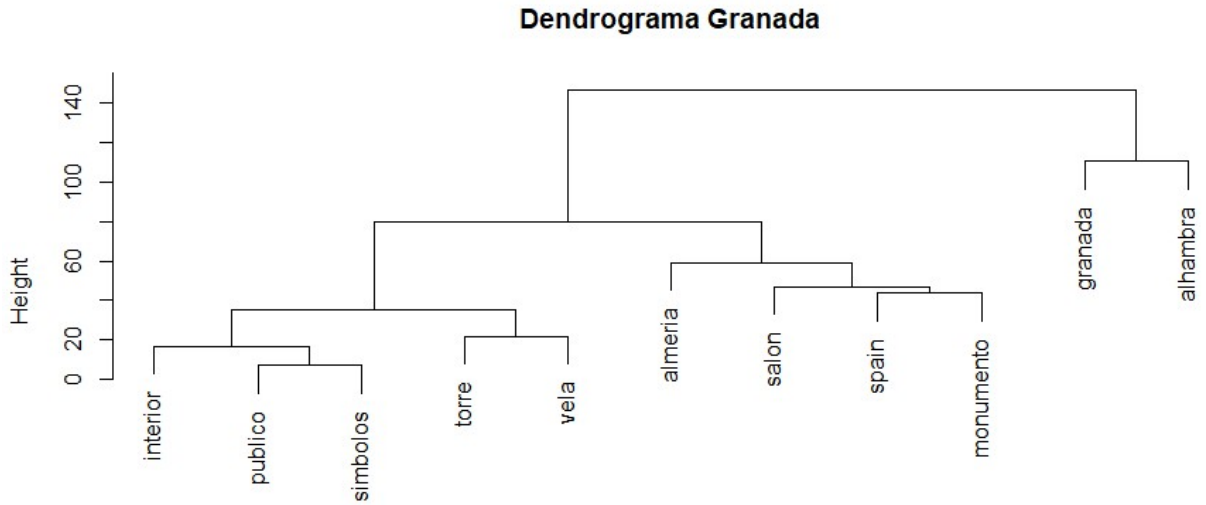


Figura 2.15: Dendrograma. Granada

En el gráfico anterior se puede observar que *Granada* y *Alhambra* forman un grupo entre sí. Además, hay otro grupo formado por las palabras *torre*, *vela*, *interior*, *público* y *símbolos*, debido a que en La Alhambra de Granada hay una torre llamada "Torre de la Vela", y su interior se ha abierto al público durante el mes de marzo.

2.1.2.5. Madrid

Como ya se ha comentado, en Madrid, las cuentas que se han analizado son *@ visita_madrid* y *@ turismoMadrid* de forma conjunta. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
madrid	5.893	27,97	bernabeu	301	1,43
visita	5.051	23,97	primera	292	1,39
real	1.301	6,18	museoalmudena	287	1,36
presidente	656	3,11	marzo	276	1,31
spain	499	2,37	final	271	1,29
madrida	403	1,91	sabado	261	1,24
barcelona	401	1,90	peru	258	1,22
fotos	375	1,78	atletico	251	1,19
lisboa	371	1,76	gran	251	1,19
condebarajas	361	1,71	culturacmadrid	248	1,18
palacio	347	1,65	mañana	246	1,17
hostalpersal	344	1,63	arteenmadrid	233	1,11
granviademadrid	342	1,62	casareal	233	1,11
semana	335	1,59	plaza	225	1,07
visitspain	312	1,48	honor	222	1,05

Cuadro 2.8: Número de repeticiones y porcentaje de uso de cada palabra. Madrid.

En el siguiente gráfico de barras se pueden observar las palabras más utilizadas y sus respectivas frecuencias.

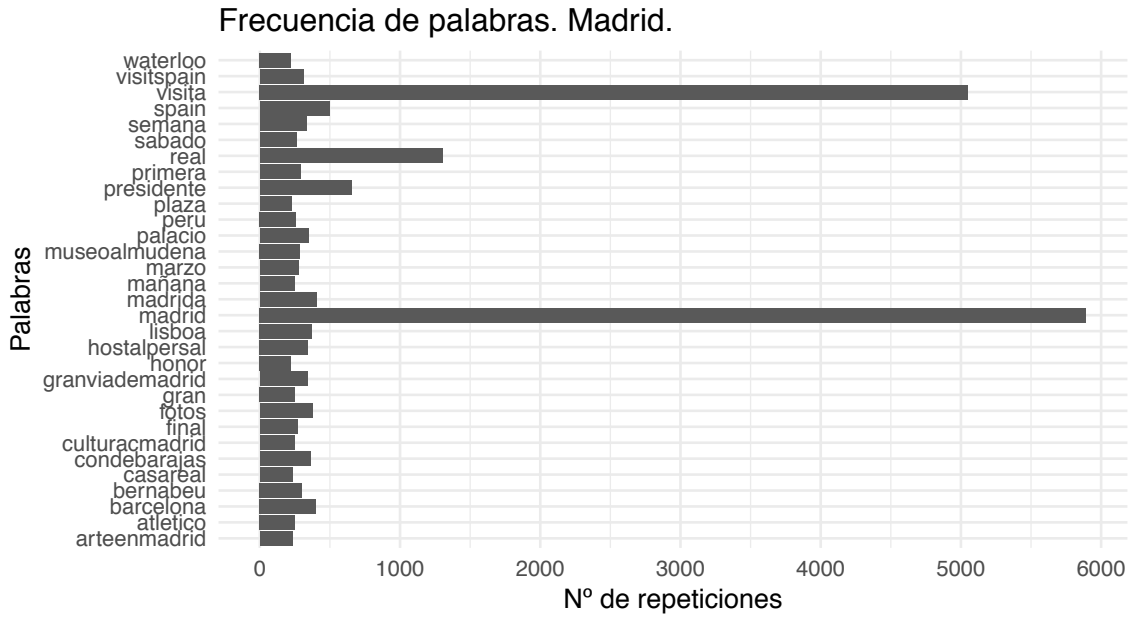


Figura 2.16: Frecuencia de palabras más utilizadas. Madrid

A continuación, se tiene una nube de palabras.

Figura 2.17: Nube de palabras. Madrid

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 220 veces. En la Tabla 2.8. se observa que las palabras más utilizadas en Madrid respecto al turismo son *Madrid*, *visita* y *real* con 5.893, 5.051 y 1.301 repeticiones, respectivamente. Se observa que aunque la palabra más utilizada sea *Madrid*, no hay gran diferencia con la palabra *visita*. Las palabras menos utilizadas son *casareal*, *plaza* y *honor* con 233 , 225 y 222 repeticiones, respectivamente. También, cabe destacar palabras como *Lisboa* y *Barcelona*. Esta última podría ser debido a que como se ha dicho anteriormente, el 27 de febrero fue

la vuelta de la semifinal de la copa del rey entre el Real Madrid y el F. C. Barcelona, además, también jugaron un partido de liga el día 2 de marzo en Madrid. Además, el 14 de febrero se abrió el período de inscripción a el *NonStop Madrid-Lisboa 2019*, que es una carrera en bicicleta que va desde Madrid hasta Lisboa, por lo que parece que este es el principal motivo por el que *Lisboa* aparece como palabra destacada.

A continuación, se tiene un gráfico en el que se establecen las relaciones entre las palabras formando diferentes grupos.

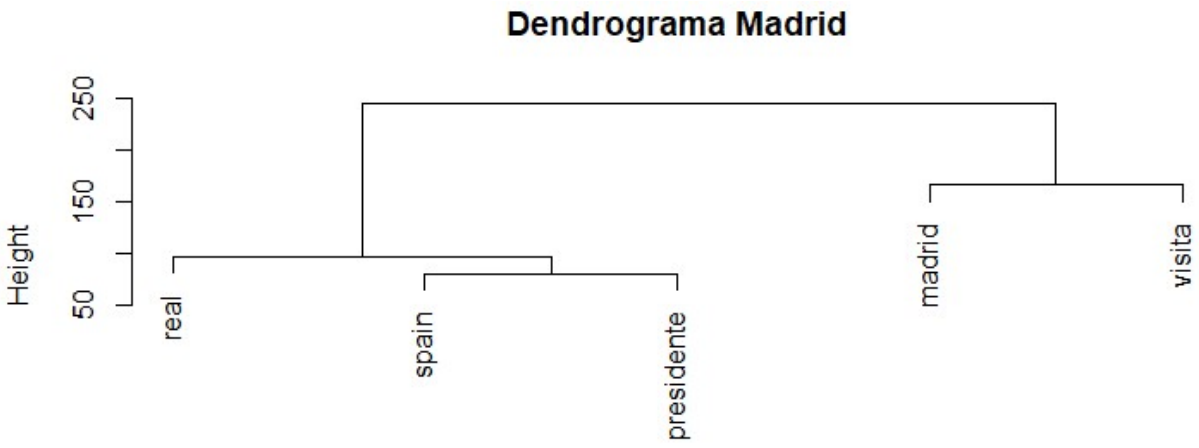


Figura 2.18: Dendrograma. Madrid

En el anterior dendrograma, al igual que en el dendrograma creado para las cuentas de Córdoba, no existe gran relación entre las palabras. Tenemos relacionadas las palabras *visita* y *Madrid*. Se podría concluir que uno de los temas que se habla en Madrid durante el período del estudio es del presidente del Real Madrid, ya que existe cierta relación entre *presidente*, *Spain* y *Real*, esto puede ser debido a que parece que hay una crisis en el equipo.

2.1.2.6. Barcelona

Como ya se ha comentado, en Barcelona, las cuentas que se han analizado son *@ turismoBCN* y *@ sagradafamilia* de forma conjunta. En la siguiente tabla, se observan las palabras que se ha utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
sagradafamilia	1.743	48,92	comparativa	161	4,52
barcelona	472	13,25	basilica	93	2,61
gaudi	216	6,06	travel	84	2,36
sagrada	186	5,22	spain	76	2,13
familia	183	5,14	torres	73	2,05
impresionante	165	4,63	luz	57	1,60

Cuadro 2.9: Número de repeticiones y porcentaje de uso de cada palabra. Barcelona.

En el siguiente gráfico se pueden observar las palabras más utilizadas y su respectiva frecuencia.

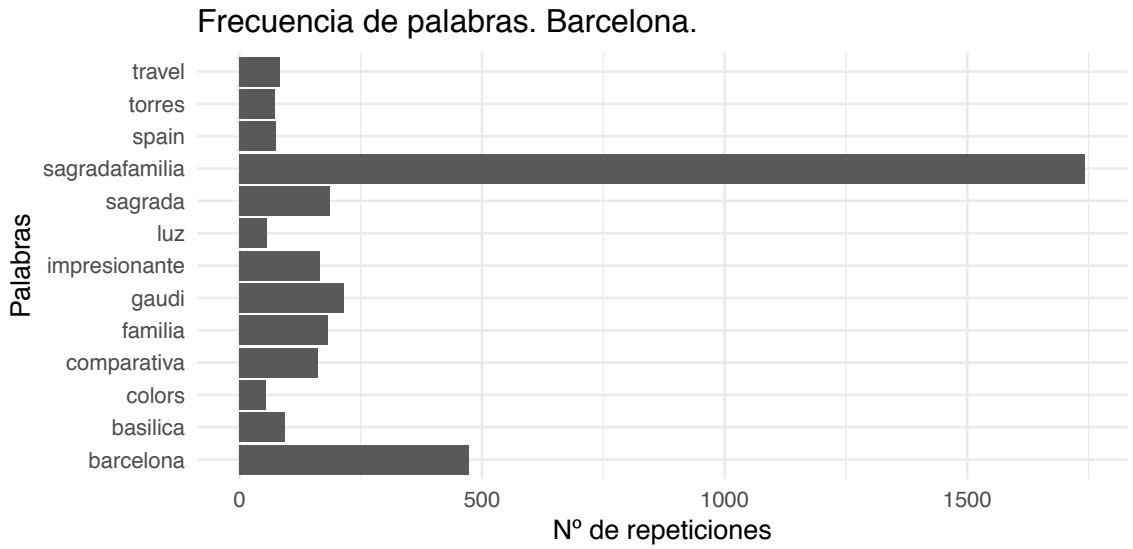


Figura 2.19: Frecuencia de palabras más utilizadas. Barcelona

A continuación, se tiene una nube de palabras.

Figura 2.20: Nube de palabras. Barcelona

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 52 veces. En la Tabla 2.9. se observa que por un lado, las palabras más utilizadas en Barcelona respecto al turismo son *sagradafamilia*, *Barcelona* y *Gaudi* con 1.743, 472 y 216 repeticiones, respectivamente. También cabe destacar las palabras *sagrada* y *familia* con 186 y 183 veces repetidas, respectivamente, por lo que finalmente, *sagrada familia* se pone con cerca de 2.000 repeticiones. Por otro lado, las palabras menos utilizadas son *torres*, *Spain* y *luz*con 76, 73 y 57 repeticiones, respectivamente.

A continuación, se tiene un gráfico en el que se establecen las relaciones entre las palabras formando diferentes grupos.

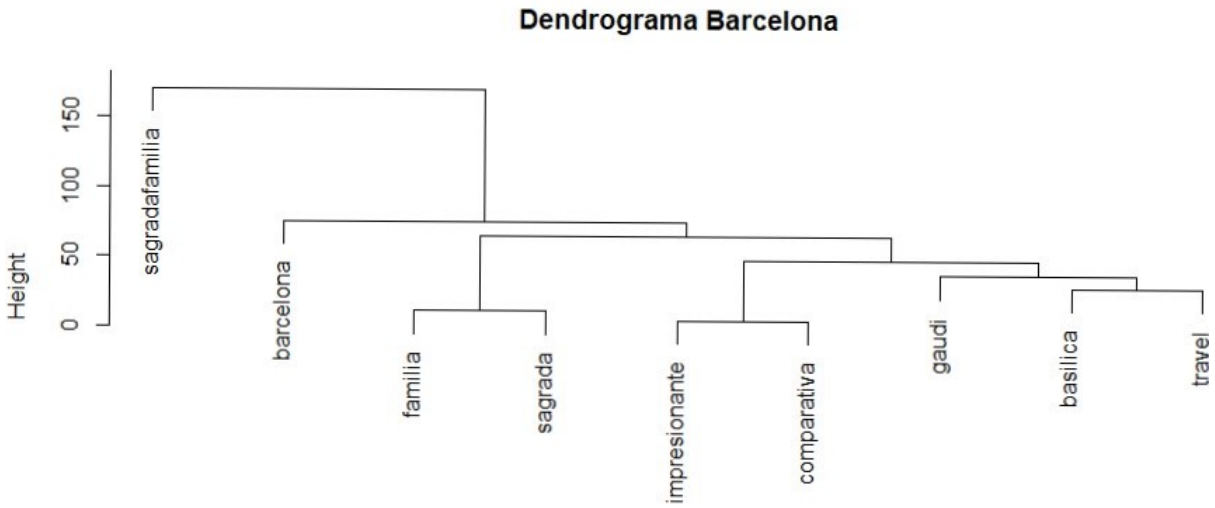


Figura 2.21: Dendrograma. Barcelona

Se puede observar que palabras como *impresionante* y *comparativa* están relacionadas con *Gaudi*, *Basílica* y *travel* que también están relacionadas en menor medida con *sagrada* y *familia*, esto es debido a que en Barcelona uno de los principales sitios turísticos es la Basílica de la Sagrada Familia de Antonio Gaudi. Por otro lado, se tiene *sagradafamilia* por separado, esto puede ser debido a que una de las cuentas estudiadas en Barcelona es *@ sagradafamilia* todo junto.

2.1.2.7. Valencia

Recordamos que las cuentas que se han analizado son *@ Valenciaturismo* y *@ c_valenciana* de forma conjunta. En la siguiente tabla, se observan las palabras que se han utilizado más en los tweets.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
valenciana	1.988	18,75	generalitat	335	3,16
valencia	1.618	15,26	felicidades	292	2,75
cvalenciana	1.465	13,81	felis	291	2,74
fallas	1.059	9,99	valenciaplaza	249	2,35
spain	738	6,96	electomania	248	2,34
renfe	728	6,86	canto	224	2,11
comunidades	362	3,41	billete	224	2,11
asturias	349	3,29	comunidad	219	2,07

Cuadro 2.10: Número de repeticiones y porcentaje de uso de cada palabra. Valencia.

En el siguiente gráfico se pueden observar las palabras más utilizadas y su respectiva frecuencia.

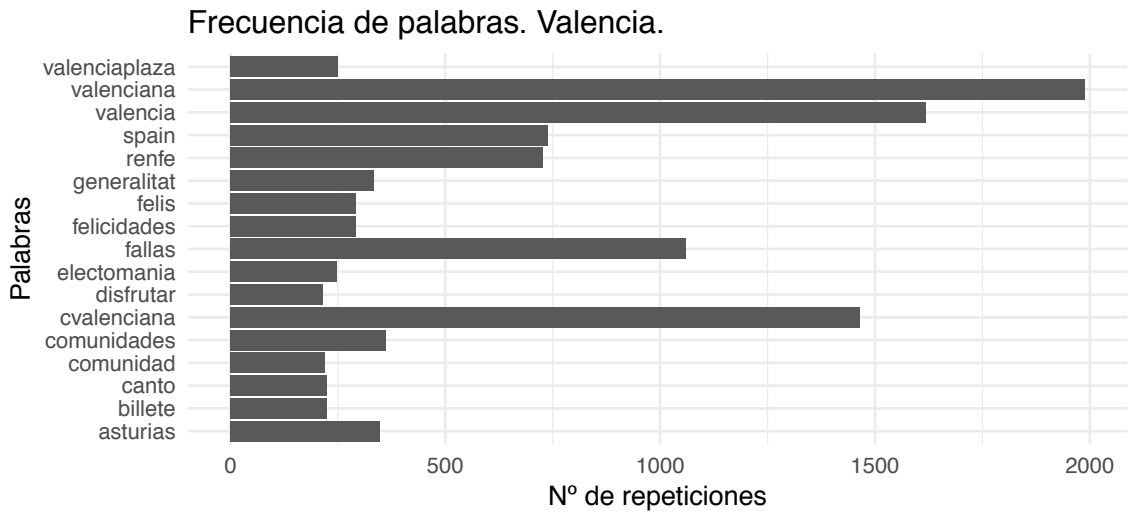


Figura 2.22: Frecuencia de palabras más utilizadas. Valencia

A continuación, se tiene una nube de palabras.

Figura 2.23: Nube de palabras. Valencia

Para hacer el análisis se ha seleccionado las palabras que se repiten más de 213 veces. En la Tabla 2.10. se observa que las palabras más utilizadas en Valencia respecto al turismo son *valenciana*, *Valencia* y *cvalenciana* con 1.988,1.618 y 1.465 repeticiones, respectivamente. También cabe destacar las palabras *fallas* y *electromanía*, ya que las Fallas de Valencia han tenido lugar en la primera quincena de marzo. Las palabras menos utilizadas son *billete*, *canto* y *comunidad* con 224, 224 y 219 repeticiones, respectivamente. Observando el gráfico de barras y la nube de puntos, también se observa que a parte de las palabras dichas anteriormente se destaca *renfe*, esto puede ser debido a que la ciudad de Valencia dispone de gran cantidad de comunicaciones ferroviarias con gran parte del resto de ciudades españolas.

A continuación, se tiene un gráfico en el que se comparan las relaciones entre palabras.

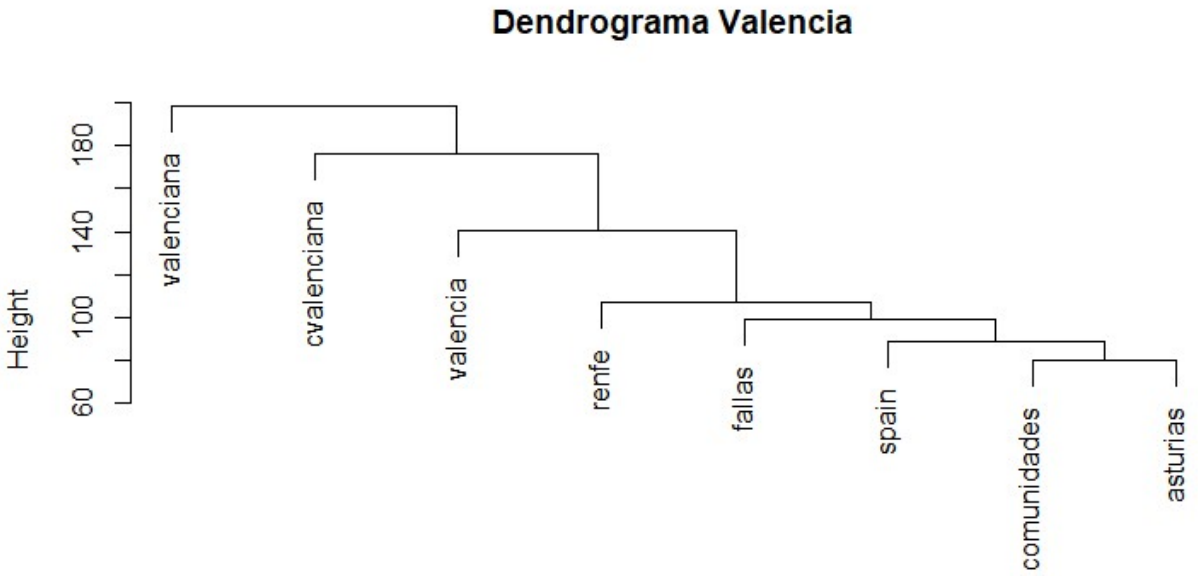


Figura 2.24: Dendrograma. Valencia

Se puede observar que se forman dos grupos, uno formado por *valenciana* y otro por *cvalenciana*, *valencia*, *renfe*, *fallas*, *Spain*, *comuniadades* y *Asturias*, este último grupo puede ser debido a la ampliación del número de trenes entre Valencia y las ciudades de las demás comunidades autónomas.

2.2. Análisis de sentimientos

El análisis de sentimientos es una tarea de clasificación masiva de documentos de manera automática en función de la connotación positiva o negativa del lenguaje utilizado en el fichero de datos, en este caso las palabras de los tweets. Se han analizado los tweets que se han recopilado de las menciones de las dos cuentas seleccionadas en la ciudad de Sevilla y posteriormente, se ha realizado una comparación con los tweets de las menciones de las distintas ciudades que se están estudiando

2.2.1. Sevilla

Para realizar el análisis de sentimientos se han catalogado cada una de las palabras en una escala de **-5** a **5** siendo **-5 muy negativa** y **5 muy positiva**. El siguiente conjunto de gráficas representa las palabras positivas y negativas realizadas en las menciones de cada cuenta y el número de tweets en el que aparecen.

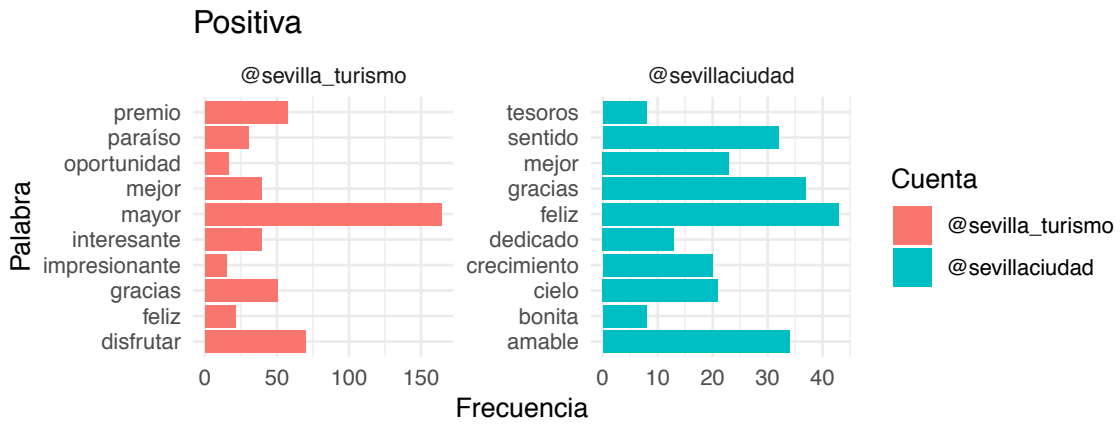


Figura 2.25: Frecuencia de palabras positivas por menciones de cuenta. Sevilla

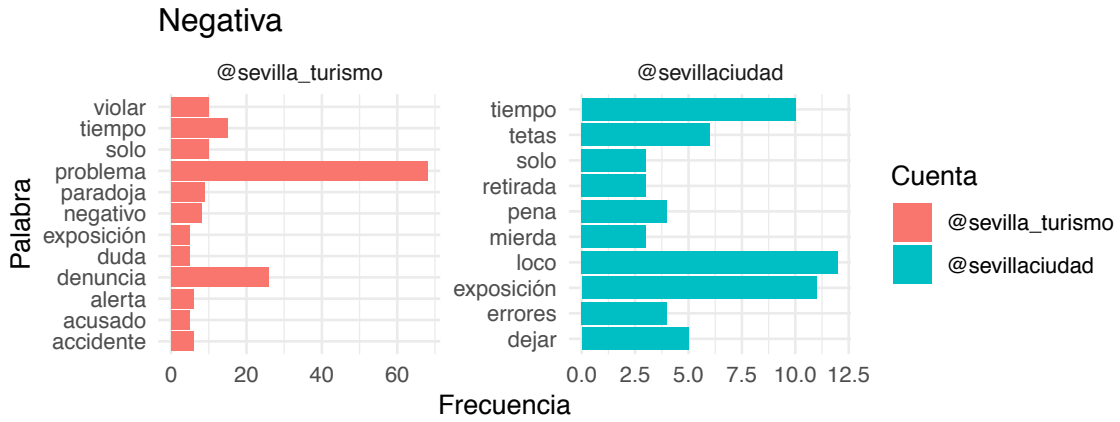


Figura 2.26: Frecuencia de palabras negativas por menciones de cuenta. Sevilla

Como se ha podido ver en las anteriores gráficas, en las menciones a ambas cuentas se utilizan un mayor número de palabras positivas.

A continuación, se va a calcular la media de positividad/negatividad de cada día estudiado, esto es de utilidad para ver si en alguno de los días se habla con mayor positividad o negatividad, no en el conjunto total como se calculará posteriormente con porcentajes. Para representar dichos calculos, en el siguiente gráfico se observa la media de positividad/negatividad de los tweets por día diferenciando las dos cuentas estudiadas.

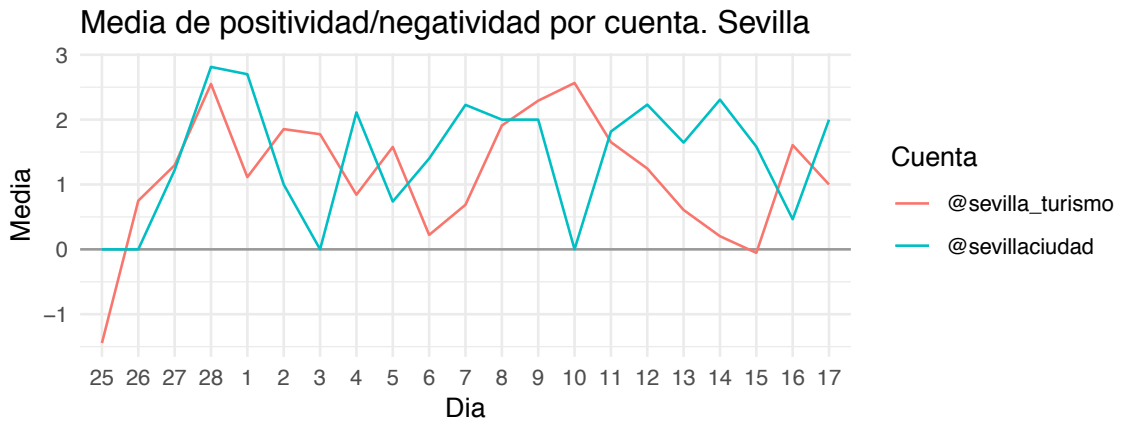


Figura 2.27: Media de positividad/negatividad de los tweets por día del mes y cuenta

Se observa que la media de la puntuación de ambas cuentas suele ser mayor a cero, es decir, de carácter positivo independientemente del día. Se observa que la cuenta de *@ sevilla_turismo* tiene puntuación media mayor a cero durante todos los días, a excepción del 25 de febrero y el 15 de marzo.

A continuación, se ha realizado un gráfico en el que se observa el porcentaje de tweets negativos y positivos de cada una de las cuentas.

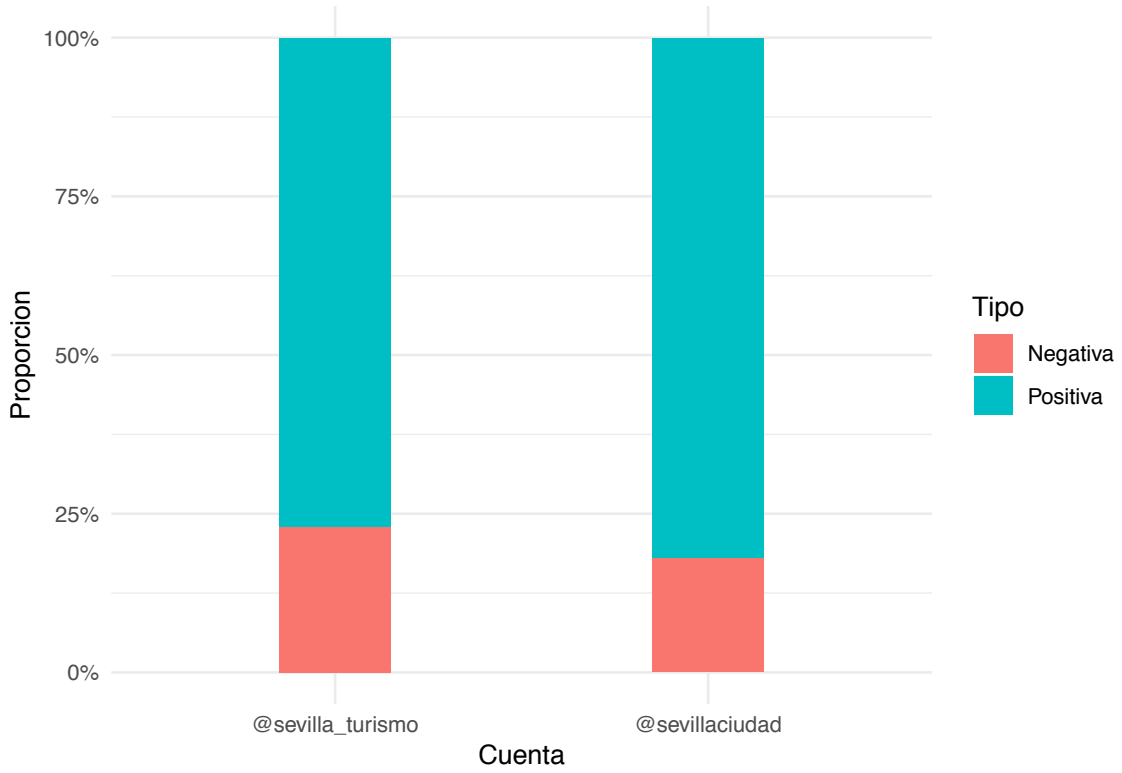


Figura 2.28: Porcentaje de tweets positivos y negativos por cuenta

Se observa que la cuenta *@ sevillaciudad* tiene un mayor número de tweets positivos que la cuenta de *@ sevilla_turismo*, sin embargo la diferencia es muy pequeña ya que

ambas tienen aproximadamente el 75 % de los tweets positivos. Como se observó en el gráfico anterior, la mayor parte de los tweets en ambas cuentas son de carácter positivo.

2.2.2. Análisis conjunto de Sevilla y ciudades turísticas estudiadas

Se ha realizado un análisis de sentimientos con las menciones para cada cuenta que se ha estudiado. Para cada ciudad, se ha analizado de forma conjunta los tweets que se han recopilado de las menciones de las cuentas estudiadas.

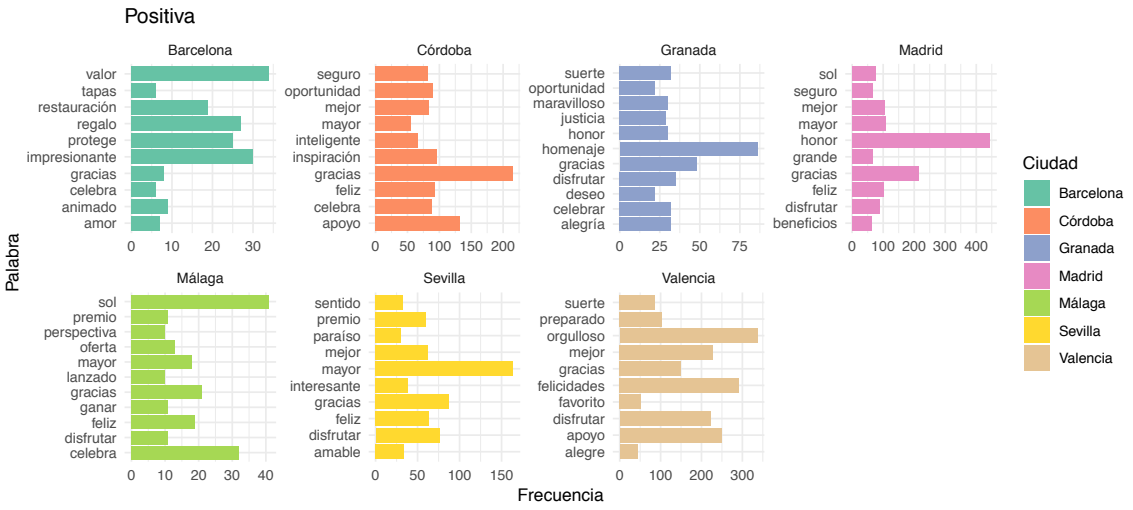


Figura 2.29: Frecuencia de palabras positivas por ciudad

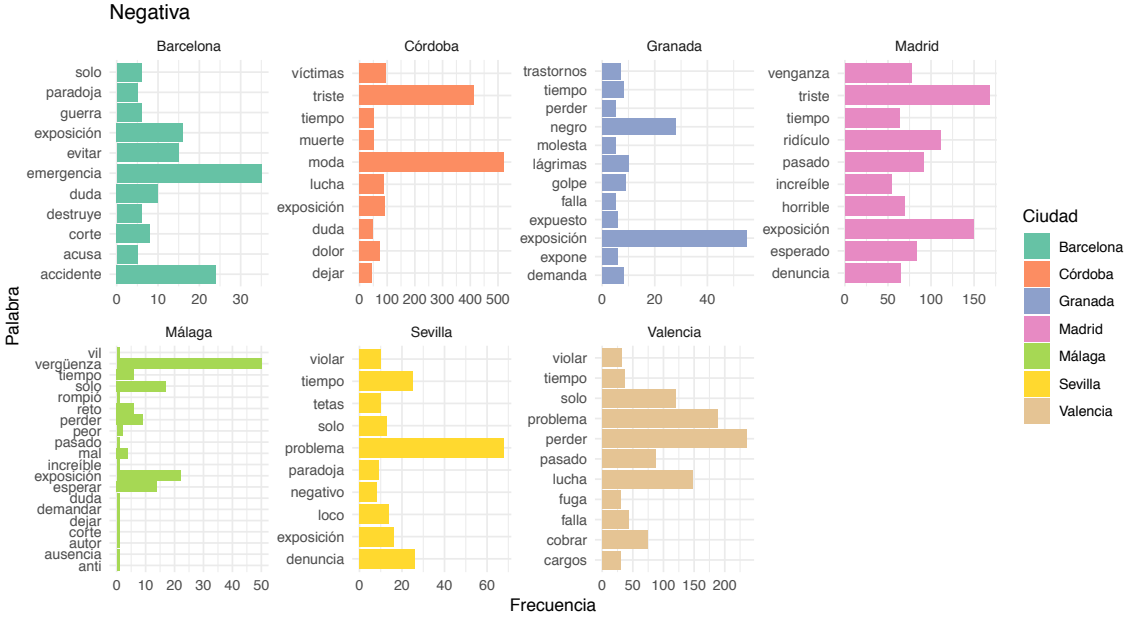


Figura 2.30: Frecuencia de palabras negativas por ciudad

En las gráficas anteriores, se puede observar que existe diferencia entre el número de

veces que aparecen las palabras en cada ciudad, esto es debido a que en unas cuentas se han escrito más tweets que en otras.

Si se atiende a las palabras positivas, en las ciudades de Córdoba, Granada, Madrid y Sevilla, utilizan *gracias* y en Córdoba y Valencia la palabra *apoyo*. A nivel general, palabras como *feliz*, *disfrutar* y demás palabras que expresan alegría se encuentran en casi todas las ciudades. También hay palabras como *mejor* y *mayor* que se utilizan en Sevilla, Córdoba y Madrid.

Si se atiende a las palabras negativas, en las ciudades de Valencia y Sevilla, utilizan *emergencia* y *accidente* y en Córdoba y en Madrid la palabra *triste*. Además, la palabra *tiempo* se encuentra en todas las ciudades excepto Barcelona. También, cabe destacar palabras como *denuncia*, *violar* y *víctimas*, que son palabras que se encuentran en Madrid, Córdoba, Sevilla y Valencia, esto puede ser debido a que en el período estudiado está el 8M, es decir, el 8 de marzo que es el día de la mujer.

El siguiente gráfico representa la media de positividad/negatividad de los tweets de cada día diferenciando las siete ciudades estudiadas.

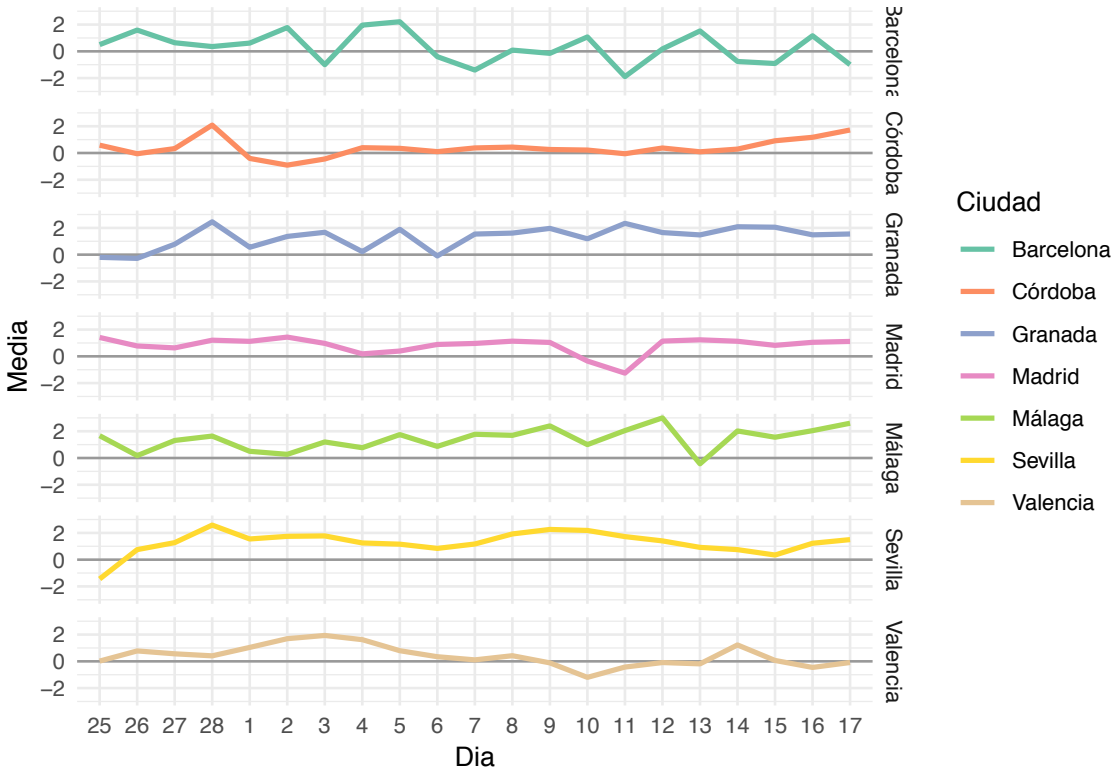


Figura 2.31: Media de positividad/negatividad de los tweets por día del mes y ciudad

Se observa que en ciudades como Granada, Madrid, Málaga y Sevilla, sobre todo hay menciones con palabras positivas, mientras que en Valencia y Barcelona va variando dependiendo del día. Además está Córdoba, que empieza como Valencia y Barcelona, pero a partir del día 4 de marzo la media de las menciones podría decirse que es neutra, es decir, hay aproximadamente el mismo número de tweets negativos que positivos.

A continuación, hay un gráfico en el que se observa el porcentaje de tweets negativos y positivos de cada una de las cuentas.

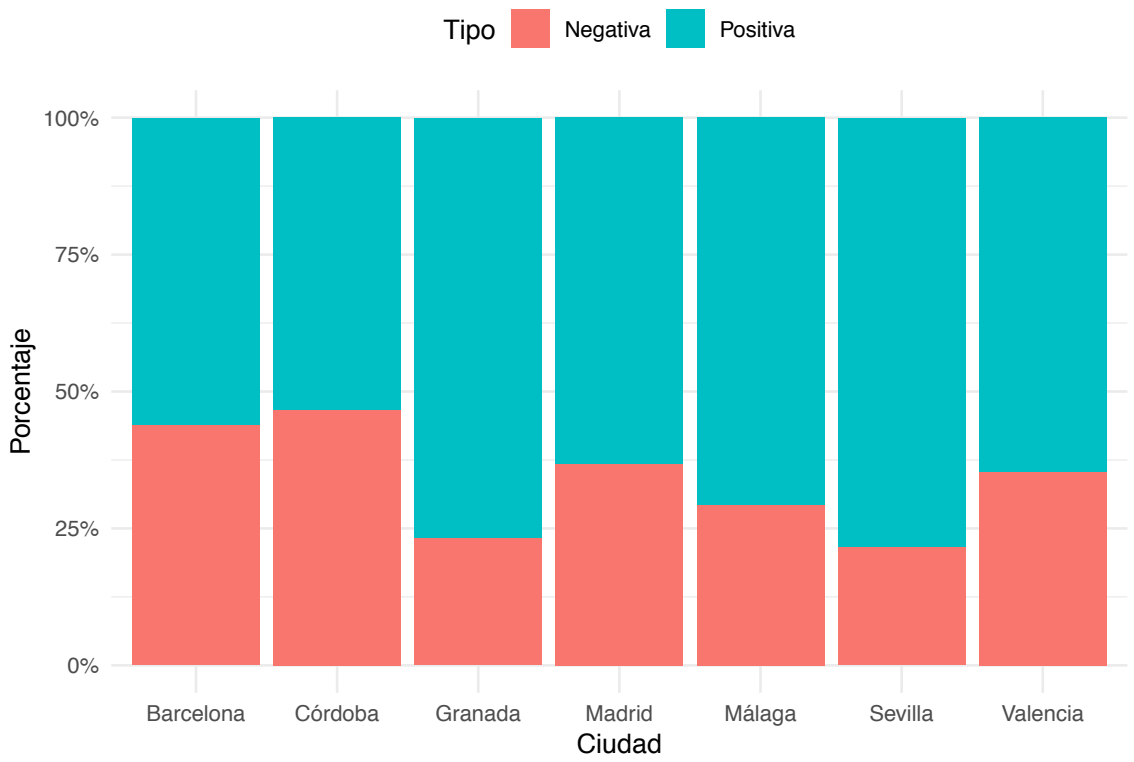


Figura 2.32: Porcentaje de tweets positivos y negativos por ciudad

Se puede observar que todas las ciudades tienen un mayor número de tweets positivos que negativos, sin embargo, Sevilla es la ciudad con mayor número de tweets de carácter positivo, seguida de Granada y Málaga. Además, se observa que Marid y Valencia tienen aproximadamente el mismo número de tweets positivos, pero es menor al de las ciudades citadas anteriormente. Finalmente, destaca Córdoba como la ciudad con menor número de tweets de carácter positivo, seguida de Barcelona. Ambas ciudades tienen mayor número de tweets de carácter positivo que negativo pero con muy poca diferencia.

2.3. Metodología

En este apartado se van a explicar los comandos de RStudio utilizados para la creación de los gráficos y tablas anteriores, así como la limpieza y la transformación de la base de datos descargada.

2.3.1. Análisis del número de tweets

Se cargan todos los datos de las dos cuentas de Sevilla. Se realiza el mismo procedimiento para las dos cuentas seleccionadas de cada una de las ciudades. Para ver la metodología, se cargan los datos de las cuentas de Sevilla y se crean las variables "Dia" y "Mes" a raíz de la variable *created*. Se selecciona para cada una de las cuentas los datos desde el 25 de febrero de 2019 al 17 de marzo de 2019.

```
datos_sevilla_turismo <- read.csv("datos_tweets_sevilla_turismo.csv")
datos_sevilla_turismo$created <- as.Date(
```

```
  datos_sevilla_turismo$created)
datos_sevilla_turismo$Dia <- as.numeric(
  format(datos_sevilla_turismo$created,"%d"))
datos_sevilla_turismo$Mes <- as.numeric(
  format(datos_sevilla_turismo$created,"%m"))

datos_sevilla_turismo_estudio <- datos_sevilla_turismo %>%
  filter(Mes==02 & between(Dia,25,28)|
           Mes==03 & between(Dia,1,17))

datos_sevillaciudad <- read.csv("datos_tweets_sevillaciudad.csv")
datos_sevillaciudad$created <- as.Date(datos_sevillaciudad$created)
datos_sevillaciudad$Dia <- as.numeric(
  format(datos_sevillaciudad$created,"%d"))
datos_sevillaciudad$Mes <- as.numeric(
  format(datos_sevillaciudad$created,"%m"))

datos_sevillaciudad_estudio <- datos_sevillaciudad %>%
  filter(Mes==02 & between(Dia,25,28)|
           Mes==03 & between(Dia,1,17))
```

Después de cargar los datos, se utiliza la variable "created" para obtener las variables "dia", "mes", "año" y "dia_semana". A continuación, se unen los data.frame de cada ciudad añadiendo una columna llamada "Cuenta" que diferencie cada una de las cuentas.

```
fecha_tweets_1 <- datos_sevilla_turismo_estudio$created
fecha_tweets_2 <- datos_sevillaciudad_estudio$created
cuenta <- c(rep("@sevilla_turismo",length(fecha_tweets_1)),
            rep("@sevillaciudad",length(fecha_tweets_2)))
dia <- as.numeric(format(c(fecha_tweets_1,fecha_tweets_2),"%d"))
mes <- as.numeric(format(c(fecha_tweets_1,fecha_tweets_2),"%m"))
año <- as.numeric(format(c(fecha_tweets_1,fecha_tweets_2),"%Y"))
dia_semana <- factor(format(c(fecha_tweets_1,fecha_tweets_2),"%A"),
                     levels = c("lunes","martes","miércoles","jueves",
                                "viernes","sábado","domingo"),
                     labels=c("Lunes", "Martes","Miércoles","Jueves",
                              "Viernes","Sábado","Domingo"))
tabla_sevilla <- data.frame(cuenta,dia,mes,año,dia_semana)
```

Después de realizar el mismo procedimiento para las 7 seleccionadas, se unen los data.frame de cada una de las ciudades y se crea una nueva variable llamada "Ciudad" que diferencia a qué ciudad se refiere cada tweet.

```
tabla <- rbind(tabla_barcelona, tabla_cordoba, tabla_granada,
               tabla_madrid,tabla_malaga, tabla_sevilla,
               tabla_valencia)
tabla$Ciudad <- factor(c(rep("Barcelona", nrow(tabla_barcelona)),
               rep("Cordoba",  nrow(tabla_cordoba)),
               rep("Granada", nrow(tabla_granada)),
```

```
                rep("Madrid", nrow(tabla_madrid)),
                rep("Málaga", nrow(tabla_malaga)),
                rep("Sevilla", nrow(tabla_sevilla)),
                rep("Valencia", nrow(tabla_valencia))
                ))
```

Por lo tanto, se tiene un data.frame en el que se tiene el número de tweets por ciudad, cuenta, día, mes, día de la semana y año.

A continuación, se crea un gráfico en el que se pueda observar el número de tweets por ciudad.

```
g_tweet_ciudad <- ggplot(tabla, aes(Ciudad,fill = Ciudad)) +
  geom_bar() +
  xlab("") +
  ylab("Número de tweets") +
  ggtitle("Total de tweets por ciudad*") +
  labs(fill = "Ciudades") +
  scale_fill_brewer(palette="Set2")+
  scale_y_continuous(breaks=seq(0, 12000, 1000))+
  theme_minimal()
```

A continuación, se crea una tabla en la que se represente el número de tweet por cuenta y ciudad.

```
grupos <- group_by(tabla, cuenta, Ciudad)
n_cuenta_ciudad <- summarise(grupos,
          n_tweets = n()
)
n_cuenta_ciudad_tabla <- as.data.frame(n_cuenta_ciudad)
colnames(n_cuenta_ciudad_tabla) <- c("Cuenta", "Ciudad",
                                     "Número de tweets")
n_cuenta_ciudad_tabla_dis <- format(n_cuenta_ciudad_tabla,
                                    decimal.mark=",",big.mark=".",
                                    scientific=FALSE)
```

A continuación, se crea una tabla en la que se vea el número de tweets por día de la semana.

```
n_diasemana <- tapply(tabla$dia, tabla$dia_semana,length)
n_diasemana <- data.frame(names(n_diasemana), n_diasemana)
colnames(n_diasemana) <- c("Día de la semana","Número de tweets")
n_diasemana_dis<-format(n_diasemana, decimal.mark=",",
                        big.mark=".", scientific=FALSE)
```

Se crea un gráfico en el que se pueda observar el número de tweets por dia de la semana y ciudad. Y otro gráfico en el que se observe el número de tweet por día del mes y ciudad.

```
g_diasemana_ciudad <- ggplot(data = n_diasemana_ciudad,
                             aes(x=dia_semana, y=n_tweets,
                                 group=Ciudad, colour=Ciudad)) +
  geom_line() +
```

```
  geom_point() +
  geom_abline(intercept=9, slope=0.5)+
  ggtitle("Número de tweets de cada ciudades por día de la semana")+
  xlab("Día de la semana") +
  ylab("Número de tweets") +
  theme_minimal()

g_dia_ciudad <- ggplot(data = n_dia_ciudad,
                       aes(x=dia, y=n_tweets, group=Ciudad,
                           colour=Ciudad)) +
  geom_line() +
  geom_point() +
  xlab("Día") +
  ylab("Número de tweets") +
  ggtitle("Total tweets por ciudad* y por día")+
  scale_fill_brewer(palette="Set2","Día de la semana")+
  scale_x_discrete(breaks=seq(1, 28, 1))+
  scale_y_continuous(breaks=seq(0, 1300, 100))+
  scale_color_brewer(palette="Set2","Ciudad")+
  theme_minimal()
```

Para calcular el número de tweets por dispositivo se hace lo siguiente. Se selecciona la variable "source" de la base de datos, después, dentro de source se selecciona todo lo que venga despues de Twitter ya que la forma en la que viene expresado el tweet viene de esta manera. A continuación, se suma el número de tweets por dispositivos para tener el total

```
dispositivo_sevilla_turismo <-datos_sevilla_turismo_estudio %>%
  select(statusSource) %>%
  extract(statusSource, "source", "Twitter (.*?)<") %>%
  group_by(source) %>%
  summarise(n_dispositivo = n())
# Le cambiamos el nombre a las variable
colnames(dispositivo_sevilla_turismo) <- c("Dispositivo", "Número tweets")
# Renombramos una variable vacía
dispositivo_sevilla_turismo[1,1] <- "Ordenador"
# Creamos dos nuevas variables que son Cuenta y Ciudad para
# la tabla final
dispositivo_sevilla_turismo$Cuenta <-
  rep("@sevilla_turismo", nrow(dispositivo_sevilla_turismo))
dispositivo_sevilla_turismo$Ciudad <-
  rep("Sevilla", nrow(dispositivo_sevilla_turismo))
```

Se realiza el mismo procedimiento para cada una de las cuentas estudiadas. Una vez realizado todo, se unen todas las base de datos en una llamada "dispositivos".

```
dispositivos <- rbind(dispositivo_sevilla_turismo,
                      dispositivo_sevillaciudad,
                      dispositivo_cordobaESP,
                      dispositivo_patiosdecordoba,
```

```
                        dispositivo_vivecostadelsol,
                        dispositivo_turismodemalaga,
                        dispositivo_alhambracultura,
                        dispositivo_turgranada,
                        dispositivo_turismoMadrid,
                        dispositivo_visita_madrid,
                        dispositivo_turismoBCN,
                        dispositivo_sagradafamilia,
                        dispositivo_c_valenciana,
                        dispositivo_Valenciaturismo)
```

Se cambia la columna de "Dispositivo" en muchas columnas, una por cada campo tipo de dispositivo

```
h <- dispositivos %>%
  spread(Dispositivo, `Número tweets`) %>%
  arrange(Ciudad)
```

A continuación, se crea una funcion que sustituye NA por 0, ya que es posible que en alguna de las cuentas no exista uno de los tipos de dispositivos. DespuSe aplica la funcion al data.frame y se pasan los valores a número entero para poder tratar con ellos.

```
hacer_cero <- function(x){
  ifelse(is.na(x),0,x)}

dispositivos=data.frame(sapply(h,hacer_cero))

for(i in 3:ncol(dispositivos)){
  dispositivos[,i] <- as.integer(as.character(dispositivos[,i]))
}
```

Por un lado, se unen las filas con pocos datos en el campo "otros" y por otro lado, se unen "Web.App" y "Web.Client" en una llamada "Web". Finalmente se modifica el nombre de las variables y se da formato a los números para mejorar la estética.

```
dispositivos <- dispositivos %>%
  mutate(Otros=Ads + `Ads.Composer` + Assistant + `editor.Andalucia` +
           `editor.Turismo` + `for.BlackBerry.`+ for.BlackBerry +
           `Media.Dashboard` + `Media.Studio` + Ordenador +
           Thing.App + Update.News,
         Web = `Web.App` + `Web.Client`,
         for.Android = for..Android+for.Android) %>%
  select(Ciudad, Cuenta, `for.Android`, `for.iPhone`,
         `for.iPad`, Web, Otros)

colnames(dispositivos) <- c("Ciudad", "Cuenta", "Android", "iPhone",
                            "iPad","Web","Otros")

dispositivos_dis <-format(dispositivos, decimal.mark=",",
                          big.mark=".", scientific=FALSE)
```

2.3.2. Análisis por cuentas de twitter

En este apartado, la metodología utilizada va a centrarse en el caso de Sevilla, pero el procedimiento es similar en todas las cuentas, a excepción de la eliminación de palabras sin significado, que en cada cuenta hay unas distintas. Para empezar, se leen todos los datos de sevilla y se crean las variables "Dia" y "Mes" a raíz de la variable "created". Posteriormente se seleccionan los tweets de los días estudiados y se crea ña variable "texto_sevilla" en la que se encuentra el contenido de todos los tweets seleccionados.

```
datos_sevilla <- read.csv("datos_tweets_sevilla.csv")
datos_sevilla$created <- as.Date(datos_sevilla$created)
datos_sevilla$Dia <- as.numeric(format(datos_sevilla$created,"%d"))
datos_sevilla$Mes <- as.numeric(format(datos_sevilla$created,"%m"))

datos_sevilla_estudio <- datos_sevilla %>%
  filter(Mes==02 & between(Dia,25,28)|
         Mes==03 & between(Dia,1,17))

texto_sevilla <- datos_sevilla_estudio$text
```

Se carga la librería "tm" para poder crear el "corpus" con el texto. Y posteriormente se realiza la limpieza de la base de datos quitando enlaces, espacios en blancos innecesarios, transformar todas las letras en minúsculas y eliminar las palabras que carezcan de significado.

```
library(tm)
sevilla_corpus <- Corpus(VectorSource(texto_sevilla))
# Se eliminan los enlaces.
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeURL))
# Se eliminan los espacios en blanco.
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeNumPunct))
sevilla_corpus <- tm_map(sevilla_corpus, stripWhitespace)
# Se transforman todas las letras a minúsculas.
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(tolower))
# Se eliminan todas las palabras sin significado en español y en inglés,
# y demás palabras que carezcan de significado para el estudio.
myStopwords <- c(stopwords('spanish'),stopwords('english'),
                 "juanespadassvq","fufuub","fufuu","sevillaciudad",
                 "sevilla_turismo","sudigastro","one","know","cordobaesp",
                 "aquí","viveandalucia","uufef","spainurban",
                 "salporsevilla","prodetur","maría","maicarivera",
                 "marzo","gracias", "año","hoy","wttc","vendrá")

sevilla_corpus <- tm_map(sevilla_corpus, removeWords, myStopwords)
```

Continuando con el proceso de depuración, se eliminan los acentos.

```
removeAccents <- function(x) {
  gsub("á","a",x)
}
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeAccents))

removeAccents <- function(x) {
  gsub("é","e",x)
}
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeAccents))

removeAccents <- function(x) {
  gsub("í","i",x)
}
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeAccents))

removeAccents <- function(x) {
  gsub("ó","o",x)
}
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeAccents))

removeAccents <- function(x) {
  gsub("ú","u",x)
}
sevilla_corpus <- tm_map(sevilla_corpus,
                         content_transformer(removeAccents))
```

Para acabar el proceso de depuración de la base de datos se agrupan las palabras de igual significado. Por ejemplo, las palabra "spain" y "españa" significan lo mismo pero en diferente idioma, por lo que se pueden unir.

```
removeWord <- function(x) {
  gsub("locations","lugar",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("spain","españa",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("cultural","cultura",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))
```

```
removeWord <- function(x) {
  gsub("sevillatourism","sevillaturismo",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("electrico","patinete",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("españacul ","cultura",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("sevillaturismo","turismo",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("viajes","viaje",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))

removeWord <- function(x) {
  gsub("culturatura","cultura",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))


removeWord <- function(x) {
  gsub("seville","sevilla",x)
}
sevilla_corpus <- tm_map(sevilla_corpus, content_transformer(removeWord))
```

Se crea la TermDocumentMatrix y se transforma a matriz. Después, se obtiene la frecuencia de cada una de las palabras y se seleccionan las que se repitan más del 2,5 % del número total de tweets. Se crea un data.frame con dichas palabras y frecuencias, y se calcula el porcentaje que ocupa esa palabra respecto a las demás.

```
sevilla_tdm = TermDocumentMatrix(sevilla_corpus,
                          control = list(removePunctuation = T,
                                         removeNumbers = TRUE))

sevilla_m = as.matrix(sevilla_tdm)

sevilla_word_freqs = sort(subset( rowSums(sevilla_m),
```

```
                                    rowSums(sevilla_m) >=
                                      nrow(datos_sevilla_estudio)*0.025),
                            decreasing=TRUE)

sevilla_dm_1 = data.frame(word=names(sort(sevilla_word_freqs,
                                          decreasing=T)),
                  freq=sort(sevilla_word_freqs,decreasing = T))
sevilla_dm_1[,"freq"] = as.integer(sevilla_dm_1[,"freq"])
sevilla_dm_1 <- sevilla_dm_1 %>%
  mutate(perc = round((freq/sum(freq))*100,2))

colnames(sevilla_dm_1) <- c("Palabra","Nº de repeticiones",
                            "Porcentaje de uso")
```

Se realiza una modificación para mejorar el aspecto de la tabla, es decir, para que la tabla se divida en dos partes y además, poner "." cuando se refiere a "big.mark" y "," cuando se refiere a "decimal.mark".

```
sevilla_dm_1_modificado <-
  data.frame(sevilla_dm_1[1:(nrow(sevilla_dm_1)/2),],
             sevilla_dm_1[(nrow(sevilla_dm_1)/2+1):nrow(sevilla_dm_1),])
colnames(sevilla_dm_1_modificado) <- c("Palabra","Repeticiones","%",
                                       "Palabra","Repeticiones","%")
sevilla_dm_1_modificado_dis<-format(sevilla_dm_1_modificado,
                                    decimal.mark=",",
                                    big.mark=".", scientific=FALSE)
```

Se realiza un gráfico en el que se vea la frecuencia de las palabras más utilizadas utilizando el paquete "ggplot2", paquete que pertenece al paquete "tidyverse".

```
library(ggplot2)
g_frec_sevilla_1 <- ggplot(sevilla_dm_1,
       aes(x=names(sevilla_word_freqs),
           y=sevilla_word_freqs)) +
  ggtitle("Frecuencia de palabras. Sevilla")+
  geom_bar(stat="identity") +
  xlab("Palabras") +
  ylab("Nº de repeticiones") +
  coord_flip() +
  theme(axis.text=element_text(size=7))+
  theme_minimal()
```

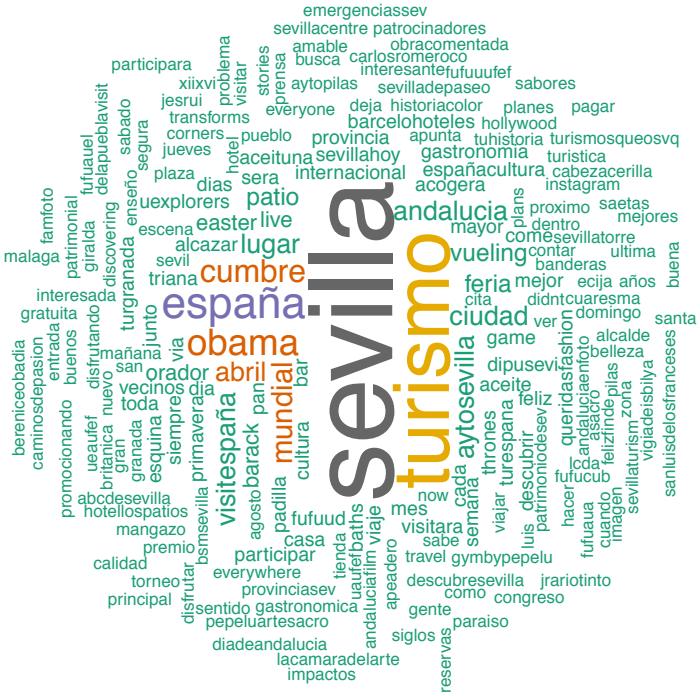
Se realiza la nube de palabras, en la que como máximo se incluirán 50 palabras.

```
library(wordcloud)
library(RColorBrewer)
wordcloud(sevilla_dm_1$Palabra, max.words = 50,
          sevilla_dm_1$`Nº de repeticiones`, random.order=F,
          colors=brewer.pal(8, "Dark2"),rot.per=0.35)
```

Se puede realizar la nube de palabras de otra forma en la que se reflejen la totalidad de las palabras, pero desde mi punto de vista se ve más claro y se obtiene más información de la anterior forma. Aún así, dicho gráfico se realizaría de la siguiente manera.

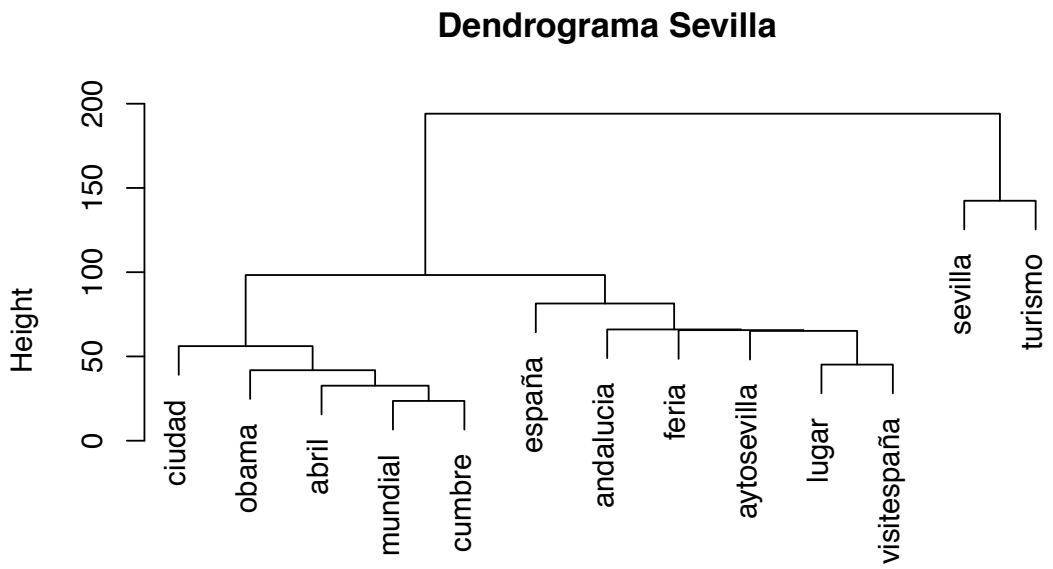
Se calculan las frecuencias de las palabras, pero en este caso no seleccionamos las más frecuentes, sino que posteriormente, con la función que utilizamos para la creación de la nube de palabras (wordcloud) se seleccionan las 200 palabras más frecuentes.

```
sevilla_word_freqs_nueva = sort(rowSums(sevilla_m),
                                decreasing=TRUE)

sevilla_dm_nueva = data.frame(word=names(sort(sevilla_word_freqs_nueva,
                                              decreasing=T)),
                              freq= sort(sevilla_word_freqs_nueva,
                                         decreasing = T))
sevilla_dm_nueva[,"freq"] = as.integer(sevilla_dm_nueva[,"freq"])
sevilla_dm_nueva <- sevilla_dm_nueva %>%
  mutate(perc = round((freq/sum(freq))*100,2))

colnames(sevilla_dm_nueva) <- c("Palabra","Nº de repeticiones",
                                "Porcentaje de uso")

wordcloud(sevilla_dm_nueva$Palabra, max.words = 200,
          sevilla_dm_nueva$`Nº de repeticiones`,
          random.order=F, colors=brewer.pal(8, "Dark2"),
          rot.per=0.35)
```



Finalmente, se crea un dendrograma, para ello, se realiza un análisis de cluster utilizando el método de "Ward.D".

```
sevilla_tdm2=removeSparseTerms(sevilla_tdm,sparse=0.96)
sevilla_m2=as.matrix(sevilla_tdm2)
sevilla_distMatrix=dist(scale(sevilla_m2))
sevilla_fit_1=hclust(sevilla_distMatrix, method="ward.D")

plot(sevilla_fit_1, main = "Dendrograma Sevilla", sub = "", xlab = "")
```



2.3.3. Análisis de sentimientos

La metodología utilizada en el análisis de sentimientos se va a explicar con el ejemplo de la comparación de las dos cuentas de Sevilla. Como ya se ha dicho anteriormente, se va a utilizar uno de los enfoques semánticos.

Para realizarlo, primero se tiene que transformar la base de datos añadiendo variables para su posterior manejo. Se crea la variable "Cuenta", que incluye las cuentas correspondientes a la ciudad de Sevilla, de esta forma, tendremos los tweets de los días estudiados con la cuenta a la que pertenece.

```
cuenta <- c(rep("@sevilla_turismo",nrow(datos_sevilla_turismo_estudio)),
            rep("@sevillaciudad",nrow(datos_sevillaciudad_estudio)))
datos <- rbind(datos_sevilla_turismo_estudio,datos_sevillaciudad_estudio)
datos <- cbind(datos, cuenta)
```

Se carga el fichero "lexico_afinn.en.es.csv" que es el diccionario AFINN que como ya se ha dicho anteriormente es el que se va a utilizar para la realización del análisis de sentimientos. Y posteriormente, se separa la variable "created" en Año, Mes y Día.

```
afinn <- read.csv("lexico_afinn.en.es.csv",
                  stringsAsFactors = F, fileEncoding = "latin1") %>%
  tbl_df()

tweets <- datos %>%
  separate(created, into = c("Anho", "Mes", "Día"), sep = "-",
           remove = FALSE) %>%
  mutate(text = tolower(text),
```

```
        # Se transforma a factor la variable "Dia" para que los días
        # al ser de diferente mes salgan en orden cronológico.
        Dia = factor(Dia),
        Dia = factor(Dia, levels(Dia)[c(seq(18,21,1),seq(1,17,1))]))
```

Se convierten los tweets en palabras, es decir, se separa todo en palabras individuales para después compararlas con el fichero AFINN y finalmente se crean columnas que nos digan si es sentimiento positivo o negativo según la puntuación que tenga la palabra.

```
tweets_afinn <-
  tweets %>%
  unnest_tokens(input = "text", output = "Palabra") %>%
  inner_join(afinn, ., by = "Palabra") %>%
  mutate(Tipo = ifelse(Puntuacion > 0, "Positiva", "Negativa")) %>%
  rename("Cuenta" = cuenta)
```

Se eliminan las palabras "sí" y "no" ya que su presencia no tiene por qué determinar si el tweet es positivo o no. Posteriormente, se unen las palabras de igual significado, al igual que ya se ha hecho anteriormente.

```
tweets_afinn <-
  tweets_afinn %>%
  filter(Palabra != "no") %>%
  filter(Palabra != "sí")

removeWord <- function(x) {
  gsub("premios","premio",x)
}
tweets_afinn$Palabra<- removeWord(tweets_afinn$Palabra)

removeWord <- function(x) {
  gsub("disfrutando","disfrutar",x)
}
tweets_afinn$Palabra<- removeWord(tweets_afinn$Palabra)
```

Se agrupan las palabras por tweets y se pone una puntuación a cada tweet.

```
tweets <-
  tweets_afinn %>%
  group_by(id) %>%
  # se hace la media de las puntuaciones de las palabras de cada tweet.
  summarise(Puntuacion_tweet = mean(Puntuacion)) %>%
  # se unen los tweets con su respectiva puntuación
  left_join(tweets, ., by = "id") %>%
  # Si hay algún tweet sin puntuación, es decir, que no tenga palabras
  # catalogadas, se le da media 0, ya que es un tweets sin
  # carácter positivo ni negativo.
  mutate(Puntuacion_tweet = ifelse(is.na(Puntuacion_tweet),
                                   0, Puntuacion_tweet)) %>%
  rename("Cuenta" = cuenta)
```

Se crean dos gráficos en el que se observe por un lado las palabras positivas y su frecuencia por cada cuenta, y por otro lado, las palabras negativas y su frecuencia por cuenta.

```
g_pos_neg_sevilla <- map(c("Positiva", "Negativa"),
                         function(sentimiento) {
  tweets_afinn %>%
    filter(Tipo ==  sentimiento) %>%
    group_by(Cuenta) %>%
    count(Palabra, sort = T) %>%
    top_n(n = 10, wt = n) %>%
    ggplot() +
    aes(Palabra, n, fill = Cuenta) +
    ylab("Frecuencia")+
    geom_col() +
    facet_wrap("Cuenta", scales = "free") +
    coord_flip() +
    labs(title = sentimiento) +
    theme_minimal()
})
```

Se agrupan los tweets por fecha y cuenta, para ver la positividad o negatividad de cada una de los días estudiados.

```
tweets_afinn_fecha <-
  tweets_afinn %>%
  group_by(id) %>%
  mutate(Suma = mean(Puntuacion)) %>%
  group_by(Cuenta, Dia) %>%
  summarise(Media = mean(Puntuacion))
```

Se añaden valores 0 a los días 25 y 26 y 3 de la cuenta @ sevillaciudad, ya que no hay palabras consideradas ni positivas ni negativas, por lo que la media en esos días es nula.

```
tweets_afinn_fecha <- tweets_afinn_fecha%>%
  complete(Dia)
tweets_afinn_fecha$Media[is.na(tweets_afinn_fecha$Media)] <- 0
```

Se crea un gráfico en el que se observe la positividad/negatividad de cada uno de los días estuadiados, separado por cuentas.

```
g_media_cuentas_sevilla <- tweets_afinn_fecha %>%
  ggplot() +
  ggtitle("Media de positividad/negatividad por cuenta. Sevilla")+
  aes(Dia, Media, color = Cuenta, group=Cuenta) +
  geom_hline(yintercept = 0, alpha = .35) +
  geom_line() +
  theme_minimal() +
  scale_fill_brewer("Día de la semana")
```

Para finalizar el análisis de sentimientos, se crea un gráfico en el que se observe la proporción de tweets negativos y positivos por cada una de las cuentas.

```
g_proporcion_sevilla <- tweets_afinn %>%
  count(Cuenta, Tipo) %>%
  group_by(Cuenta) %>%
  mutate(Proporcion = n / sum(n)) %>%
  ggplot() +
  ylab("Porcentajae")+
  aes(Cuenta, Proporcion, fill = Tipo,group=1) +
  geom_bar(stat = "identity",width=0.3) +
  scale_y_continuous(labels = percent_format()) +
  theme_minimal() +
  theme(legend.position = "right")
```

Capítulo 3

Análisis de la actualidad política en Twitter: Elecciones generales de España, 28 abril de 2019

Uno de los principales motivos por el que se ha realizado un análisis de la política electoral es porque es un tema social de actualidad, que desde mi punto de vista, es de vital importancia que la gente joven esté interesada en él. Además, las elecciones generales se celebraron el 28 de abril de 2019, fecha que encajó a la perfección con la programación del trabajo.

El estudio se ha centrado en la comparación de las cuentas de los representantes de los partídos políticos seleccionados. Las cuentas seleccionadas son las siguientes:

1. Pablo Iglesias: Cuenta: @ Pablo_Iglesias_. Representante del partido Unidas Podemos.

2. Pedro Sánchez: Cuenta: @ sanchezcastejon. Representante del Partido Socialista Obrero Español (PSOE).

3. Albert Rivera: Cuenta: @ Albert_Rivera. Representante del partido Ciudadanos.

4. Pablo Casado: Cuenta: @ pablocasado_. Representante del Partido Popular (PP).

5. Santiago Abascal: Cuenta: @ Santi_ABASCAL. Representante del partido VOX.

Se han descargado los tweets desde el día 12 de abril hasta el día 5 de mayo, es decir, desde que comenzó la campaña electoral hasta una semana despúes de la realización de las elecciones. Pero sólo se han seleccionado para hacer el análisis los tweets de los días 22 y 23 de abril, que son los días en los que se realizó el debate en Atresmedia y en RTVE, respectivamente. También se han estudiado los tweets del día 28 de abril, es decir, el día en el que se produjeron las elecciones.

3.1. Análisis de tweets

En este apartado se ha realizado un análisis del número de tweets que se ha recopilado de las menciones que se han realizado de las diferentes cuentas. Se ha analizado el número de tweets por cuenta de cada representante de los partidos seleccionados, por día del mes y el correspondiente dispositivo de uso. Además, también, se ha analizado qué palabras/temas

se han utilizado en las diferentes cuentas que se han analizado, si están relacionadas entre sí y el posible motivo de dicha relación.

3.1.1. Número de tweets

Para poder realizar el análisis del número de tweets se tiene que crear una tabla para cada representante en la que se pueda observar al representante del partido, el día en el que se creó el tweet y el número de tweets correspondientes para cada día. Para así, posteriormente poder unirlas.

En la siguiente gráfica, se puede observar el número de tweets relacionados con cada una de las cuentas de los representantes de los partidos.

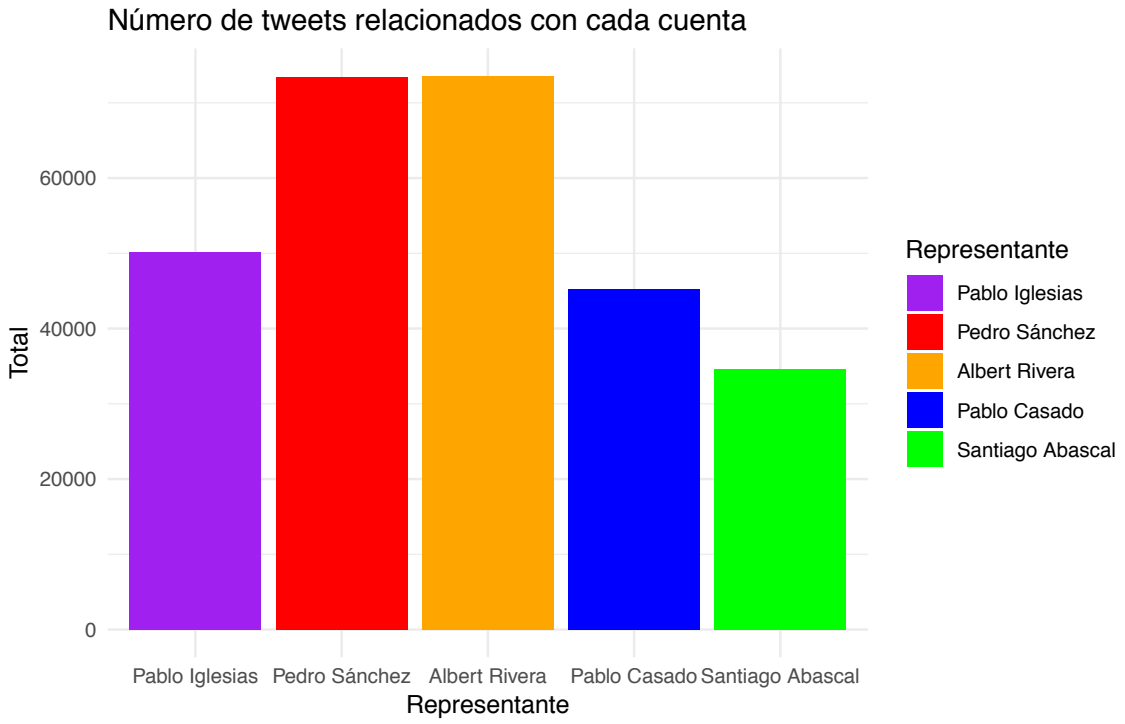


Figura 3.1: Número de tweets por cuenta

Se puede observar que los representantes que tienen mayor número de tweets relacionados con ellos son Pedro Sánchez y Albert Rivera, seguidos de Pablo Iglesias, Pablo Casado y finalmente Santiago Abascal. Esto puede ser debido a que como ya se ha dicho anteriormente, los días 22 y 23 de abril hubo debates en los que no participó Santiago Abascal, por lo que se habló con más frecuencia de los demás representantes.

A continuación, se tiene el número de tweets por representantes de los partidos y día estudiado.

Se observa que durante el día 22 hubo una actividad similar en relación a Pedro Sánchez y Albert Rivera. Sin embargo, el día 23 hubo mayor actividad en relación a Albert Rivera y el 28 en relación a Pedro Sánchez. Se puede destacar que Santiago Abascal es el representante con menor actividad a excepción del día 28 que es superado por Pablo Iglesias.

Representante	22	23	28
Pablo Iglesias	16.493	25.678	8.012
Pedro Sánchez	26.014	28.945	18.457
Albert Rivera	25.012	38.379	10.163
Pablo Casado	15.903	18.637	10.746
Santiago Abascal	14.086	9.824	10.654

Cuadro 3.1: Número de tweets de las menciones por cuenta y día. Política electoral.

A continuación se ha creado una tabla en la que se puede observar el número de tweets por representante y dispositivo utilizado. Para ello, es necesario hacer lo siguiente.

Representante	Android	iPhone	iPad	Web	Otros
Albert Rivera	40.584	18.929	1.638	11.608	795
Pablo Casado	24.128	12.059	1.063	7.194	842
Pablo Iglesias	27.744	11.020	1.174	9.587	658
Pedro Sánchez	38.422	19.511	1.915	12.456	1.112
Santiago Abascal	19.809	7.682	607	6.281	185

Cuadro 3.2: Número de tweets de las menciones por cuenta y dispositivo utilizado. Política electoral

Se puede observar que el mayor número de tweets fue realizado a través de un dispositivo Android, seguido de iPhone, iPad y finalmente por la Web. Indistintamente del representante al que vayan dirigidos los tweets.

3.2. Análisis por cuenta de twitter

3.2.1. Días 22 y 23 de abril

3.2.1.1. Pablo Iglesias

Recordar que la cuenta seleccionada es la cuenta oficial de Pablo Iglesias, es decir, @ Pablo_Iglesias_. Realizamos el mismo procedimiento que en el caso de las cuentas de turismo para el manejo y la limpieza de la base de datos.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
sanchezcastejon	3.062	15,08	populares	1.038	5,11
pablocasado	2.499	12,31	equiparacionya	895	4,41
ahorapodemos	2.166	10,67	debate	808	3,98
albertrivera	1.937	9,54	constitución	585	2,88
eldebateenrtve	1.566	7,71	pablo	504	2,48
jusapol	1.460	7,19	bien	393	1,94
psoe	1.432	7,05	hoy	380	1,87
eldebatedecisivo	1.205	5,93	único	375	1,85

Cuadro 3.3: Número de repeticiones y porcentaje de uso de cada palabra 22 y 23 de abril. Pablo Iglesias

A continuación, se crea un gráfico de barras en el que se observen las palabras más utilizadas y su frecuencia.

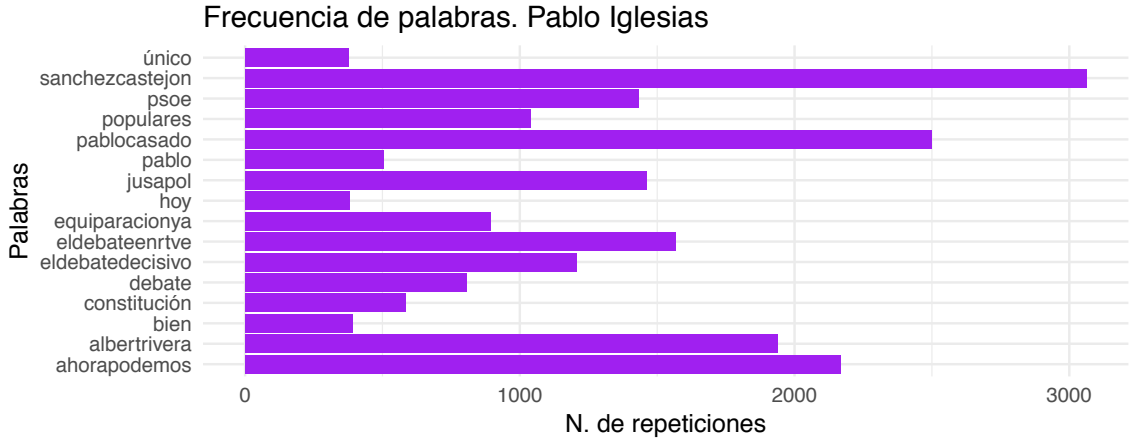


Figura 3.2: Frecuencia de palabras 22 y 23 abril. Pablo Iglesias.

En el siguiente gráfico se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.3: Nube de palabras 22 y 23 de abril. Pablo Iglesias.

Tanto en la tabla de frecuencias como en el gráfico de barras y la nube de palabras, se observa que la palabra más utilizada en relación a **@ pablo_iglesias_** los días 22 y 23 de abril es *sanchezcastejon* con 3.062 repeticiones, seguido de *pablocasado*, *ahorapodemos* y *albertrivera* con 2.499, 2.166 y 1.937 repeticiones, respectivamente. Estas palabras corresponden a las cuentas oficiales de los secretarios generales de los partidos más importantes y a la cuenta oficial del partido de Pablo Iglesias, esta alta frecuencia puede ser debida a que como se dijo anteriormente, los días 22 y 23 de abril se realizaron dos debates entre los secretarios señalados anteriormente. Además también tenemos palabras como *eldebateenrtve*, *eldebatedecisivo* y *debate* con 1.566, 1.205 y 808 repeticiones, respectivamente, que corroboran la deducción anterior. Las palabras menos utilizadas son *único*, *bien* y *hoy* con 411, 396 y 380 repeticiones, respectivamente.

Finalmente, se crea un dendrograma en el que se observe las posibles relaciones entre las palabras más utilizadas.

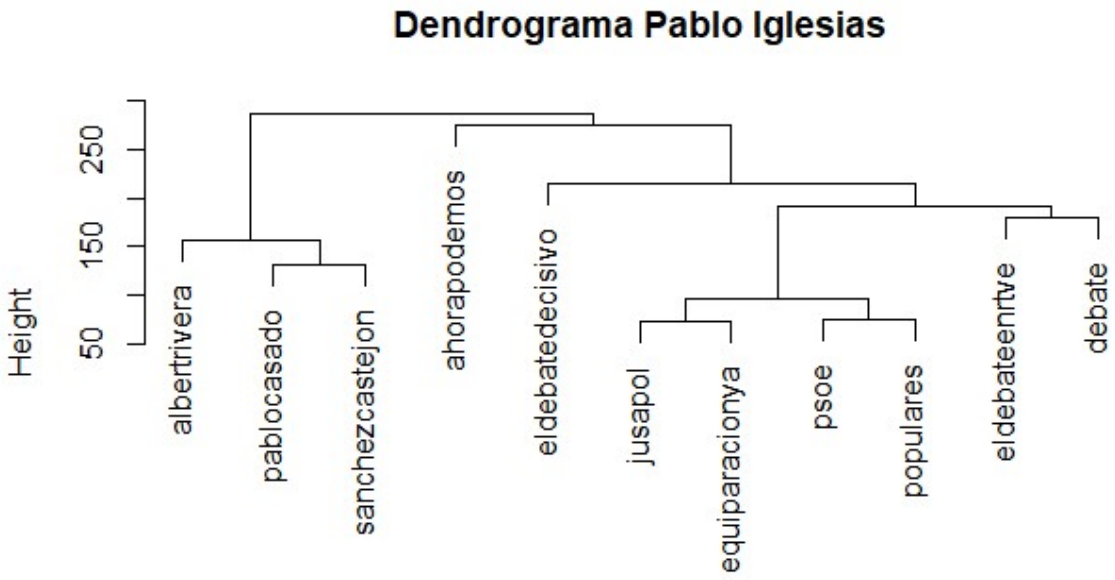


Figura 3.4: Dendrograma 22 y 23 abril. Pablo Iglesias.

Por un lado se puede observar que las palabras *sanchezcastejon*, *albertrivera* y *pablocasado* están relacionadas entre sí, esto es debido a que dichos candidatos junto con Pablo Iglesias fueron contrincantes en los debates realizados los días 22 y 23 de abril. Por otro lado, las otras ramas ponen de manifiesto el gran interés que dichos debates suscitaron, como partidos políticos aparecen expresamente mencionados *Ahora Podemos*, *PSOE* y *PP*. En esta cuenta en concreto, aparece el propio partido, es decir, *Ahora Podemos* y dos de los partidos contrincantes, que son *PP* y *PSOE*

3.2.1.2. Pedro Sánchez

Recordar que la cuenta seleccionada es la cuenta oficial de Pedro Sánchez, es decir, @sanchezcastejon. Realizamos el mismo procedimiento que en los casos anteriores para el manejo y la limpieza de la base de datos.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
psoe	4.114	21,86	eldebatedecisivo	860	4,57
pablocasado	2.480	13,18	equiparacionya	835	4,44
albertrivera	2.204	11,71	debate	716	3,80
pabloiglesias	2.031	10,79	ahorapodemos	449	2,39
eldebateenrtve	1.613	8,57	españa	429	2,28
jusapol	1.362	7,24	sanchez	426	2,26
populares	916	4,87	pedro	383	2,04

Cuadro 3.4: Número de repeticiones y porcentaje de uso de cada palabra 22 y 23 de abril. Pedro Sánchez

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

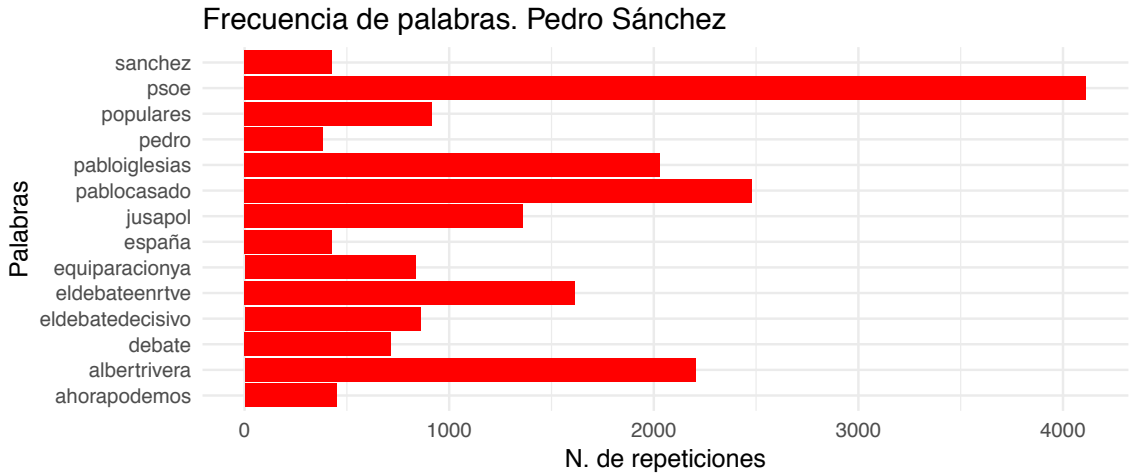


Figura 3.5: Frecuencia de palabras 22 y 23 de abril. Pedro Sánchez.

En el siguiente gráfico se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.6: Nube de palabras 22 y 23 abril. Pedro Sánchez.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 15000 veces. En la tabla se observa que la palabra más utilizada es *psoe* con 4.114 repeticiones, debido a que Pedro Sánchez es el Secretario general del PSOE. Le sigue *pablocasado*, *albertrivera*, *pabloiglesias* y *eldebateenrtve* con 2.480, 2.204, 2.031 y 1.613 repeticiones, respectivamente. Esto es debido a que como ya se ha comentado anteriormente, Pablo Casado, Albert Rivera, Pablo Iglesias junto con Pedro Sánchez formaron parte del debate de TVE. Las palabras menos utilizadas son *españa*, *sanchez* y *pedro* con 429, 426 y 383 repeticiones, respectivamente. Esto es debido a que estamos análizando los tweets de Pedro Sánchez.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

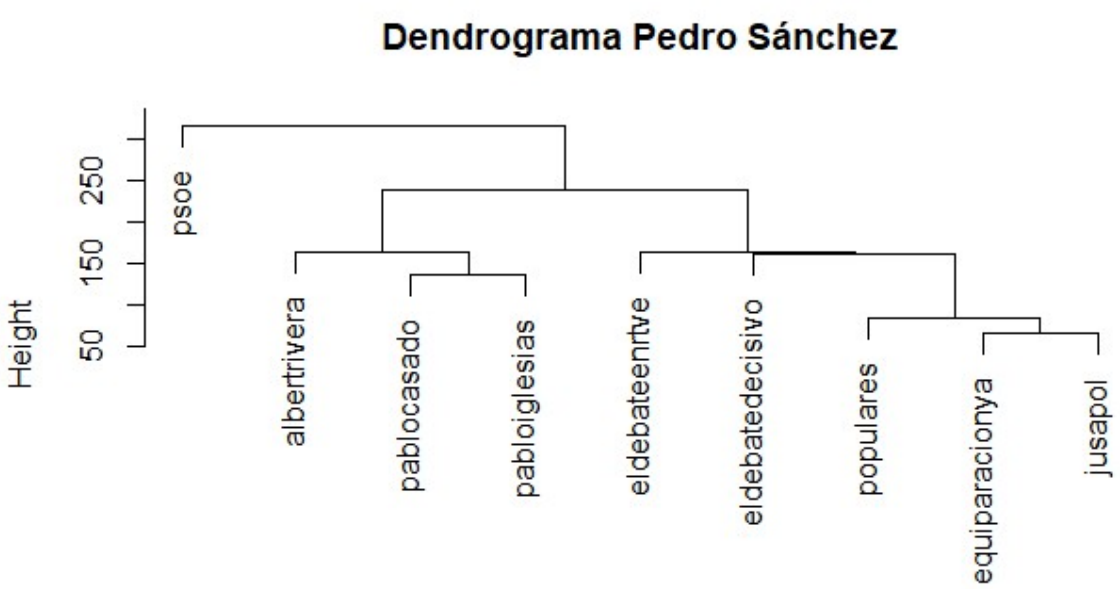


Figura 3.7: Dendrograma 22 y 23 abril. Pedro Sánchez.

En el gráfico anterior se puede observar que *albertrivera*, *pablocasado* y *pabloiglesias* forman un grupo entre sí, ya que como se ha dicho anteriormente, son los representantes de Ciudadanos, PP y Unidas Podemos, respectivamente, y fueron contrincantes en los debate de los días estudiados. Además, hay otro grupo formado por *psoe* que como ya se ha dicho anteriormente es la cuenta oficial del PSOE. Y por último, hay un grupo formado por *eldebateenrtve*, *eldebatedecisivo*, *ppopulares*, *equiparacionya* y *jusapol*, esto es debido a que en el debate se estuvo hablando sobre el motivos del hashtag #equiparacionya (creado por la cuenta @ jusapol, cuenta oficial de la justicia salarial de la policia) este es un movimiento creado para la equiparación salarial en la policía. Además, estos hicieron una concentración el día 23 de abril en la puerta del lugar donde se produjo el debate.

3.2.1.3. Albert Rivera

Recordar que la cuenta seleccionada es la cuenta oficial de Albert Rivera. Realizamos el mismo procedimiento que en los casos anteriores para el manejo y la limpieza de la base de datos.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
sanchezcastejon	2.563	14,91	eldebatedecisivo	1.257	7,31
pablocasado	2.458	14,30	psoe	1.049	6,10
ciudadanoscs	1.931	11,23	jusapol	1.025	5,96
pabloiglesias	1.573	9,15	populares	741	4,31
inesarrimadas	1.496	8,70	equiparacionya	636	3,70
eldebateenrtve	1.410	8,20	debate	633	3,68

Cuadro 3.5: Número de repeticiones y porcentaje de uso de cada palabra 22 y 23 de abril. Albert Rivera

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

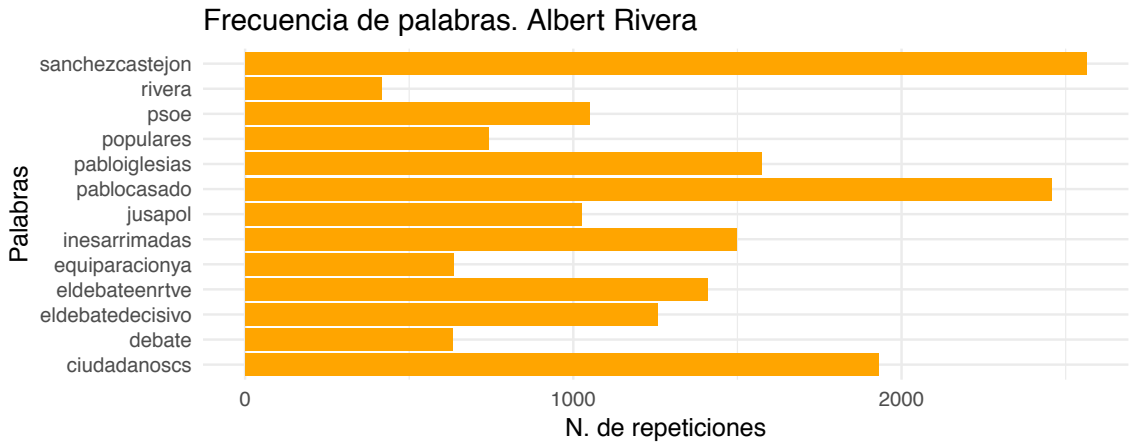


Figura 3.8: Frecuencia de palabras 22 y 23 abril. Albert Rivera.

En el siguiente gráfico se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

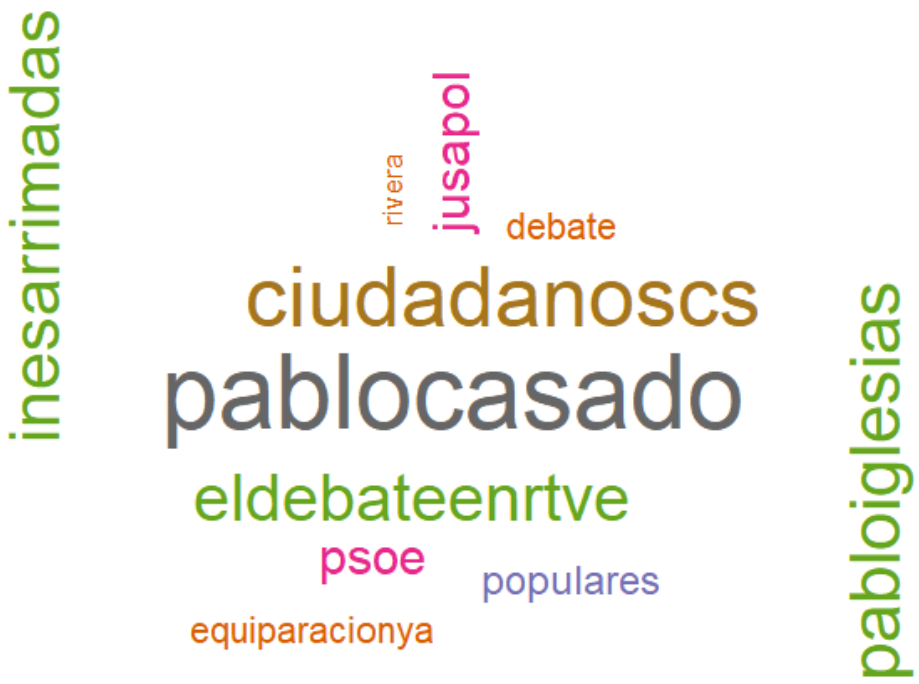


Figura 3.9: Nube de palabras 22 y 23 abril. Albert Rivera.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 15000 veces. En la tabla se observa que las palabras más utilizadas son *sanchezcastejon*, *pablocasado*, *ciudadanoscs*, *pabloiglesias* y *inesarrimadas* con 2.563, 2.458, 1.931, 1.573 y 1.496 repeticiones, respectivamente. Esto puede ser debido a que como ya se ha comentado anteriormente, Pedro Sánchez, Pablo Casado, Pablo Iglesias junto con Albert Rivera fueron contrincantes en el debate de RTVE y el de Antena 3. Además, están Ciudadanos e Inés Arrimadas que son los referentes a su propio partido. Las palabras menos utilizadas son *equiparacionya* y *debate* con 636 y 633 repeticiones, respectivamente. Esto es debido a lo ya comentado anteriormente. Al igual que en los casos anteriores hay palabras como *eldebateenrtve*, *eldebatedecisivo* y *jusapol* que se encuentran con frecuencia.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

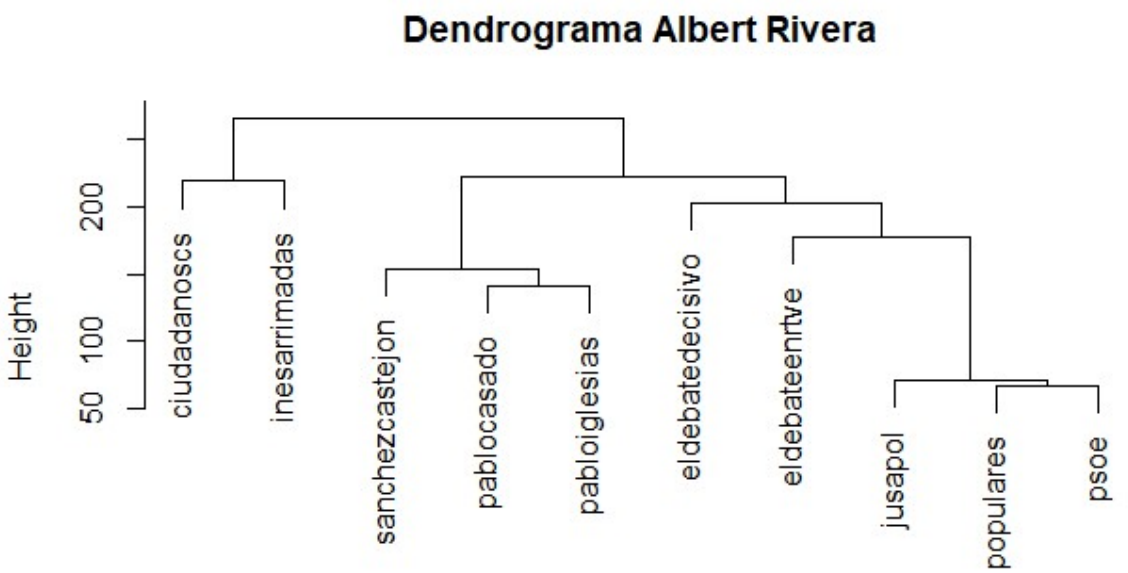


Figura 3.10: Dendrograma 22 y 23 abril. Albert Rivera.

En el gráfico anterior se puede observar que *sanchezcastejon*, *pablocasado* y *pabloiglesias* forman un grupo entre sí, ya que como se ha dicho anteriormente, son los representantes de PSOE, PP y Unidas Podemos, respectivamente, y fueron los contrincantes que formaron parte de los debate realizados los días 22 y 23 de abril. Además, hay otro grupo formado por *inesarrimadas* y *ciudadanoscs*, que es debido a que Inés Arrimadas es la Secretaria de Formación y Portavoz de Ciudadanos. Y por último, hay un grupo formado por *eldebateenrtve*, *eldebatedecisivo*, *eldebateenrtve*, *jusapol* y *psoe* esto es debido a que como ya se ha dicho anteriormente, en el debate se estuvo hablando sobre el movimiento creado para la equiparación salarial en la policía. Además, estos hicieron una concentración el día 23 de abril en la puerta del lugar donde se produjo el debate.

3.2.1.4. Pablo Casado

Recordar que la cuenta seleccionada es la cuenta oficial de Pablo Casado, es decir, @ pablocasado_ Realizamos el mismo procedimiento que en los casos anteriores para el manejo y la limpieza de la base de datos.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
populares	3.667	15,65	eldebatedecisivo	1.315	5,61
sanchezcastejon	3.654	15,60	equiparacionya	1.124	4,80
albertrivera	3.004	12,82	ahorapodemos	591	2,52
pabloiglesias	2.309	9,86	debate	579	2,47
eldebateenrtve	2.111	9,01	sánchez	551	2,35
psoe	1.816	7,75	españa	488	2,08
jusapol	1.799	7,68	casado	417	1,78

Cuadro 3.6: Número de repeticiones y porcentaje de uso de cada palabra 22 y 23 de abril. Pablo Casado

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

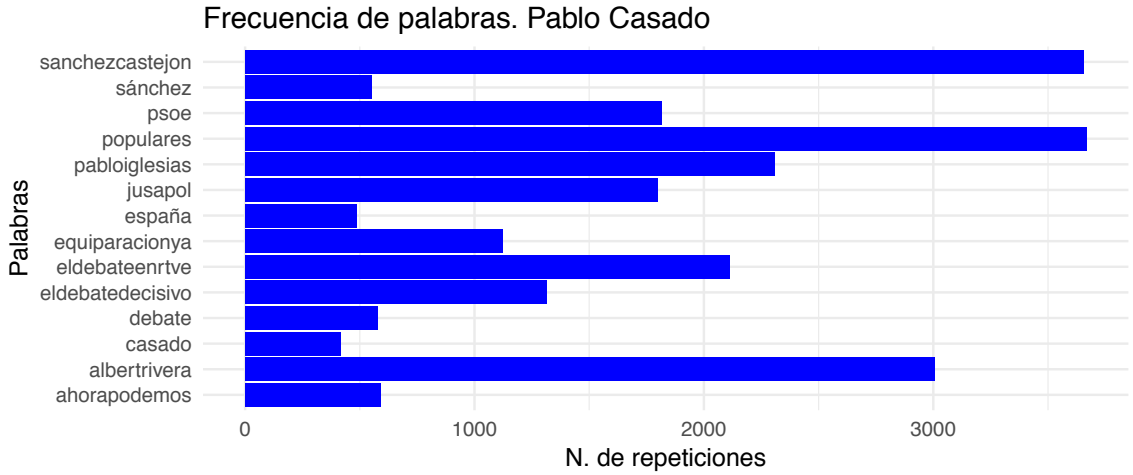


Figura 3.11: Frecuencia de palabras 22 y 23 abril. Pablo Casado.

En el siguiente gráfico se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.12: Nube de palabras 22 y 23 abril. Pablo Casado.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 15000 veces. En la tabla se observa que las palabras más utilizadas son *populares*, *sanchezcastejon*, *albertrivera* y *pabloiglesias* con 3.667, 3.655, 3.004 y 2.309 repeticiones, respectivamente. Esto puede ser debido a que como ya se ha comentado anteriormente, Pedro Sánchez, Albert Rivera, Pablo Iglesias junto con Pablo Casado formaron parte del debate de RTVE y el de Antena 3, además Pablo Casado es el representante del PP, cuyo usuario de twitter es @ populares. Las palabras menos utilizadas son *españa* y *casado* con 488 y 417 repeticiones, respectivamente.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

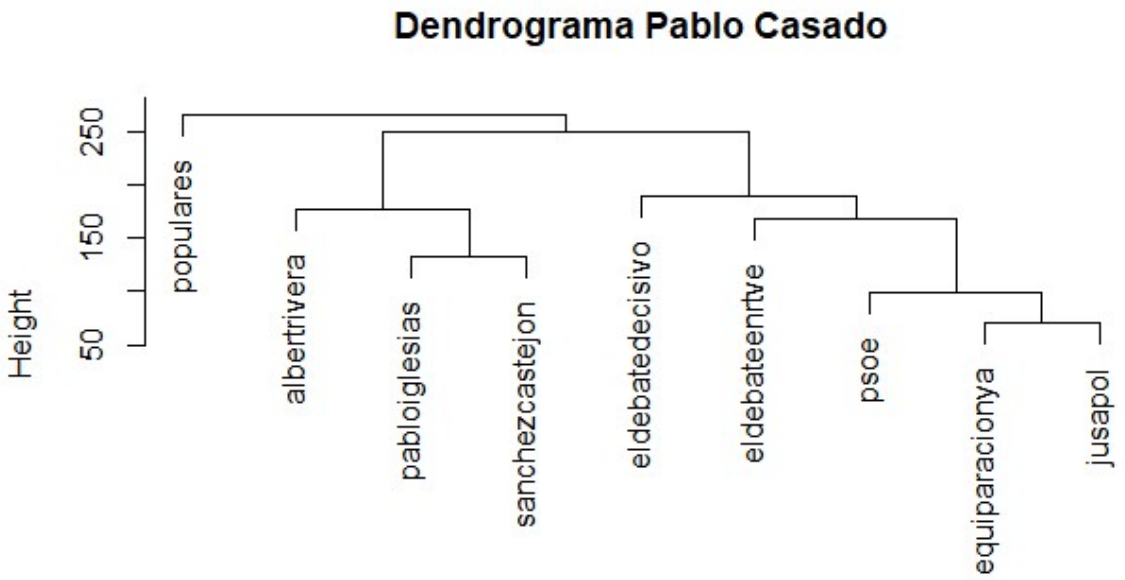


Figura 3.13: Dendrograma 22 y 23 abril. Pablo Casado.

En el gráfico anterior se puede observar que *albertrivera*, *pabloiglesias* y *sanchezcastejon* forman un grupo entre sí, ya que como se ha dicho anteriormente, son los representantes de Ciudadanos, Unidas Podemos y PSOE, respectivamente, y fueron los que formaron parte de los debate realizados los días 22 y 23 de abril. Además, hay otro grupo formado por *populares* que como se ha dicho anteriormente es la cuenta oficial del PP. Y por último, hay un grupo formado por *eldebatedecisivo*, *eldebateenrtve*, *psoe*, *equiparacionya* y *jusapol* esto es debido a que se estuvo hablando sobre el motivos del hashtag #equiparacionya, como se ha dicho anteriormente, este es un movimiento creado para la equiparación salarial en la policía. Además, estos hicieron una concentración el día 23 de abril en la puerta del lugar donde se produjo el debate.

3.2.1.5. Santiago Abascal

Recordar que la cuenta seleccionada es la cuenta oficial de Santiago Abascal, es decir, @ SANTI_Abascal. Realizamos el mismo procedimiento que en los casos anteriores para el manejo y la limpieza de la base de datos.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las

demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
voxes	4.402	16,28	ortegasmith	828	3,06
jusapol	2.584	9,56	ivanedlm	773	2,86
sanchezcastejon	2.439	9,02	espejopublico	693	2,56
psoe	2.067	7,64	monasterior	686	2,54
pablocasado	2.045	7,56	ahorapodemos	639	2,36
populares	1.646	6,09	albertrivera	604	2,23
eldebateenrtve	1.533	5,67	debate	561	2,07
equiparacionya	1.499	5,54	españa	504	1,86
pabloiglesias	1.304	4,82	susannagriso	498	1,84
vox	885	3,27	elmundoes	467	1,73

Cuadro 3.7: Número de repeticiones y porcentaje de uso de cada palabra 22 y 23 de abril. Santiago Abascal

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

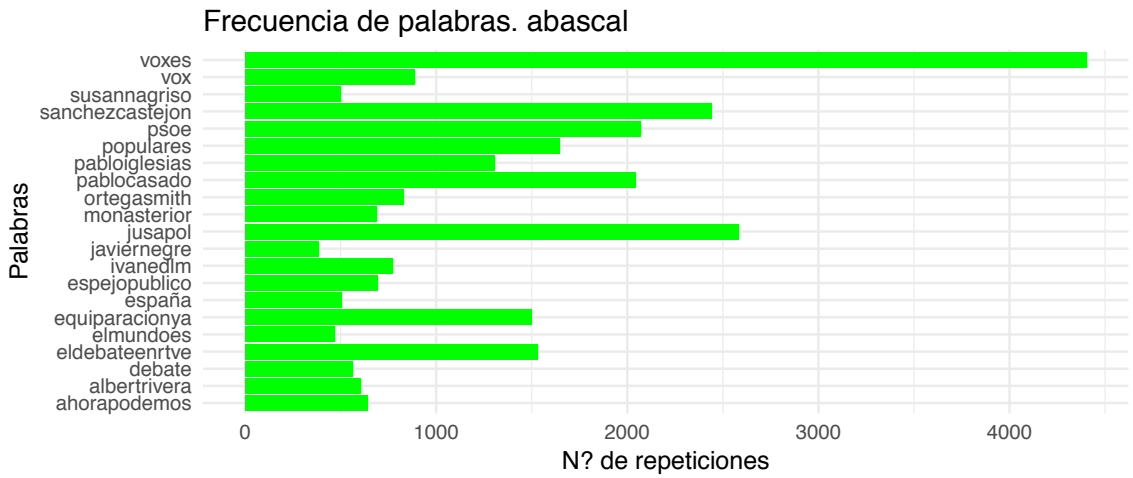


Figura 3.14: Frecuencia de palabras 22 y 23 abril. Santiago Abascal.

En el siguiente gráfico se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.15: Nube de palabras 22 y 23 abril. Santiago Abascal.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 15000 veces. En la tabla se observa que la palabra más utilizada es *voxes* con 4.402 repeticiones, debido a que Santiago Abascal es el Secretario general de VOX y @ voxes es la cuenta oficial de dicho partido. Le sigue *jusapol*, *sanchezcastejon* y *psoe* con 2.584, 2.439 y 2.067 repeticiones, respectivamente. También hay palabras como *equiparacionya* con 1.501 repeticiones que es un hashtag creado por la cuenta @ jusapol que busca la equiparación salarial en la policía. Las palabras menos utilizadas son *debate*, *españa*, *susannagriso* y *elmundoes* con 561, 504, 498 y 467 repeticiones, respectivamente. Por un lado, se conoce que @ Susannagriso es la cuenta oficial de la periodista Susanna Griso que tiene una sección en el programa "Espejo Público" llamada "Un café con Susana" en la que el día 22 de abril estuvo de invitado Santiago Abascal. Por otro lado, está *elmundoes* que es la cuenta oficial del periódico El Mundo que habló mucho durante estos días de los debates y de la ausencia de Santiago Abascal en dichos debates.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

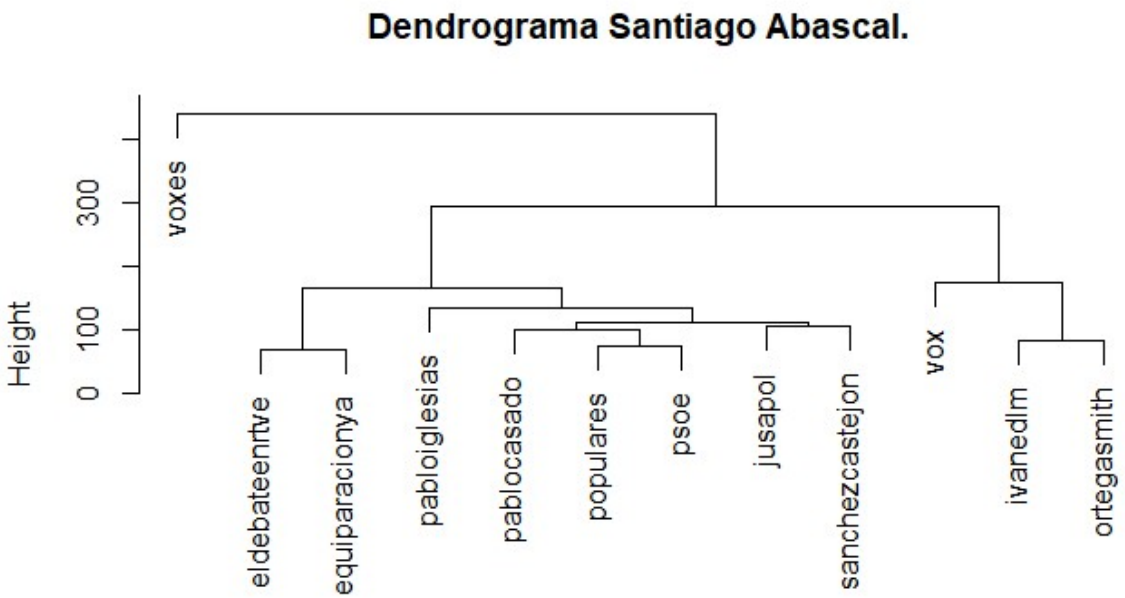


Figura 3.16: Dendrograma 22 y 23 abril. Santiago Abascal.

En el gráfico anterior se puede observar que *vox*, *ivanedlm* y *ortegasmith* forman un grupo entre sí, ya que Ivan Espinosa y Javier Ortega Smith son diputados de Vox. Además, hay otro grupo formado por *voxes* que como ya se ha dicho anteriormente es la cuenta oficial de Vox. Y por último, hay un grupo formado por *eldebateenrtve*, *equiparacionya*, *pabloiglesias*, *pablocasado*, *ppopulares*, *psoe*, *jusapol* y *sanchezcastejon*, esto es debido a que en el debate se estuvo hablando sobre el hashtag #equiparacionya, como se ha dicho anteriormente, este es un movimiento creado para la equiparación salarial en la policía, pero se acabó extrapolando a la ausencia de Vox en el debate, es un hashtag con variedad de significados. También cabe destacar que a diferencia de los demás representantes, Santiago Abascal no participó en los debates, por lo que se ve una gran diferencia en el dendrograma respecto a los realizados anteriormente para los otros representantes.

3.2.2. Día 28 de abril (Elecciones generales en España)

3.2.2.1. Pablo Iglesias

Realizamos el mismo procedimiento pero en este caso para los del día 28 de abril, es decir, para los tweets realizados el día de las elecciones generales.

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
ahorapodemos	1.652	27,49	albertrivera	246	4,09
sanchezcastejon	786	13,08	pnique	230	3,83
psoe	552	9,19	hxbimin	228	3,79
pablo	365	6,07	sostaxis	227	3,78
podemos	300	4,99	colegio	226	3,76
españa	289	4,81	pablocasado	217	3,61
puede	285	4,74	ahora	203	3,38

Cuadro 3.8: Número de repeticiones y porcentaje de uso de cada palabra 28 de abril. Pablo Iglesias

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

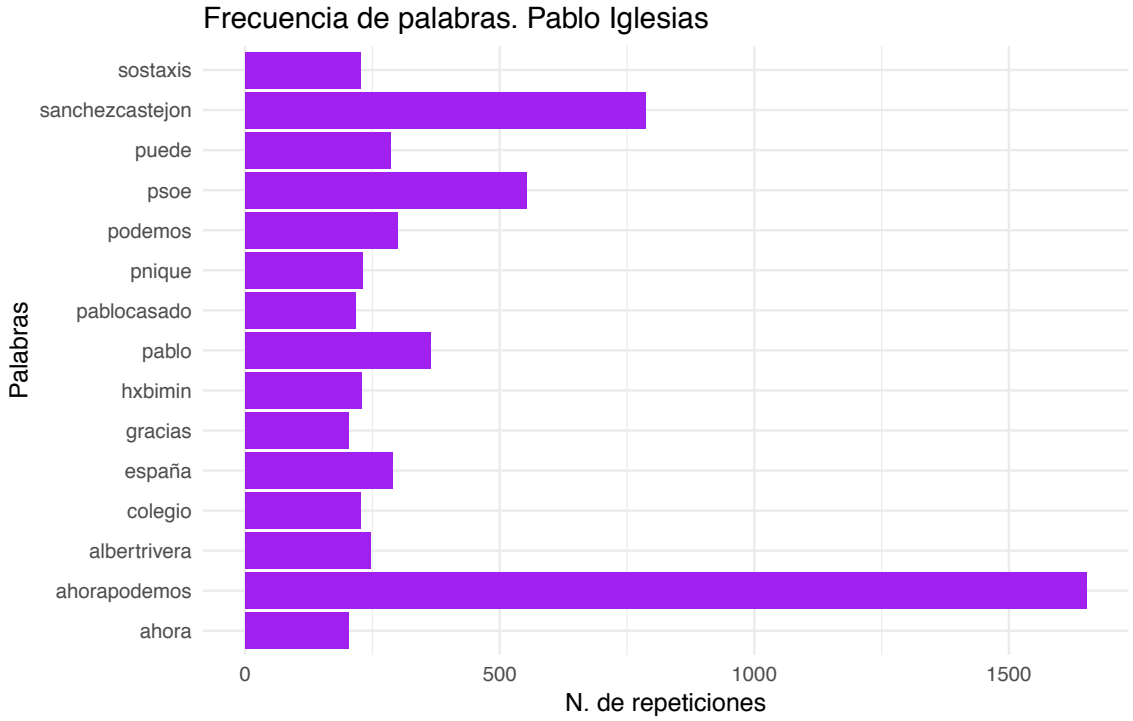


Figura 3.17: Frecuencia de palabras 28 de abril. Pablo Iglesias.

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.18: Nube de palabras 28 de abril. Pablo Iglesias.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 375 veces. En la tabla se observa que las palabras más utilizadas son *ahorapodemos*, *sanchezcastejon* y *psoe*, con 1.652, 786 y 552 repeticiones, respectivamente. Esto puede ser debido a que según los pactómetros y la mayor parte de los medios de comunicación, PSOE y Unidas Podemos podrían formar gobierno, tema del que se habló mucho el día de las elecciones. Las palabras menos utilizadas son *colegio*, *pablocasado* y *ahora* con 226, 217 y 203 repeticiones, respectivamente. Además, se tienen palabras como *pnique* con 230 repeticiones que es el usuario de Pablo Echenique, el Secretario de Organización de Unidas Podemos, y también se tiene *hxbimin* que es el usuario de una chica que causó polémica en twitter con un tweet que publicó refiriéndose a que toda su familia era fiel a Unidas Podemos y a Pablo Iglesias.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

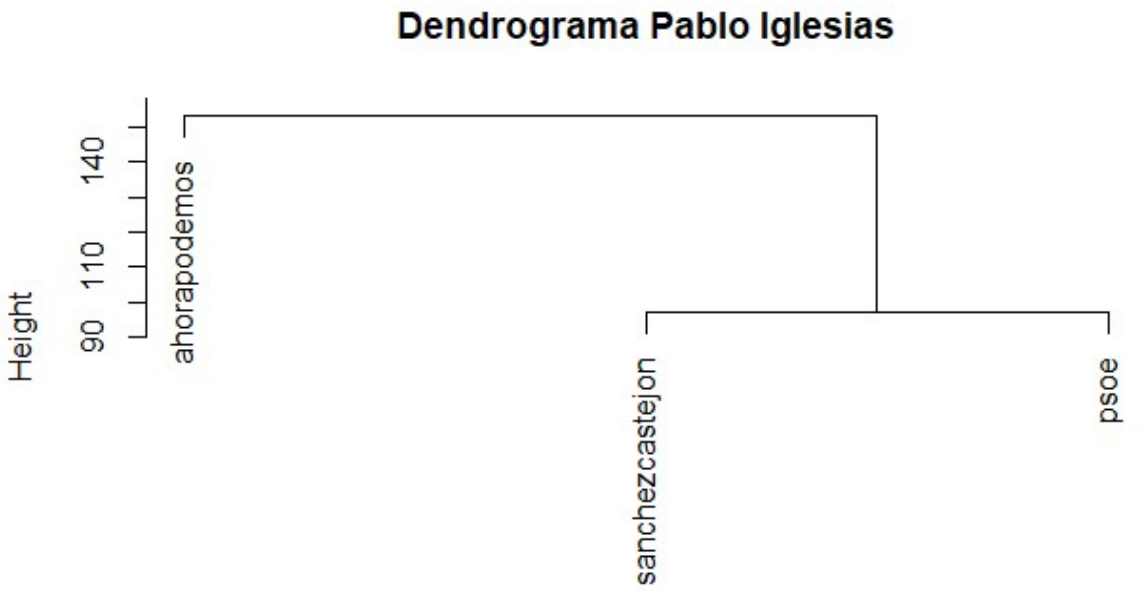


Figura 3.19: Dendrograma 28 de abril. Pablo Iglesias.

Se puede observar que por un lado tenemos la palabra *ahorapodemos*. Por otro lado hay un grupo formado por las palabras *sanchezcastejon* y *psoe*, esto es debido a que como ya se ha comentado anteriormente, @ sanchezcastejon es el usuario de Pedro Sánchez que es el Secretario General del PSOE.

3.2.2.2. Pedro Sánchez

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
psoe	5.342	34,97	presidente	568	3,72
rivera	1.817	11,90	enhorabuena	468	3,06
españa	872	5,71	claro	448	2,93
albertrivera	797	5,22	ciudadanoscs	442	2,89
conriverano	661	4,33	voxes	434	2,84
ahora	636	4,16	espero	408	2,67
pedro	628	4,11	sanchez	395	2,59
pabloiglesias	584	3,82	gobierno	390	2,55

Cuadro 3.9: Número de repeticiones y porcentaje de uso de cada palabra 28 de abril. Pedro Sánchez

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

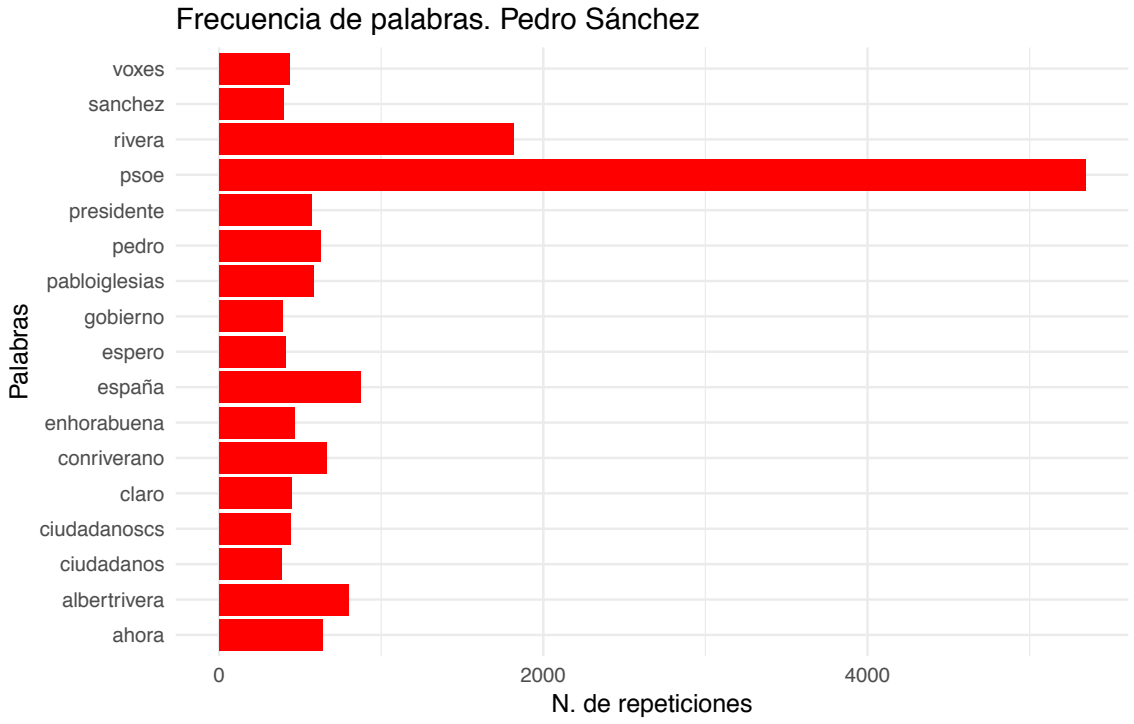


Figura 3.20: Frecuencia de palabras 28 de abril. Pedro Sánchez.

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.21: Nube de palabras 28 de abril. Pedro Sánchez.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 375 veces. En la tabla se observa que la palabra más utilizada es *psoe* con 5.342 repeticiones, ya que como se ha comentado anteriormente es la cuenta oficial del PSOE. Seguida de *rivera* y *españa*, con 1.817 y 872 repeticiones, respectivamente. Las palabras menos utilizadas son *sanchez* y *gobierno* con 395 y 390 repeticiones, respectivamente. Esto es debido a que Pedro Sánchez consiguió un número elevado de escaños y gracias a eso parece que puede volver a ser presidente del gobierno, siempre y cuando consiga hacer pactos. Además, se tienen palabras como *conriverano* con 661 repeticiones que es lo que decía el público en Ferraz cuando ya se supo el número de escaños que había obtenido el PSOE.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

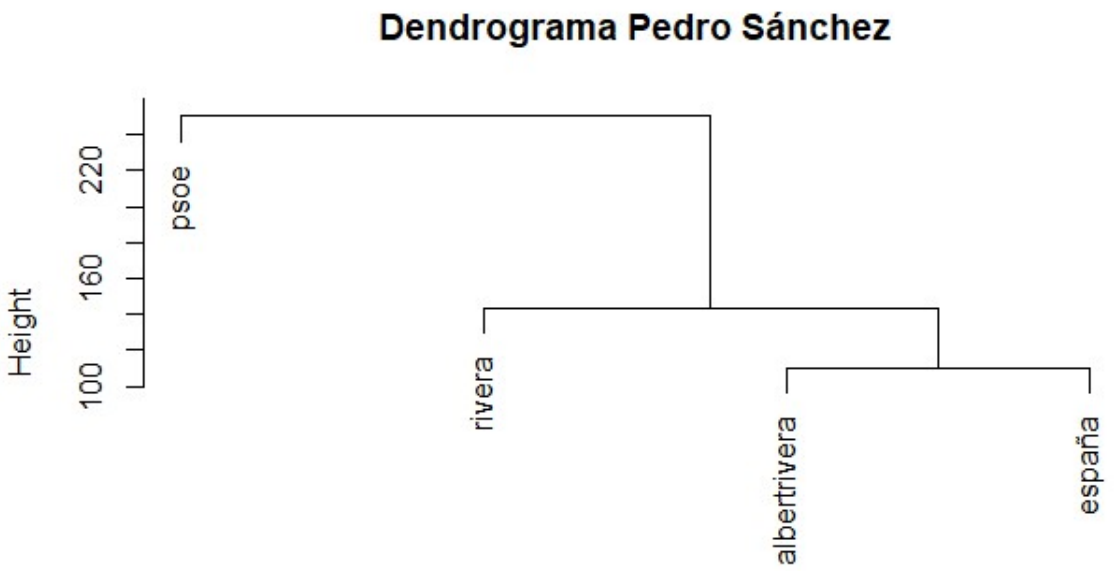


Figura 3.22: Dendrograma 28 de abril. Pedro Sánchez.

Se puede observar que por un lado tenemos la palabra *psoe*. Por otro lado hay un grupo formado por las palabras *rivera*, *albertrivera* y *españa* esto puede ser debido a que como ya se ha comentado anteriormente el día de las elecciones en la sede del PSOE en Ferraz se gritó al unísono "con rivera no" por lo que las personas que estaban allí no querían coalición de PSOE con Ciudadanos.

3.2.2.3. Albert Rivera

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
ciudadanoscs	1.533	23,47	mano	397	6,08
sanchezcastejon	1.003	15,36	españa	387	5,93
psoe	946	14,48	rivera	345	5,28
pablocasado	597	9,14	ahora	283	4,33
inesarrimadas	493	7,55	ser	279	4,27

Cuadro 3.10: Número de repeticiones y porcentaje de uso de cada palabra 28 de abril. Albert Rivera

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

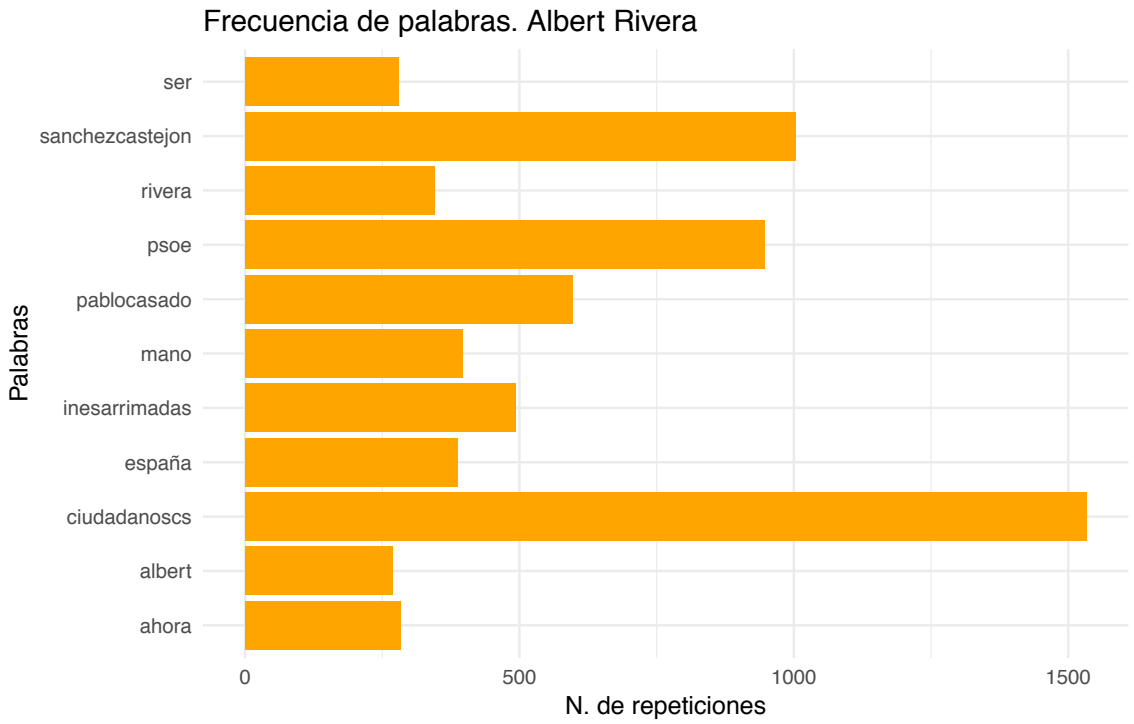


Figura 3.23: Frecuencia de palabras 28 de abril. Albert Rivera.

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.24: Nube de palabras 28 de abril. Albert Rivera.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 254 veces. En la tabla se observa que la palabra más utilizada es *ciudadanoscs* con 1.533 repeticiones, ya que como se ha comentado anteriormente es la cuenta oficial de Ciudadanos. Seguida de *sanchezcastejon* y *psoe*, con 1.003 y 946 repeticiones, respectivamente. Las palabras menos utilizadas son *albert* y *sanchez* con 269 y 263 repeticiones, respectivamente. Además, se tienen palabras como

inesarrimadas con 493 repeticiones al igual que ocurrió en los días 22 y 23 de abril, ya que Inés Arrimadas es un gran pilar del partido.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

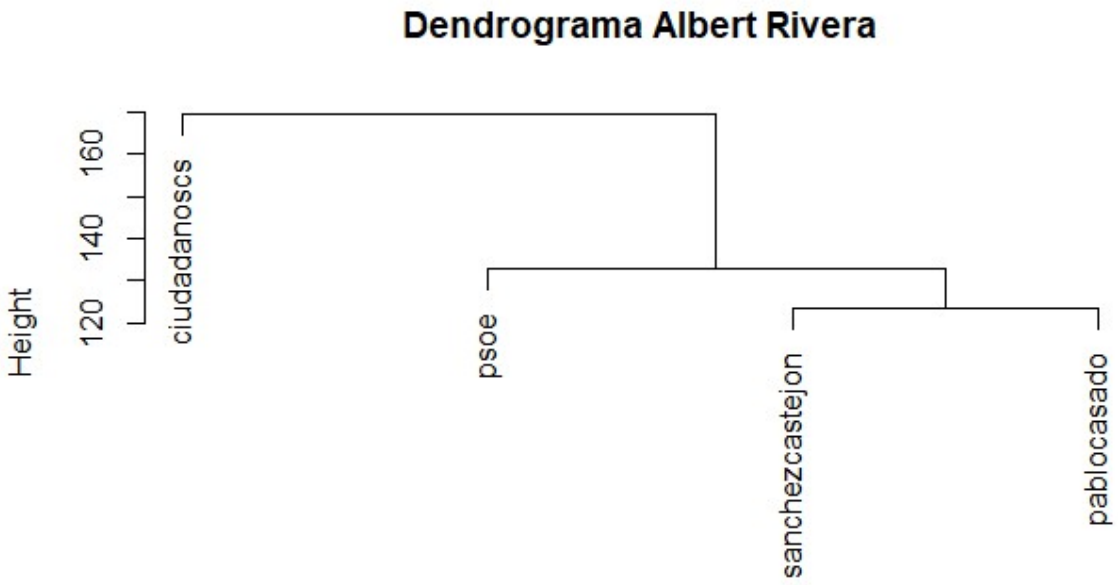


Figura 3.25: Dendrograma 28 de abril. Albert Rivera.

Se puede observar que por un lado tenemos la palabra *ciudadanoscs*. Por otro lado hay un grupo formado por las palabras *psoe*, *sanchezcastejon* y *pablocasado*.

3.2.2.4. Pablo Casado

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
populares	2.816	39,79	sanchezcastejon	378	5,34
albertrivera	556	7,86	teogarciaegea	371	5,24
cayetanaat	459	6,49	eleccionesgeneralesa	340	4,80
pablo	447	6,32	hoy	312	4,41
ahora	424	5,99	eleccionesl	296	4,18
españa	388	5,48	casado	290	4,10

Cuadro 3.11: Número de repeticiones y porcentaje de uso de cada palabra 28 de abril. Pablo Casado

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

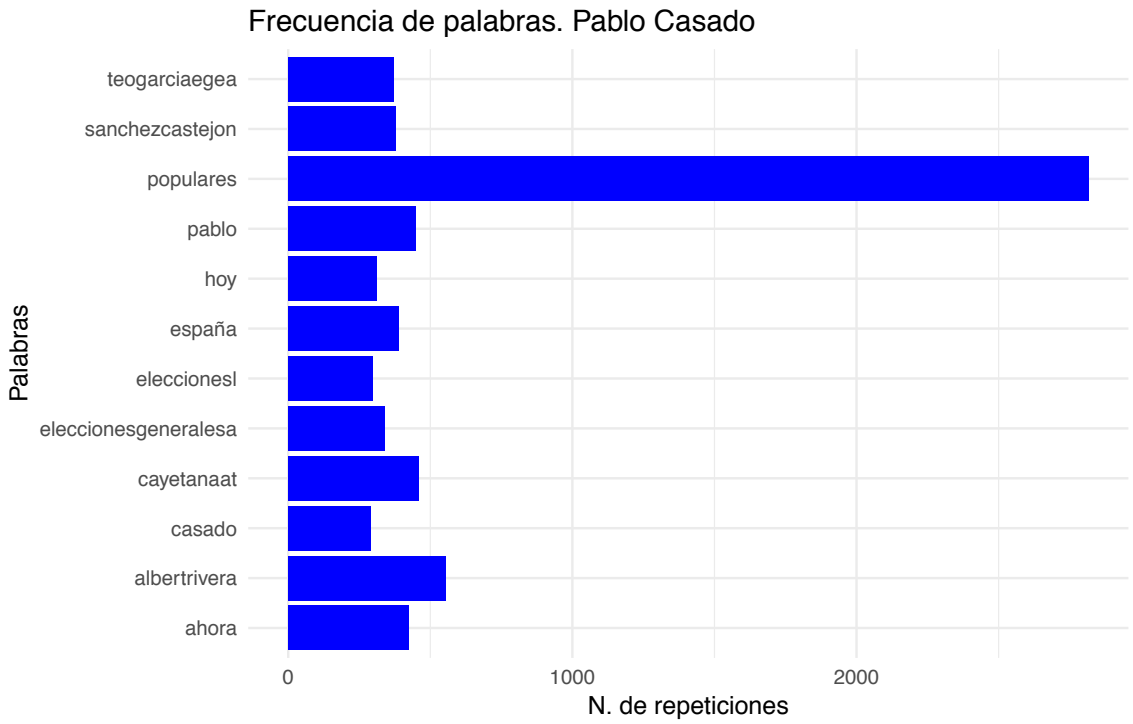


Figura 3.26: Frecuencia de palabras 28 de abril. Pablo Casado.

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.27: Nube de palabras 28 de abril. Pablo Casado.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 254 veces. En la tabla se observa que la palabra más utilizada es *populares* con 2.816 repeticiones, ya que como se ha comentado anteriormente es la cuenta oficial del Partido Popular. Seguida de *albertrivera* y *cayetanat*, con 556 y 459 repeticiones, respectivamente. Esto puede ser debido a que Cayetana Alvarez de Toledo es diputada del Partido Popular por Barcelona. Las palabras menos utilizadas son *eleccionesl* y *casado* con 269 y 263 repeticiones, respectivamente. Además, se tienen palabras como

teogarciaegea con 371 repeticiones, ya que es el nombre de usuario de Teodoro García Egea el Secretario General del Partido Popular.

En el caso de Pablo Casado para el día de las elecciones generales no se puede realizar dendrograma, ya que hay poca información y no existe relación entre las palabras.

3.2.2.5. Santiago Abascal

A continuación, se comienza con una tabla de frecuencias en la que se representan las palabras más utilizadas, el número de veces utilizada y el porcentaje de uso respecto a las demás palabras.

Palabra	Repeticiones	%	Palabra	Repeticiones	%
voxes	5.168	39,78	vox	744	5,73
ortegasmith	1.154	8,88	antoniomaestre	537	4,13
ivanedlm	1.147	8,83	gracias	410	3,16
monasterior	1.086	8,36	hoy	308	2,37
voxnoticiases	799	6,15	pablocasado	303	2,33
españa	782	6,02	psoe	284	2,19

Cuadro 3.12: Número de repeticiones y porcentaje de uso de cada palabra 28 de abril. Santiago Abascal

A continuación, se puede ver un gráfico en el que se representan las palabras más utilizadas y su respectiva frecuencia.

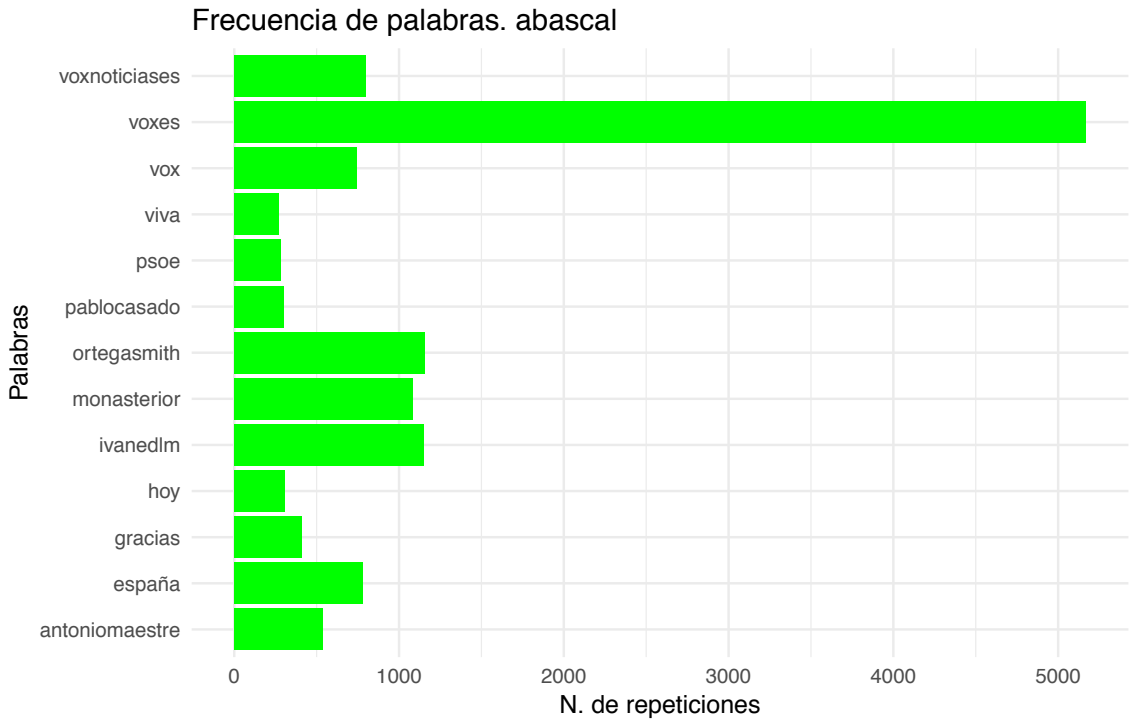


Figura 3.28: Frecuencia de palabras 28 de abril. Santiago Abascal.

En el siguiente gráfico, se representa una nube de palabras en el que las palabras de mayor tamaño son las que se repiten un mayor número de veces.

Figura 3.29: Nube de palabras 28 de abril. Santiago Abascal.

Como ya se ha comentado anteriormente, para hacer el análisis se han seleccionado las palabras que se repiten más de 266 veces. En la tabla se observa que la palabra más utilizada es *voxes* con 5168 repeticiones, ya que como se ha comentado anteriormente es la cuenta oficial de Vox. Seguida de *ortegasmith*, *ivanedlm* y *monasterior* con 1.154, 1.147 y 1.086 repeticiones, respectivamente. Esto es debido a que tanto como Javier Ortega Smith, como Ivan Espinosa y Rocío Monasterio pertenecen a Vox. Las palabras menos utilizadas son *pablocasado* y *psoe* con 303 y 284 repeticiones, respectivamente.

A continuación, se tiene un dendrograma en el que se establecen las relaciones entre las palabras formando diferentes grupos.

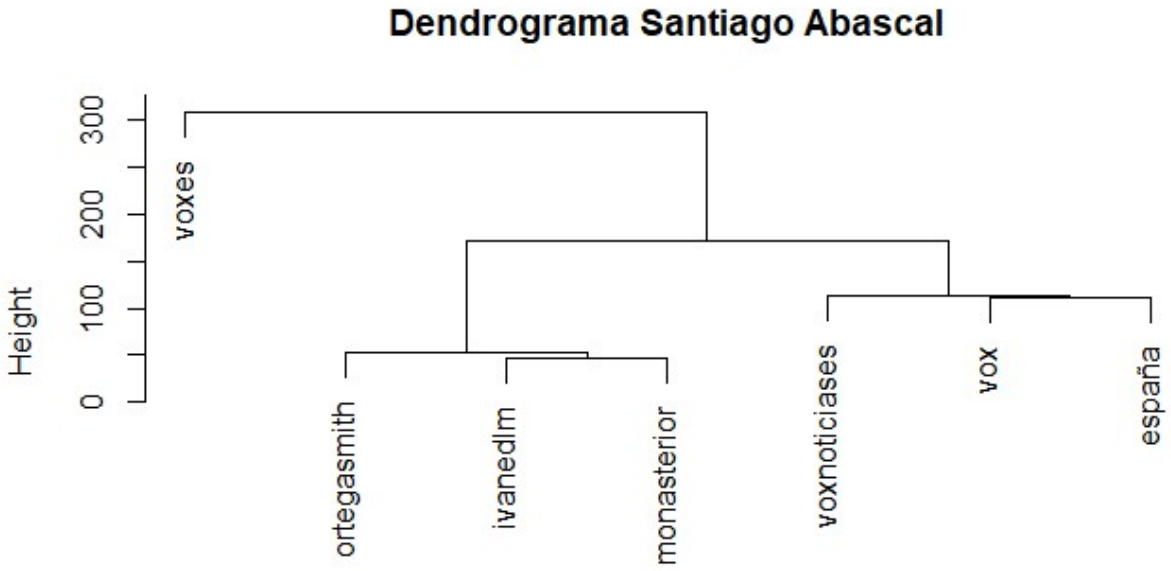


Figura 3.30: Dendrograma 28 de abril. Santiago Abascal.

Se puede observar que por un lado tenemos la palabra *voxes*. Por otro lado hay un grupo formado por las palabras *ortegasmith*, *ivanedlm* y *monasterior* por lo ya explicado anteriormente, y por último un grupo formado por *voxnoticiases*, *vox* y *españa*.

3.3. Análisis de Sentimientos

Para realizar el análisis de sentimientos se han catalogado cada una de las palabras en una escala de **-5** a **5** siendo **-5 muy negativa** y **5 muy positiva**. El siguiente conjunto de gráficas representa las palabras positivas y negativas realizadas en las menciones de cada cuenta y el número de tweets en el que aparece.

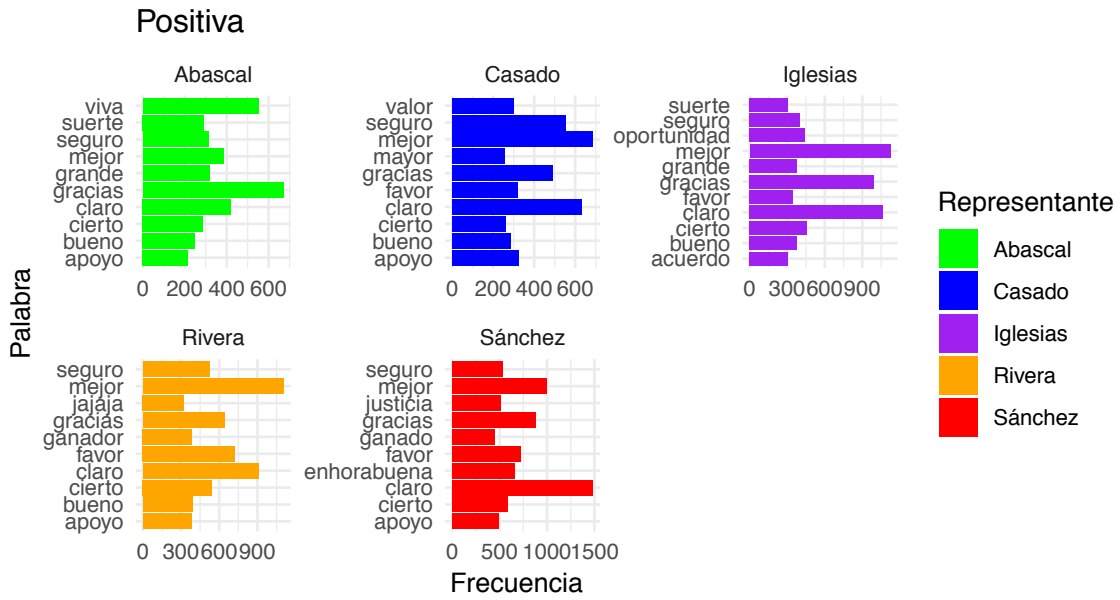


Figura 3.31: Frecuencia de palabras positivas por menciones de cuenta. Representantes.

Si se atiende a las palabras positivas, se observa que palabras como *gracias*, *claro*, *seguro* y *mejor* se encuentran relacionadas con las cuentas de los cinco representantes. Además, se pueden destacar palabras como *oportunidad* relacionada con Pablo Iglesias, *enhorabuena* relacionada con Pedro Sánchez, *ganador* relacionada con Albert Rivera, *valor* con Pablo Casado y *viva* con Santiago Abascal.

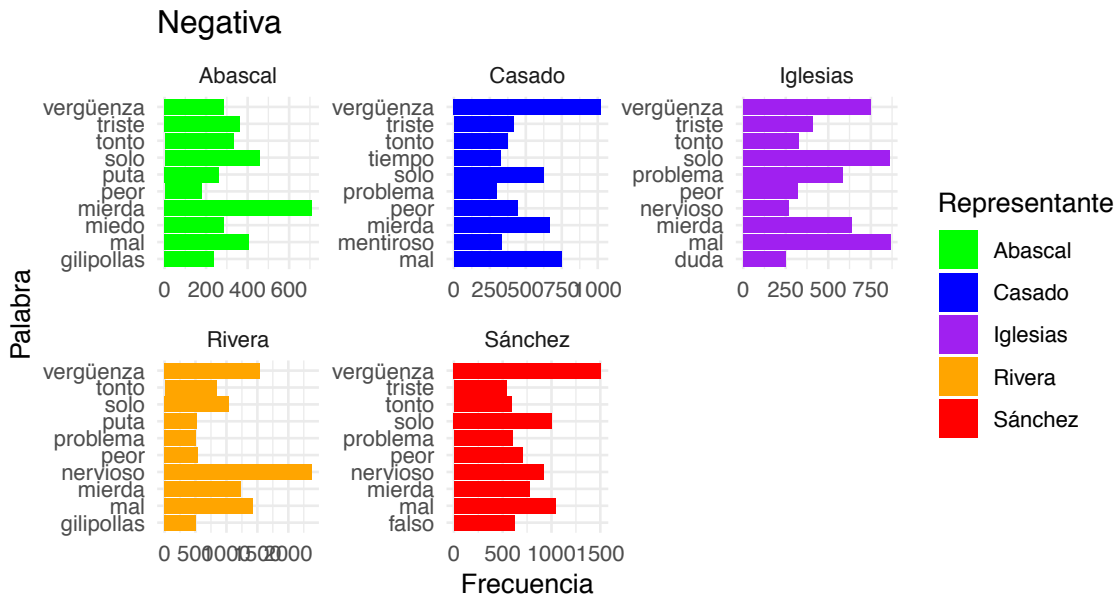


Figura 3.32: Frecuencia de palabras negativas por menciones de cuenta. Representantes.

Si se atiende a las palabras con connotación negativa, se observa que la mayor parte de las palabras relacionadas con los cinco representantes son insultos, como por ejemplo, *tonto* y *gilipollas*. Además, hay otras palabras que se encuentran relacionados con todos, como son *vergüenza*, *mal*, *peor* y *mierda*. También se puede destacar palabras como *miedo* relacionado con Santiago Abascal, *falso* relacionada con Pedro Sánchez y *mentiroso* con Albert Rivera.

Como se ha podido ver en las anteriores gráficas, en las menciones a las cuentas de Albert Rivera y Pablo Casado, se utilizan mayor número de palabras negativas. Sin embargo, en relación con Pedro Sánchez y Santiago Abascal son similares el número de palabras con connotación positiva y negativa. Pero en relación con Pablo Iglesias hay mayor número de palabras positivas que negativas.

A continuación, para poder realizar el siguiente gráfico, se tiene que agrupar las palabras por días, ya que se quiere ver la media de positividad/negatividad de cada uno de los días estudiados.

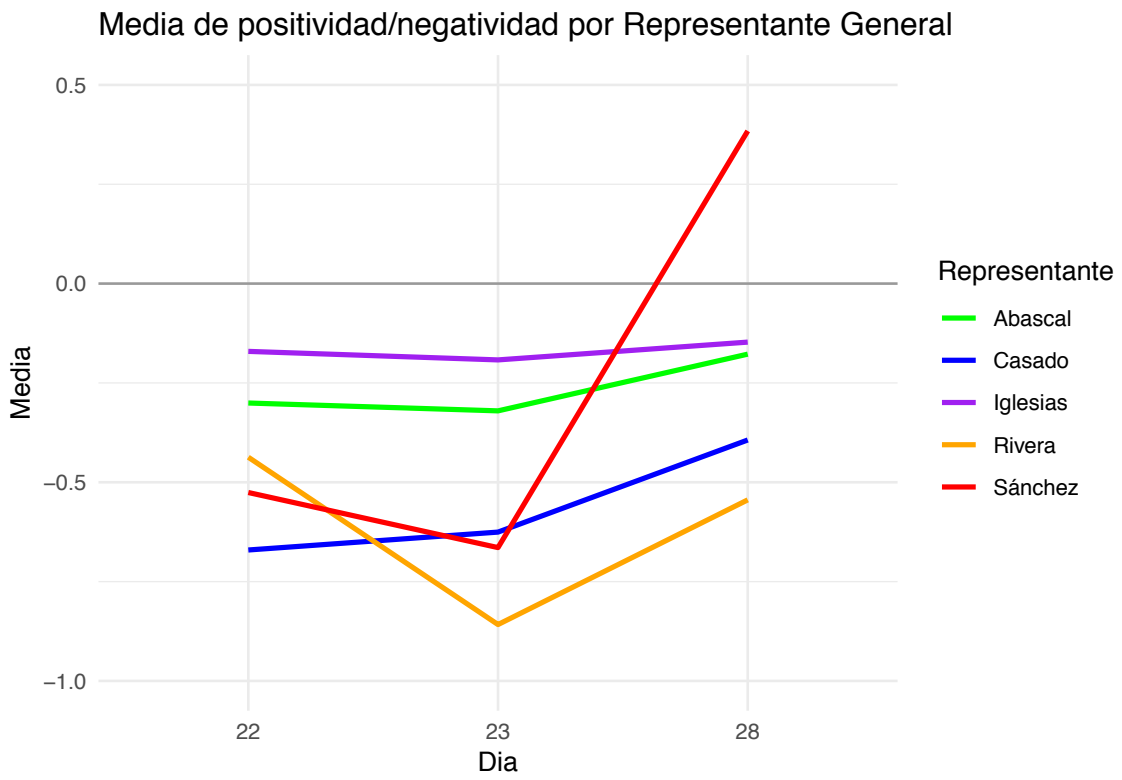


Figura 3.33: Media de positividad/negatividad de los tweets por día del mes y cuenta

Se puede observar que todos se mantienen en media negativa, es decir, la media de sus tweets es de carácter negativo. A excepción de Pedro Sánchez el día 28 de abril que su media aumenta hasta ser de carácter positivo. Esto puede ser debido a que fue el que mayor número de votos tuvo en las elecciones generales. Por un lado, se tiene a Pablo Iglesias y Santiago Abascal que su media está cercana a cero, es decir, hay tanto tweets positivos como negativos. Por otro lado, se observa que Albert Rivera disminuye mucho el día 23 ya que según distintos medios de comunicación y distintos analistas del grupo de expertos de Agenda Pública creen que Albert Rivera fue el perdedor en este debate de Antena 3. A continuación, citamos textualmente sus comentarios: "Ocho analistas, cinco mujeres y tres hombres, del grupo de expertos de Agenda Pública analizan en EL PAÍS a quién consideran ganador y a quién perdedor del debate electoral que se ha celebrado este martes en Atresmedia entre los cabezas de lista del PP, PSOE, Unidas Podemos y Ciudadanos para las elecciones del 28-A. Los expertos consideran por unanimidad que ha ganado Pablo Iglesias y una mayoría cree que ha perdido Albert Rivera. Este mismo análisis se realizó en el debate de TVE, consulte aquí las conclusiones a las que llegaron en este caso." (Agenda Pública, El País, 24 de abril de 2019 "Ganadores y Perdedores del último debate" Disponible en https://elpais.com/politica/2019/04/23/actualidad/1556023173_939577.html).

A continuación, se ha realizado un gráfico en el que se observa el porcentaje de tweets negativos y positivos en relación a cada una de las cuentas estudiadas.

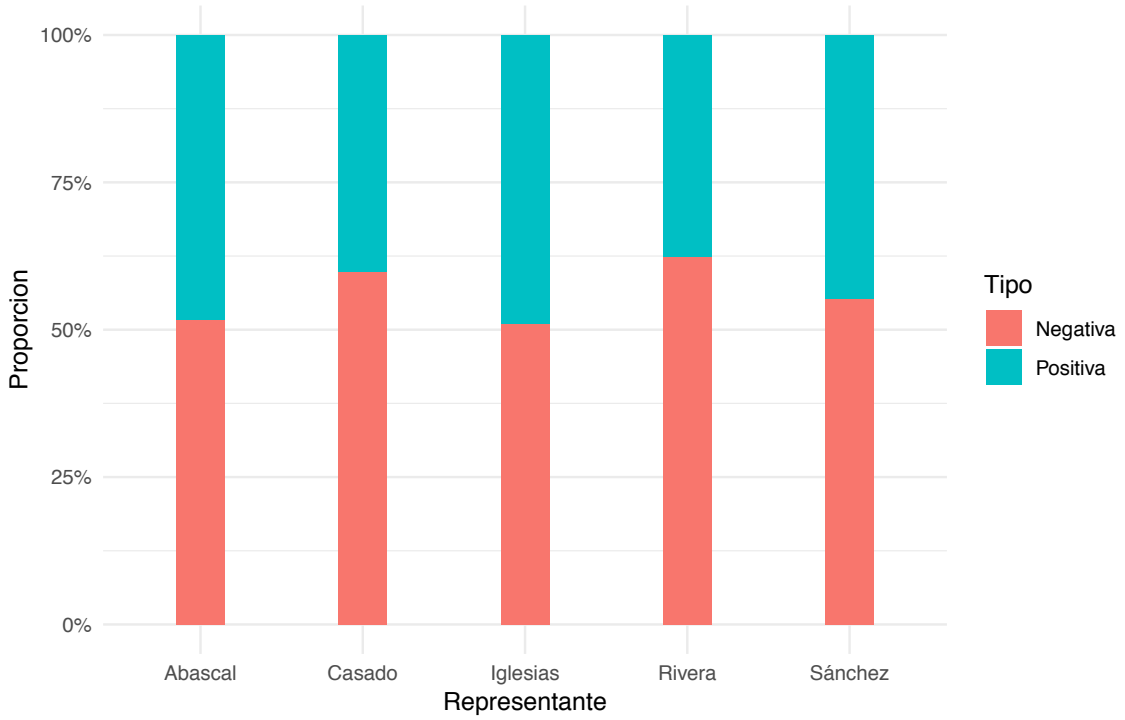


Figura 3.34: Porcentaje de tweets positivos y negativos por cuenta

Se puede observar, al igual que en las gráficas anteriores, que en todas las cuentas hay mayor número de tweets de carácter negativo que positivo. Sin embargo, en cuentas como la de Pablo Iglesias, Santiago Abascal y Pedro Sánchez ese porcentaje es cercano al 50 %. Las cuentas de Albert Rivera y Pablo Casado se encuentran sobrepasando el 60 % de tweets con connotación negativa.

Bibliografía

[1] Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J. and Chang, W. 2018. *Rmarkdown: Dynamic documents for r.*

[2] Amat Rodrigo, J. 2017. "Text mining con R: Ejemplo práctico twitter". Disponible en https://rpubs.com/Joaquin_AR/334526.

[3] Cubiles de la Vega, M.D. "Análisis de conglomerados".

[4] Dahl, D.B., Scott, D., Roosen, C., Magnusson, A. and Swinton, J. 2018. *Xtable: Export tables to latex or html.*

[5] Editorial, E. 2018. Análisis de sentimiento para una empresa. Disponible en https://reportedigital.com/transformacion-digital/analisis-de-sentimientos-para-una-empresa/.

[6] Gallardo San Salvador, J.A. "Métodos jerárquicos de análisis cluster". Disponible en https://www.ugr.es/~gallardo/pdf/cluster-3.pdf.

[7] (jpuigde), J. "Ggplot2 the easiest path to graphics". Disponible en https://rpubs.com/jpuigde/Ggplot2.

[8] Kabacoff, R. 2018. *Data visualization with* R.

[9] Luque-Calvo, P.L. 2017. *Escribir un trabajo fin de estudios con* R *markdown.* Disponible en http://destio.us.es/calvo.

[10] Martínez Martínez, F. J. 2017. *"Análisis de sentimiento en twitter de las principales compañías del sector asegurador español".*

[11] Mendoza Vega, J.B. "Introdución a la minería de textos con R". Disponible en https://rpubs.com/jboscomendoza/mineria-de-textos-con-r.

[12] País, A.P.E. 2019. Ganadores y perdedores del último debate. Disponible en https://elpais.com/politica/2019/04/23/actualidad/1556023173_939577.html.

[13] R Core Team *Package: "RColorBrewer".* R Foundation for Statistical Computing.

[14] R Core Team *Package: "Scales".* R Foundation for Statistical Computing.

[15] R Core Team *Package: "Tidytext".* R Foundation for Statistical Computing.

[16] R Core Team *Package: "Tidyverse".* R Foundation for Statistical Computing.

[17] R Core Team *Package: "Tm".* R Foundation for Statistical Computing.

[18] R Core Team *Package: "TwitteR".* R Foundation for Statistical Computing.

[19] R Core Team *Package: "Wordcloud".* R Foundation for Statistical Computing.

[20] R Core Team 2016. *R: A language and environment for statistical computing.* R

Foundation for Statistical Computing.

[21] RStudio Team 2015. *RStudio: Integrated development environment for* R. RStudio, Inc.

[22] Salazar, C. "Extraer tweets en R". Disponible en https://rpubs.com/camilamila/tweets2.

[23] Techopedia "Definition - what does business intelligence (bi) mean?" Disponible en https://www.techopedia.com/definition/345/business-intelligence-bi.

[24] Valencia Cabrera, L. "Conjunto de datos relacionales".

[25] Valencia Cabrera, L. "Manipulación de conjuntos de datos".

[26] Villena, J. 2015. Introducción al análisis de sentimientos (minería de opiniones).

[27] Vision, S. Integración con twitter desde r obtenido el 2017-08-21 04:16:10 -0400, desde el sitio web de synergy vision: /corpus/tecnologia/2017-08-21-twitter.html.

[28] Wei Xu y Tabassum, J. "Twitter api tutorial". Disponible en http://socialmedia-class.org/twittertutorial.html.

[29] Wickham, H. 2018. *Stringr: Simple, consistent wrappers for common string operations*.

[30] Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C. and Woo, K. 2018. *Ggplot2: Create elegant data visualisations using the grammar of graphics*.

[31] Wickham, H., François, R., Henry, L. and Müller, K. 2019. *Dplyr: A grammar of data manipulation*.

[32] Wikipedia "Twitter". Disponible en https://es.wikipedia.org/wiki/Twitter.

[33] Wikipedia. "Dendrograma". Disponible en https://es.wikipedia.org/wiki/Dendrograma.

[34] Xie, Y. 2018. *Knitr: A general-purpose package for dynamic report generation in r*.

[35] "Lenguaje R". Disponible en https://lenguajesdeprogramacion.net/r/.