

DETECCIÓN DE PATRONES, TENDENCIAS Y
TEMÁTICAS DURANTE LA CRISIS DEL COVID-19
EN TWITTER:

ANÁLISIS DE LAS CUENTAS DE POLÍTICOS Y GOBIERNOS DE ESPAÑA

AUTOR: IÑAKI URRUTIA SÁNCHEZ
26 DE JUNIO DE 2020

Resumen

Desde principios de este año, el mundo está sufriendo las consecuencias de la pandemia del COVID-19. Este virus ha puesto en jaque a la sociedad tal y como la conocemos, y ha provocado que muchos gobiernos tomen medidas que, en otras circunstancias, serían impensables.

Las consecuencias de esta pandemia están siendo especialmente duras en España, donde se han tomado medidas tan excepcionales como la declaración del segundo estado de alarma en la historia de nuestra democracia o el decreto de un estricto confinamiento a nivel nacional con el fin de frenar la expansión del virus, paralizando por completo la actividad del país.

Por otro lado, las redes sociales, y, especialmente, Twitter, son canales de comunicación que cada vez tienen más relevancia, especialmente en el plano de lo político, gracias a su capacidad de difusión, a la posibilidad que ofrece de conectar directamente con el ciudadano y a su inmediatez.

De la combinación de estas características de Twitter con esta situación tan inusual que vivimos nace la idea de este proyecto: aprovechar la creciente pero ya gran presencia de los políticos en Twitter para visualizar como han ido desarrollándose las tendencias, ideas y temáticas en sus discursos respecto de los diferentes contextos temporales que estamos viviendo en esta crisis.

Índice general

1. Introducción	1
1.1. Motivación	2
1.2. Propuesta de proyecto	3
1.3. Estructura de la memoria	3
2. Contexto y estado del arte	4
2.1. Twitter	5
2.1.1. Interacción básica	5
2.1.2. Tweet	7
2.2. API	9
2.2.1. API de Twitter	9
2.3. Procesamiento del lenguaje natural	11
2.3.1. Lenguaje natural	11
2.3.2. Procesamiento del lenguaje natural	11
2.3.3. Expresiones regulares	12
2.3.4. Preprocesamiento de texto	14
2.3.5. Limpieza del texto	14
2.3.6. Normalización del textos	14
2.3.7. <i>Tokenización</i>	15
2.3.8. <i>Stop words</i>	15
2.3.9. <i>Stemming</i>	15
2.4. Modelado de temáticas	20
2.4.1. <i>Latent Dirichlet Allocation</i>	20
2.4.2. Distribuciones de Dirichlet	21
2.5. Análisis de sentimientos en textos	23
3. Objetivo	24
4. Metodología	26
4.1. Planificación	27
4.2. Identificación de cuentas	29
4.3. Descarga de tweets	29
4.3.1. Autenticación	29
4.3.2. Descarga de tweets	30

4.4.	Preprocesamiento de texto	33
4.4.1.	<i>Tokenización</i>	36
4.4.2.	Stopwords	36
4.4.3.	Stemming	36
4.5.	Análisis de sentimientos	37
4.5.1.	Preparación del <i>dataset</i>	37
4.5.2.	Creación del modelo	39
4.6.	Marco experimental	41
4.6.1.	Recursos	41
4.7.	Herramientas	41
5.	Resultados	42
5.1.	Identificación de los principales actores políticos en Twitter	43
5.1.1.	Ministerios	43
5.1.2.	Gobiernos autonómicos	45
5.1.3.	Miembros del Congreso de los Diputados	46
5.2.	Conteo general	49
5.3.	Resultados por periodos temporales	56
5.4.	Periodo 1 (1 de enero - 31 de enero)	57
5.5.	Periodo 2 (1 de febrero - 16 de febrero)	61
5.6.	Periodo 3 (17 de febrero - 8 de marzo)	65
5.7.	Periodo 4 (9 de marzo - 15 de marzo)	69
5.8.	Periodo 5 (16 de marzo - 26 de marzo)	73
5.9.	Periodo 6 (27 de marzo - 5 de abril)	77
5.10.	Periodo 7 (6 de abril - 22 de abril)	81
5.11.	Periodo 8 (23 de abril - 10 de mayo)	86
5.12.	Periodo 9 (11 de mayo - 26 de mayo)	90
5.13.	Periodo 10 (27 de mayo - 8 de junio)	94
5.14.	Periodo 11 (9 de junio - 21 de junio)	99
6.	Conclusiones	104

Índice de figuras

2.1. Esquema básico de la interacción entre un usuario de Twitter y sus seguidores	6
2.2. Tweet de ejemplo	7
2.3. Tweet con <i>hashtag</i> de ejemplo	7
2.4. Tweet con mención de ejemplo	8
2.5. Ejemplo básico de <i>tokenización</i>	15
2.6. Esquema del LDA	21
2.7. Ejemplo de una distribución de Dirichlet de 3 categorías y su comportamiento según α	22
4.1. Diagrama de Gantt	28
4.2. Estructura de ficheros para almacenar tweets descargados	31
4.3. Diagrama del preprocesamiento de los Tweets descargados	35
4.4. Dos ejemplos de tweets con la estructura básica de los documentos del TASS	37
4.5. Ejemplo de la vectorización de los mensajes	38
4.6. Output del algoritmo de entrenamiento mientras busca los mejores hiperparámetros	39
4.7. Validación cruzada con $K = 5$	40
5.1. Composición actual del Congreso de los Diputados	47
5.2. Media de tweets diarios, agrupado por semanas	49
5.3. Media de tweets diarios que hacen referencia al COVID-19, agrupado por semanas	50
5.4. Porcentaje de tweets diarios que hacen referencia al COVID-19, agrupado por semanas	51
5.5. Hashtags enero	52
5.6. Hashtags febrero	53
5.7. Hashtags marzo	54
5.8. Hashtags abril	54
5.9. Hashtags mayo	55
5.10. Principales temáticas del PSOE en el primer periodo	57
5.11. Principales temáticas del PP en el primer periodo	58
5.12. Principales temáticas de VOX en el primer periodo	59
5.13. Principales temáticas de UP en el primer periodo	59
5.14. Análisis de sentimientos sobre la investidura de enero	60

5.15. Principales temáticas del PSOE en el segundo periodo	61
5.16. Principales temáticas del PP en el segundo periodo	62
5.17. Principales temáticas de VOX en el segundo periodo	62
5.18. Principales temáticas de UP en el segundo periodo	63
5.19. Análisis de sentimientos sobre partidos de ideología opuesta	64
5.20. Principales temáticas del PSOE en el tercer periodo	65
5.21. Principales temáticas del PP en el tercer periodo	66
5.22. Principales temáticas de VOX en el tercer periodo	66
5.23. Tweet de Santiago Abascal sobre el control de las fronteras en el inicio de la expansión del coronavirus	67
5.24. Principales temáticas de UP en el tercer periodo	67
5.25. Análisis de sentimientos sobre feminismo	68
5.26. Principales temáticas del PSOE en el cuarto periodo	69
5.27. Principales temáticas del PP en el cuarto periodo	70
5.28. Principales temáticas de VOX en el cuarto periodo	70
5.29. Principales temáticas de UP en el cuarto periodo	71
5.30. Análisis de sentimientos sobre el sistema sanitario	72
5.31. Principales temáticas del PSOE en el quinto periodo	73
5.32. Principales temáticas del PP en el quinto periodo	74
5.33. Principales temáticas de VOX en el quinto periodo	74
5.34. Principales temáticas de UP en el quinto periodo	75
5.35. Tweet de la cuenta oficial del Ministerio de Derechos Sociales, agradeciendo a la ciudadanía y al sector sanitario	75
5.36. Análisis de sentimientos sobre la gestión	76
5.37. Principales temáticas del PSOE en el sexto periodo	77
5.38. Principales temáticas del PP en el sexto periodo	78
5.39. Principales temáticas de VOX en el sexto periodo	78
5.40. Principales temáticas de UP en el sexto periodo	79
5.41. Análisis de sentimientos sobre la temática del material sanitario	80
5.42. Principales temáticas del PSOE en el séptimo periodo	81
5.43. Tweet de la cuenta oficial del Ministerio de Asuntos Exteriores hablando de la lucha contra la desinformación	82
5.44. Principales temáticas del PP en el séptimo periodo	82
5.45. Tweet de Pablo Casado criticando las intervenciones de Pedro Sánchez	83
5.46. Principales temáticas de VOX en el séptimo periodo	83
5.47. Principales temáticas de UP en el séptimo periodo	84
5.48. Análisis de sentimientos sobre la desescalada	85
5.49. Principales temáticas del PSOE en el octavo periodo	86
5.50. Tweet de Pablo Casado exigiendo tests masivos y mascarillas	87
5.51. Principales temáticas del PP en el octavo periodo	87
5.52. Principales temáticas de VOX en el octavo periodo	87
5.53. Tweet de Santiago Abascal culpando al Gobierno del confinamiento	88
5.54. Principales temáticas de UP en el octavo periodo	88

5.55. Análisis de sentimientos sobre el Gobierno	89
5.56. Tweet de Juan López de Uralde criticando a la oposición	89
5.57. Principales temáticas del PSOE en el noveno periodo	90
5.58. Principales temáticas del PP en el noveno periodo	91
5.59. Principales temáticas de VOX en el noveno periodo	91
5.60. Principales temáticas de UP en el noveno periodo	92
5.61. Análisis de sentimientos sobre Madrid	93
5.62. Principales temáticas del PSOE en el décimo periodo	94
5.63. Principales temáticas del PP en el décimo periodo	95
5.64. Tweet de Pablo Casado pidiendo el cese de Marlaska	95
5.65. Principales temáticas de VOX en el décimo periodo	96
5.66. Retweet de Javier Ortega Smith en el que relaciona el efecto llamada con el impuesto mínimo vital	96
5.67. Principales temáticas de UP en el décimo periodo	97
5.68. Tweet de Pablo Iglesias criticando las declaraciones de Cayetana Álvarez de Toledo sobre su padre	97
5.69. Análisis de sentimientos sobre el IMV	98
5.70. Principales temáticas del PSOE en el undécimo periodo	99
5.71. Principales temáticas del PP en el undécimo periodo	100
5.72. Tweet de Pablo Casado criticando la gestión de la crisis por parte del Gobierno	100
5.73. Principales temáticas de VOX en el undécimo periodo	101
5.74. Principales temáticas de UP en el undécimo periodo	101
5.75. Tweet de Pablo Iglesias sobre la oposición	102
5.76. Análisis de sentimientos sobre la gestión de las residencias de ancianos en Madrid	103

Índice de cuadros

2.1. Ejemplo de <i>stemming</i> en algunas palabras del castellano	19
4.1. Estructura del fichero que almacena la información de los actores políticos	30
5.1. Ministerios de España, ministros/as y sus respectivas cuentas en Twitter .	45
5.2. Gobiernos autonómicos, presidentes autonómicos y sus respectivas cuentas en Twitter	46
5.3. Diputados seleccionados para el estudio	48

Capítulo 1

Introducción

1.1. Motivación

Desde principios de este año, el mundo está sufriendo las consecuencias de la pandemia del COVID-19. Este virus ha puesto en jaque a la sociedad tal y como la conocemos, y ha provocado que muchos gobiernos tomen medidas que, en otras circunstancias, serían impensables.

Las consecuencias de esta pandemia están siendo especialmente duras en España, donde se han tomado medidas tan excepcionales como la declaración del segundo estado de alarma en la historia de nuestra democracia (el primero fue el declarado durante la crisis de los controladores aéreos en 2010) o el decreto de un estricto confinamiento a nivel nacional con el fin de frenar la expansión del virus, paralizando por completo la actividad del país. Estas medidas y, en general, la gestión de esta crisis sanitaria, han afectado directamente al ecosistema político de nuestro país, haciendo que las principales figuras políticas y sus discursos sean absolutos protagonistas en esta situación tan extraordinaria.

Por otro lado, las redes sociales, y, especialmente, Twitter, son canales de comunicación que cada vez tienen más relevancia, especialmente en el plano de lo político. Tanto es así que todos los ministerios del gobierno tienen cuentas oficiales y activas, así como todos los líderes políticos de los partidos políticos con representación en el parlamento y la inmensa mayoría de los diputados del Congreso. Todo esto hace de Twitter un medio muy útil para estar constantemente informado de la última hora de la vida política de nuestro país de una manera rápida, concisa y prácticamente instantánea [1].

Estas características de Twitter en combinación a una situación tan inusual como la provocada por la pandemia dan lugar a unos datos bastante impresionantes. Sólo en las cuatro primeras semanas de enero de este año ya había más de 15 millones de publicaciones en esta red social sobre el tema del coronavirus, y sólo el día en el que la OMS declaró la pandemia por COVID-19 se registraron cerca de 10 millones de publicaciones, y, se calcula que a mediados de abril había publicados más de 417 millones de tweets acerca del virus [2].

Además, las redes sociales son medios muy utilizados en cuanto a debate y discusión sobre temas relacionados con la salud, especialmente Twitter en cuanto a difusión de información de la salud [3]. Como mostraba una encuesta de 2011, en la que se concluía que el 62 % de los usuarios adultos de Internet en los Estados Unidos (72 millones) utilizan las redes sociales para temáticas relacionadas con la salud [4].

Y es con el respaldo de estos datos como nace la idea de este proyecto: aprovechar la creciente pero ya gran presencia de los políticos en Twitter para visualizar como han ido desarrollándose las tendencias, ideas y temáticas en sus discursos respecto de los diferentes contextos temporales que estamos viviendo en esta crisis.

1.2. Propuesta de proyecto

Con el fin de llevar a cabo el objetivo planteado en el último párrafo del apartado anterior, se plantea un proyecto que se compondrá de las siguientes partes:

1. Un sistema que se encargue de descargar y almacenar de manera automática todos los tweets de las cuentas de los principales actores políticos del país publicados durante el año 2020.
2. Un programa que automáticamente se encargue de procesar a gran escala dichos tweets.
3. La aplicación del modelo generativo *Latent Dirichlet Allocation* (LDA), que nos permitirá visualizar e interpretar las tendencias, temáticas y patrones de los tweets previamente procesados.
4. La aplicación de un modelo de análisis de sentimientos en texto que nos permitirá obtener el grado de positividad o negatividad de las principales formaciones sobre temas clave, utilizando nuestro modelo LDA para la identificación de esas temáticas clave.

1.3. Estructura de la memoria

Esta memoria queda estructurada de la siguiente forma:

- En el **capítulo 1** se presenta y describe brevemente el proyecto.
- En el **capítulo 2** se presenta y explica la teoría que hay bajo el proyecto.
- En el **capítulo 3** se explican tanto el objetivo principal del proyecto como los subobjetivos que hay que cumplir para alcanzar el principal.
- En el **capítulo 4** se explica como se ha implementado en el proyecto lo explicado en el capítulo 2.
- En el **capítulo 5** se exponen y analizan los resultados obtenidos.
- En el **capítulo 6** se sintetizan algunas conclusiones generales a través de los resultados previos.

Capítulo 2

Contexto y estado del arte

2.1. Twitter

En los últimos tiempos, el uso de las redes sociales se ha extendido muchísimo, hasta el punto que muchos usuarios utilizan varias a diario. Los datos corroboran este hecho, y es que de los casi 47 millones de españoles, hasta 29 millones usan las redes sociales diariamente, llegando a pasar en ellas una media de hasta dos horas diarias. En concreto, las redes sociales que más usan los españoles son, en orden, YouTube, WhatsApp, Facebook, Instagram, y, la que nos ocupa en este proyecto, Twitter [5].

Desde su fundación en marzo de 2006, Twitter no ha parado de crecer en popularidad e importancia, siendo la trigésima primera web en el ranking de tráfico de Alexa en 2019 a nivel mundial, y la quinta en el mismo ranking en español, solo superada por Google, Gmail, Facebook y Google Translator.

Esta creciente popularidad, especialmente marcada en países de habla hispana y específicamente en España, han hecho de Twitter una herramienta muy importante en la comunicación política, gracias a que permite una comunicación clara, concisa y sin intermediarios.

Estas razones hacen de Twitter una fuente de información muy valiosa, en muchos ámbitos y, en especial, en cuanto a política. De hecho, el 64% de los usuarios utiliza Twitter al menos una vez a la semana para informarse de temas políticos, y tres de cada cuatro usuarios españoles piensan que Twitter es muy útil a la hora de mantenerse informado de la última hora política [6].

Como podemos ver, el uso de Twitter pueden ser de mucha utilidad a la hora de obtener información que nos permita analizar como se desarrolla el contexto social y político de nuestro país en una situación tan extraordinaria como la provocada por el COVID-19. A continuación, vamos a ver como funcionan las mecánicas básicas de interacción en Twitter, así como qué es un tweet y qué tipo de información puede mostrar.

2.1.1. Interacción básica

En Twitter, los usuarios pueden seguirse los unos a los otros y además, entre otras cosas, pueden publicar tweets (para más detalles, ver Sección 2.1.2). Dichos tweets aparecerán en la página de inicio de cada seguidor del usuario que publicó el tweet en cuestión.

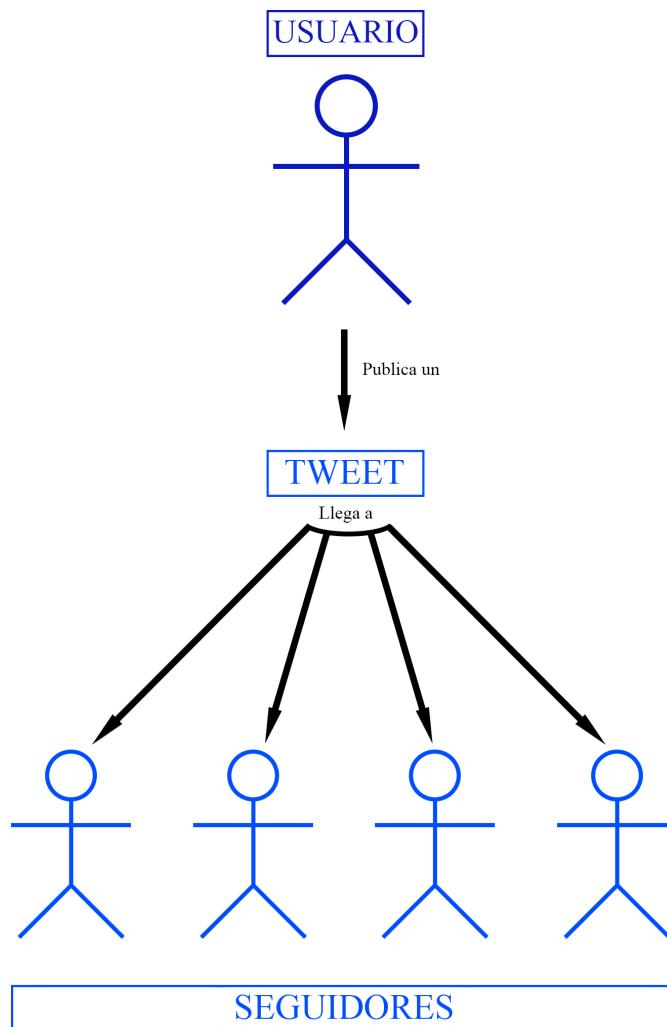


Figura 2.1: Esquema básico de la interacción entre un usuario de Twitter y sus seguidores

Esta interacción es la más importante, pero no la única. Otras posibles interacciones de un usuario hacia un tweet en concreto son las siguientes:

- **Responder:** Un usuario puede crear otro tweet como respuesta a un tweet en concreto.
- **Retwittear:** Un usuario puede publicar nuevamente un tweet, ya sea suyo o no. A esta acción se le llama *retwitear*.
- **Marcar como “Me gusta”:** Con esta función, un usuario puede indicar que un tweet le interesa.

2.1.2. Tweet

Un tweet es un mensaje de hasta 280 caracteres que, además del texto, puede contener hasta 4 imágenes, un archivo GIF o un video.



Figura 2.2: Tweet de ejemplo

Además, dentro del texto de un tweet podemos distinguir funciones especiales. Estas son dos, los *hashtags* y las menciones.

2.1.2.1. Hashtags

Los *hashtags* son palabras clave o temas dentro del texto de un tweet. Para utilizar esta función, solo hay que anteponer el símbolo # al texto que queremos marcar como *hashtag*. De esta forma, el tweet en cuestión será relacionado con el *hashtag* y aparecerá más fácilmente en la búsqueda de Twitter.



Figura 2.3: Tweet con *hashtag* de ejemplo

2.1.2.2. Menciones y respuestas

Las menciones son una función que sirve para aludir directamente a otro usuario. Para mencionar a alguien, solo hay que anteponer el símbolo @ al nombre del usuario al que queramos mencionar. Otro uso del @ en texto son las respuestas. Un usuario puede responder a un tweet específico utilizando el mismo formato que las menciones.



Figura 2.4: Tweet con mención de ejemplo

2.2. API

Una interfaz de programación de aplicaciones, más conocida por sus siglas en inglés **API**, es un conjunto de funciones, subrutinas y procedimientos que provee acceso a cierto software ocultando los detalles de implementación del mismo.

Uno de los pasos necesarios en este proyecto es la descarga a gran escala de los tweets de los principales actores políticos nacionales. Para poder llevar esto a cabo, es necesario usar la API de Twitter. Por esto, en esta sección, explicaremos cómo funciona la API de Twitter.

2.2.1. API de Twitter

Twitter da la opción a sus usuarios de solicitar que sus cuentas adquieran permisos de desarrollador. Con este permiso, el usuario en cuestión tiene acceso a unas credenciales de desarrollador, con las que tendrá podrá usar la API de Twitter [7].

Una vez tengamos acceso a la API de Twitter podremos empezar a trabajar con ella. En nuestro caso, como solo queremos descargar tweets de un grupo de personas, tendremos que usar el *endpoint*¹ *GET statuses/user_timeline* [8]. Algunas de las características y limitaciones de este *endpoint* son las siguientes:

- Al hacer una petición, nos devolverá un vector de los tweets (más adelante hablaremos de la información que podemos obtener de cada tweet) más recientes del usuario que queramos.
- Como máximo, podremos recuperar los 3200 tweets más recientes de cada usuario.
- Podemos hacer hasta 900 peticiones cada 15 minutos.

Objeto tweet

Como hemos mencionado anteriormente, el *endpoint* que usamos para recuperar tweets nos devuelve un vector de “tweets”. Sin embargo, como podemos ver en la documentación oficial de la API de Twitter [9], estos tweets no son simplemente el texto del mensaje, sino que son objetos de tipo JSON con más información de la que podría parecer a primera vista.

¹Endpoint: Proceso que recibe o devuelve información de una API

Algunos atributos que aprovecharemos son los siguientes:

- *text* nos devuelve el texto del tweet en cuestión
- *created_at* nos devuelve en qué fecha se publicó el tweet. Este atributo nos será útil para localizar en qué contexto temporal de la crisis de COVID-19 se encontraba cada tweet.
- *retweeted_status* nos dice si estamos ante un tweet propio o ante un retweet.
- *retweet_count* nos dice el número de retweets que tiene un tweet.
- *favorite_count* nos dice el número de usuarios que han indicado que les gusta un tweet.
- *city*. En el caso de que el usuario tuviese activada la opción de localización que brinda Twitter, este atributo nos mostraría dónde estaba su usuario al publicarlo.

2.3. Procesamiento del lenguaje natural

En esta sección veremos el proceso por el que pasa cada tweet que descargamos para convertirlo en información valiosa con la que podamos trabajar y los fundamentos del mismo proceso.

2.3.1. Lenguaje natural

Para entender la necesidad de la que surge el procesamiento del lenguaje natural, primero necesitamos entender qué es el lenguaje natural [10]. En términos simples, podemos definir el concepto de lenguaje natural como aquel lenguaje desarrollado y evolucionado por los humanos a través de su uso natural. De esta forma, el lenguaje natural no está diseñado de forma artificial, como lo está, por ejemplo, un lenguaje de programación, sino que se va puliendo según las necesidades de sus usuarios.

Precisamente, esta falta de un diseño artificial característica de los lenguajes naturales es lo que les hace ser difíciles de tratar en el plano de la informática. Aquí es donde entra el procesamiento del lenguaje natural.

2.3.2. Procesamiento del lenguaje natural

El procesamiento del lenguaje natural (por sus siglas en inglés, NLP) es un campo de estudio dentro de la inteligencia artificial. Su objetivo principal es el diseño y la construcción de aplicaciones y sistemas que permitan la interacción entre usuarios y máquinas usando el lenguaje natural que utilizamos los humanos.

El NLP tiene múltiples usos, como la creación de traductores, los sistemas de reconocimiento del habla, sistemas de búsqueda de respuestas, sistemas que resumen textos automáticamente, o, por supuesto, como en el caso de este proyecto, la creación de sistemas de clasificación de textos.

A continuación, desgranaremos el proceso por el cual pasa cada tweet descargado para convertirse en un documento apto para trabajar con él, pero antes hace falta introducir el concepto de expresión regular.

2.3.3. Expresiones regulares

Las expresiones regulares son secuencias de caracteres que conforman patrones de búsqueda en cadenas de caracteres. Un ejemplo muy simple podría ser la expresión regular:

Es

Que si la aplicásemos a la oración *Estamos esperando a que España supere la crisis del COVID-19*, obtendríamos que la expresión regular identifica las subcadenas subrayadas, pero no detectaría nada en la palabra *esperando*, ya que nuestra expresión regular espera una *E* mayúscula seguida de una *s* minúscula.

El ejemplo que acabamos de ver como introducción era algo trivial, ya que la verdadera potencia de las expresiones regulares reside en los caracteres especiales, que permiten una flexibilidad que nos ayuda a crear expresiones regulares mucho más complejas. A continuación vamos a ver algunos de los caracteres especiales más importantes, con algunos ejemplos aclaratorios (estamos usando la sintaxis de la biblioteca *re* de Python, puede variar respecto a otras implementaciones de expresiones regulares):

- **Asterisco (*).** Se utiliza para encontrar algo que se repita 0 o más veces. Por ejemplo, la expresión regular

*ab**

aceptaría las cadena *a* ; *ab* o *aaaaaaaaaa* ; o sea, cualquier cadena compuesta por una *a* y seguida de cualquier número de *b*.

- **Signo de suma (+).** Se utiliza para encontrar algo que se repita al menos una vez. Por ejemplo, la expresión regular

ab+

no aceptaría las cadena *a* ; pero si aceptaría *ab* o *aaaaaaaaaa* ; o sea, cualquier cadena compuesta por una *a* y seguida de al menos una *b*.

- **El punto (.).** Se utiliza para encontrar cualquier carácter que no sea un salto de línea. Por ejemplo, la expresión regular

a.c

aceptaría las cadenas *abc* ; *a!c* ; *a!c*; o sea, cualquier cadena compuesta por una *a* seguida de cualquier carácter seguido de una *c*.

- **Las llaves ({ })**. Se utilizan para determinar el rango de repeticiones permisibles de una expresión regular. Por ejemplo, la expresión regular

$$ab\{2, 4\}$$

no aceptará la cadena ab , porque al menos deben haber dos b , ni la cadena $abbbbb$, porque como máximo aparecerán cuatro b , pero sí las cadenas abb , $abbb$ o $abbbb$. También podemos especificar el número exacto de repeticiones en vez de dar un rango, así la expresión regular

$$ab\{3\}$$

sería equivalente a la expresión regular

$$abbb$$

- **Los corchetes ([])**. Se utilizan para representar clases de caracteres. Por ejemplo, la expresión regular

$$[abc]123$$

aceptará las cadenas $a123$; $b123$ o $c123$, pero no la cadena $ab123$; es decir, aceptará compuestas por una a o una b o una c seguidas de la subcadena 123 .

- **La contrabarra (\)**. Tiene dos funciones:

1. Escapa un carácter especial. Por ejemplo, en la expresión regular

$$\backslash.$$

el punto no aceptará cualquier carácter como explicamos más arriba, sino que simplemente aceptará un punto.

2. También podemos usar la contrabarra seguida de una serie de caracteres que indican grupos especiales. Estos son algunos ejemplos:

- \d: Cualquier dígito del 0 al 9.
- \D: Cualquier carácter que no sea un dígito del 0 al 9.
- \s: Un espacio en blanco.
- \S: Cualquier carácter que no sea un espacio en blanco.
- \w: Cualquier carácter alfanumérico.
- \W: Cualquier carácter no alfanumérico.
- \A: Representa la posición de inicio de una cadena.
- \Z: Representa la posición de fin de una cadena.

2.3.4. Preprocesamiento de texto

El preprocesamiento de texto es la fase en la que preparamos nuestros documentos para el procesamiento del lenguaje natural. Esta etapa es clave en el resultado final, ya que asegurar que los datos con los que trabajamos están en las mejores condiciones posibles es fundamental para obtener resultados de calidad.

2.3.5. Limpieza del texto

La limpieza del texto suele ser la primera fase del preprocesamiento, ya que es la que se encarga de dejar el texto apto para el resto de procesos.

Muchas veces los documentos que conforman nuestro *corpus* no son simples documentos de texto plano, sino que han sido que vienen dentro de etiquetas HTML, o en un JSON, o en etiquetas XML, etc. Por esto, es importante identificar y separar el texto que nos interesa del resto de elementos. Al ser una tarea tan dependiente del formato en el que estemos trabajando, no hay una sola técnica que nos permita limpiar cualquier texto, sino que depende de cada caso.

2.3.6. Normalización del textos

Otra tarea del preprocesamiento es la normalización del texto. Al igual que la limpieza del texto, la normalización depende del formato de texto sobre el que estemos trabajando y de qué queramos hacer con él.

Sin embargo, algunas de las técnicas más habituales de normalización de textos pueden ser las siguientes:

- **Eliminación de signos de puntuación.** Normalmente, los puntos, comas y demás signos de puntuación no aportan significado a los textos, por lo que suelen eliminar de los documentos.
- **Normalización de mayúsculas/minúsculas.** En la mayoría de ocasiones, el uso de mayúsculas y minúsculas no tiene ningún valor semántico. Por esto se pasa todo el texto o bien a minúsculas o a mayúsculas.

2.3.7. Tokenización

En procesamiento del lenguaje, llamamos *token* a la unidad sintáctica mínima que una máquina puede entender y procesar, por lo que debemos *tokenizar* cualquier cadena de texto que queramos procesar. La *tokenización* es el proceso por el cual dividimos el texto original en *tokens*, y la complejidad de este proceso depende en gran parte de la naturaleza del propio idioma. Por ejemplo, en el caso del inglés o el español puede ser tan fácil como usar una expresión regular que nos permita identificar números y palabras, pero en otros idiomas, como el chino o el japonés, es una tarea mucho más compleja [14].



Figura 2.5: Ejemplo básico de *tokenización*

2.3.8. Stop words

Las palabras vacías, o mejor conocidas por su nombre en inglés, *stop words*, son el conjunto de palabras sin significado, como pueden ser artículos, pronombre, preposiciones, verbos auxiliares, etc. que son eliminadas de los documentos.

2.3.9. Stemming

Llamamos lenguas flexivas a aquellas en las que existe el concepto de flexión de palabras, que básicamente es la alteración de las palabras según su función dentro de una oración y sus relaciones de dependencia con el resto de palabras. En el caso del español, podemos distinguir dos tipos de flexión de palabras:

- **Flexión nominal.** Es aquella en la que al lexema² se le añaden morfemas que definen el número, género o caso gramatical de la palabra. Por ejemplo, el lexema *niñ-* se hace más específico con los morfemas *-o* (masculino), *-a* (femenino), *-s* (plural) o la propia ausencia de morfema, que indica el singular.
- **Flexión verbal.** También conocida como conjugación, es aquella en la que participa un lexema verbal. Los morfemas que constituyen esta flexión expresan tiempo, aspecto, modo, número y persona.

Esta característica del lenguaje puede ser problemática en el campo del procesamiento del lenguaje natural, ya que normalmente no nos interesa diferenciar dos palabras por características como por ejemplo el género o el número. Aquí es donde entra en juego el *stemming*, que es el proceso que se encarga en extraer la raíz (*stem*) de cada *token*. En este caso, vamos a usar el algoritmo *Snowball* de *stemming* en castellano.

2.3.9.1. Algoritmo de *stemming Snowball* (versión castellano)

A continuación vamos a profundizar en el algoritmo que vamos a usar para hacer el *stemming* en nuestros documentos. Antes de entrar en el procedimiento en sí, es necesario definir unos cuantos conceptos [16] que usan la mayoría de los *stemmers*:

- R1. Es la zona que empieza después de la primera consonante que sigue a una vocal y acaba en el final de la palabra, o es una zona vacía en caso de no existir dicha consonante. Por ejemplo, en la palabra

plausible

, la zona subrayada es R1.

- R2. Es la zona que empieza después de la primera consonante que sigue a una vocal dentro de R1 y acaba en el final de la palabra, o es una zona vacía en caso de no existir dicha consonante. Siguiendo el ejemplo anterior, en la palabra

plausible

, la zona subrayada es R2.

- RV. Su definición tiene cuatro casos posibles según la palabra en cuestión:

²Lexema: Parte invariante de una familia de palabras, y que además refleja el significado principal de la misma.

1. Si la segunda letra es una consonante, RV es la zona que empieza después de la siguiente vocal y acaba al final de la palabra. Por ejemplo, en la palabra

oliva

, la zona subrayada es RV.

2. Si las dos primeras letras son vocales, RV es la zona que empieza después de la siguiente consonante y acaba al final de la palabra. Por ejemplo, en la palabra

áureo

, la zona subrayada es RV.

3. Si no se cumplen ninguna de las dos primera reglas, RV es la zona que empieza después de la tercera letra, y que acaba al final de la palabra. Por ejemplo, en la palabra

macho

, la zona subrayada es RV.

4. Si la palabra no satisface ninguna de las reglas anteriores, RV será una zona vacía situada al final de la palabra.

Una vez hechas estas definiciones, vamos a ver paso a paso como funciona el *stemming Snowball* para cualquier palabra *p* en castellano:

1. Pronombres enclíticos: Buscamos en *p* los siguientes sufijos:

me se sel a selo selas selos la le lo las les los nos

y de las coincidencias, eliminamos el de mayor longitud, siempre que esté precedido de uno de los siguientes grupos de morfemas dentro de RV:

- a) Morfemas acentuados:

iéndo ándo ár ér ír

En este caso, también hemos de eliminar la tilde de los morfemas anteriores. Por ejemplo: *haciéndola* pasaría a ser *haciendo*.

- b) Morfemas no acentuados:

ando iendo ar er ir

- c) Y, siempre que una *u* le preceda:

yendo

En este caso, la *u* puede estar tanto dentro como fuera de RV.

2. Eliminación de sufijo estándar. De todos los sufijos que se exponen a continuación, se busca el de mayor longitud y se aplica la regla que le corresponda:

- a) Se eliminará si está en R2:

anza anzas ico ica icos icas ismo ismos able ables ible ibles ista istas oso osa osos osas amiento amientos imiento imientos

- b) Se eliminará si está en R2, y además, si está precedido por *ic*, e *ic* también está en R2, también se eliminará:

adora ador acción adoras adores acciones ante antes ancia ancias

- c) Si está en R2, se reemplazará por *log*:

logía logías

- d) Si está en R2, se reemplazará por *u*:

encia encias

- e) Si está en R1, se eliminará. Si está en R2 precedido por *iv* y posteriormente precedido por *at*, también se eliminará *at*. También se eliminan *os*, *ic* o *ad* si están en R2 y preceden a:

amente

- f) Si está en R2, se eliminará, y, si le preceden *ante*, *able* o *ible* y se encuentran en R2, estos también se eliminan:

mente

- g) Si está en R2, se eliminará, y, si le preceden *abil*, *ic* o *iv* y se encuentran en R2, estos también se eliminan:

idad idades

- h) Si está en R2, se eliminará, y, si le precede *at* y se encuentra en R2, este también se elimina:

iva ivo ivas ivos

3. Sufijos verbales empezados en *y*. **Este paso se hace solamente cuando en el anterior no se elimina nada.** Se buscarán los siguientes sufijos en RV, y si están precedidos por *u*, se eliminarán:

ya ye yan yen yeron yendo yo yó yas yes yais yamos

4. Otros sufijos verbales. **Otra vez, este paso se hace solamente cuando en el anterior no se elimina nada.** Se buscarán y eliminarán los siguientes sufijos en RV:

en es éis emos

Además, si estaban precedidos por *gu*, eliminaremos la *u*. En cambio, si encontramos cualquiera de los siguientes sufijos, lo eliminamos sin más:

*arían arías arán arás aráis aría aréis aríamos aremos ará aré erían
erías erán erás eráis ería eréis eríamos eremos erá eré irían irías irán
irás iráis iría iréis iríamos iremos irá iré aba ada ida ía ara iera ad ed
id ase iese aste iste an aban ían aran ieran asen iesen aron ieron ado
ido ando iendo ió ar er ir as abas adas ías aras ieras ases ieses ís
áis abais íais arais ierais aseis asteis isteis ados idos amos
ábamos íamos imos áramos iéramos iésemos ásemos*

5. Sufijo residual. Buscamos los siguientes sufijos en *p*:

os a o á í ó

, y si están en RV, los eliminamos. También buscamos los siguientes sufijos:

e é

, y los eliminamos si están en RV. Además, si están precedidos por *gu* con la *u* dentro de RV, eliminamos dicha *u*.

6. Y, finalmente, eliminamos las tildes.

Palabra original	Stemmed
Niño	Niñ
Niña	Niñ
Saltar	Salt
Saltaría	Salt
Saltado	Salt
Rojo	Roj
Rojas	Roj
Torpedo	Torped
Torpedearon	Torped

Cuadro 2.1: Ejemplo de *stemming* en algunas palabras del castellano

Como podemos ver, los algoritmos de *stemming* están fuertemente ligados al idioma al que se aplican, ya que las reglas se desarrollan según su ortografía y gramática, y estas pueden ser completamente diferentes de un idioma a otro.

2.4. Modelado de temáticas

El modelado de temáticas [17] es una técnica de aprendizaje automático no supervisado basado en modelos estadísticos que es capaz de procesar un conjunto de documentos para agrupar las palabras que lo forman en grupos temáticos similares. Uno de los modelos de reconocimiento de temáticas más conocidos es el *Latent Dirichlet Allocation* (de ahora en adelante, LDA), que es el que usaremos para este proyecto.

2.4.1. *Latent Dirichlet Allocation*

El LDA es un modelo generativo cuya idea fundamental es cada documento que forma el *corpus*³ puede definirse como una distribución de temas, y cada tema puede definirse como una distribución de palabras [18]. A continuación vamos a ver como funciona el LDA. Para ello, antes tendremos que definir cierta notación:

- **k** : Número **fijo** que define el número de temas a los que un documento pertenece.
- **V** : Número total de palabras
- **M** : Número de documentos.
- **N** : Número de palabras en un documento.
- **w** : Representa una palabra en un documento
- **w**: Representa un documento (i.e. un vector de **w**) de **N** palabras.
- **D** : Es una colección de **M** documentos (i.e. el corpus)
- **z** : Es un tema definido como la ponderación del conjunto de **k** temas.

Como podemos ver en el esquema 2.6, el valor α definirá ϑ , siendo ϑ la distribución de las temáticas en los documentos. Además, tenemos M documentos, cada uno de ellos con una distribución ϑ particular. Para entender mejor la esencia del funcionamiento del LDA, vamos a asumir que $M = 1$, luego solo existe un documento w en el corpus.

³Corpus: Conjunto de datos o textos científicos, literarios, etc., que sirven de base a una investigación.

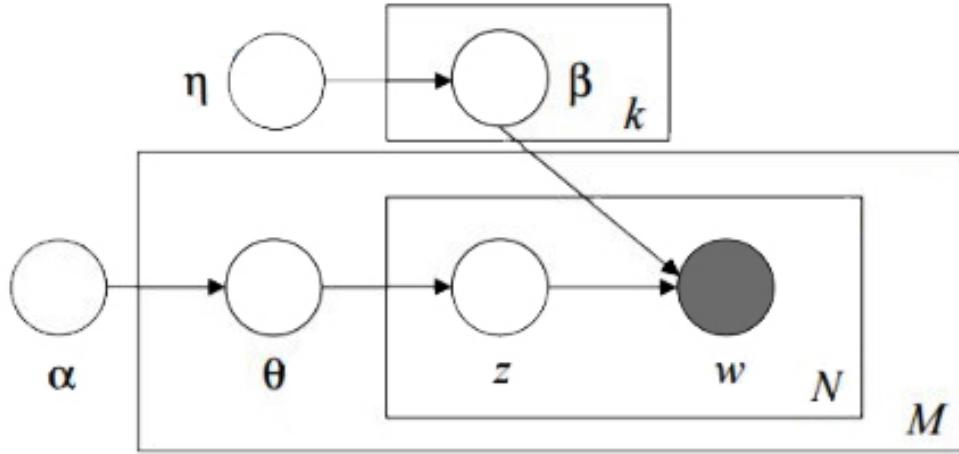


Figura 2.6: Esquema del LDA

Dicho documento w está formado por N palabras, y cada una de estas palabras debe estar relacionada a un tema z , como dijimos al principio. De esto se encarga la distribución ϑ , que asigna una temática a cada una de las palabras que conforman w .

Ya tenemos un tema asignada a cada una de las N palabras que componen nuestro documento w . Ahora nos queda asignar una palabra a cada uno de estos temas. Aquí entra en juego la distribución β , definida por η , que asignará una palabras a cada tema z , de manera que nuestro documento w queda compuesto por N palabras, donde cada palabra ha sido definida por una temática dada por β , que ha su vez ha sido definida por la distribución ϑ .

Así es como el modelo LDA genera un documento. En el caso en el que M fuera mayor que 1, sólo tendríamos que repetir el proceso para cada uno de los M documentos.

2.4.2. Distribuciones de Dirichlet

Para terminar de comprender el funcionamiento del LDA hay que saber que ϑ y β son distribuciones de Dirichlet. Ahora vamos a ver, de manera intuitiva, qué es una distribución de Dirichlet y cómo α y η afectan a ϑ y a β , respectivamente.

Una distribución de Dirichlet es la generalización multivariada de una distribución beta. Cada distribución de Dirichlet tiene dos parámetros [19]:

1. **K:** Es un entero que define el número de categorías. Volviendo al LDA, en el caso de ϑ determinará el número total de temáticas en el corpus, y en el caso de β determinará cuantas palabras hay en cada temática.
2. **Parámetros de concentración :** Es un vector de k elementos que definirá la densidad de probabilidad de cada categoría. En el caso de LDA, α es el parámetro de concentración de ϑ y η es el de β .

En la figura 2.7 podemos ver como el parámetro de concentración α afecta a una distribución de Dirichlet con $K = 3$ (i.e. con 3 categorías). A modo de resumen, podemos ver que con valores cercanos a 1, la distribución tiende a ser equiprobable; con valores menores a 1, la distribución de probabilidad tiende a agruparse más en las esquinas; y en el caso contrario, con valores mayores a 1, la densidad de probabilidad suele ser mucho mayor en zonas centrales.

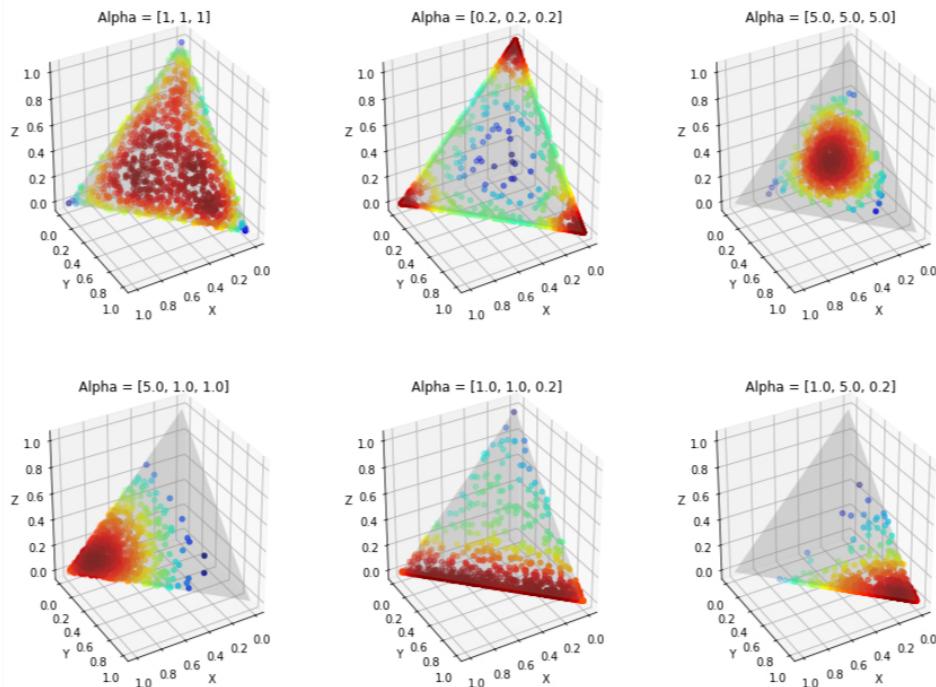


Figura 2.7: Ejemplo de una distribución de Dirichlet de 3 categorías y su comportamiento según α

Volviendo al LDA y extrapolando lo aprendido sobre distribuciones de Dirichlet, podemos ver como el problema se sitúa ahora encontrar los valores adecuados para α y η . Para lograr esto hay múltiples técnicas, entre las que destacan el muestreo de Gibbs o los métodos variacionales de Bayes.

2.5. Análisis de sentimientos en textos

El análisis de sentimientos, o también conocido como minería de opinión, es una rama de la minería de texto que trata de identificar información de carácter subjetiva de textos, como puede ser por ejemplo la polaridad (i.e: su positividad o negatividad). En los últimos tiempos, el análisis de sentimientos en textos ha ganado mucha relevancia, gracias al crecimiento de servicios de micro-blogging (como Twitter), ya que estos suelen contener textos muy concisos y con sentimientos claramente marcados [20].

Existen múltiples modelos de análisis de sentimientos en textos, pero algunos de los pasos generales que suelen compartir la mayoría de ellos son los siguientes:

- Preprocesamiento. Los documentos han de ser preprocesados, tanto los que se van a usar para entrenar el modelo como los que queremos analizar. Este procesamiento dependerá del tipo de documento que vamos a utilizar, así como del idioma de los mismos.
- Indexación. Se crea un índice con los documentos preprocesados, que suele ser un **índice invertido**.
 - Índice invertido: Índice representable mediante una matriz *término-documento*, usando el cálculo tf-idf.
 - Tf-idf: Valor numérico que expresa cómo de relevante es una palabra en un conjunto de documentos.
- Selección de características. Utilizamos un modelo de selección de características que nos permite identificar qué palabras son más influyentes a la hora de influir en la positividad/negatividad de una oración.
- Ponderación. Utilizando algoritmos de aprendizaje automático (normalmente supervisado) calibraremos en qué medida y cómo deben influir cada una de las palabras que previamente seleccionamos en el grado de positividad/negatividad (polaridad) de una oración.

Para ello, necesitamos disponer de una colección de documentos previamente clasificados según su polaridad. Es importante seleccionar bien estos documentos, ya que cuanto más amplia y de calidad sea dicha colección, mejor será nuestro modelo.

Capítulo 3

Objetivo

El objetivo principal de este proyecto es realizar un análisis de los tweets de los principales actores políticos de España durante la crisis del COVID-19 con el fin de identificar las diferentes temáticas de los principales partidos políticos según el contexto temporal, así como evaluar el grado de positividad de estos partidos políticos respecto a temáticas clave.

Este objetivo principal se divide en los siguientes objetivos secundarios:

1. Comprender las técnicas del procesamiento del lenguaje natural.
2. Analizar las diferentes técnicas de extracción de temáticas en textos.
3. Entender los pasos para crear un modelo de análisis de sentimientos en textos.
4. Identificar los principales actores políticos en España y crear un sistema para descargar y almacenar sus tweets.
5. Dividir el espacio temporal de la crisis del COVID-19 en España en intervalos significativos.
6. Aplicar técnicas de procesamiento del lenguaje natural para procesar los tweets.
7. Entrenar un modelo de análisis de sentimientos y aplicarlo a los tweets pertinentes.
8. Identificar las diferentes temáticas aplicando técnicas de extracción de temáticas en textos en los diferentes intervalos temporales.

Capítulo 4

Metodología

4.1. Planificación

El desarrollo de este proyecto se divide en tres bloques fundamentales:

1. El bloque de **aprendizaje**, que tendrá lugar durante las primeras semanas y que tendrá como objetivo comprender y asimilar la teoría de las técnicas que posteriormente se implementarán en el proyecto. Estas técnicas serán las siguientes:
 - a) Técnicas de procesamiento del lenguaje natural.
 - b) Técnicas de identificación de temáticas en textos.
 - c) Técnicas de análisis de sentimientos en textos.
2. El bloque de **desarrollo**, que empezará pasadas las primeras semanas y durará hasta casi el final del proyecto. En esta etapa se implementaran los modelos y sistemas necesarios para el objetivo del proyecto, que son las siguientes:
 - a) Sistema de descarga y almacenamiento de tweets.
 - b) Sistema de procesamiento de tweets.
 - c) Modelo de identificación de temáticas en textos (LDA).
 - d) Modelo de análisis de sentimientos.
3. El bloque de realización de la **memoria** y de la **presentación**, que comenzará prácticamente al mismo tiempo que la etapa de desarrollo y durará hasta el fin del proyecto. En este bloque se recopilarán y analizarán los resultados obtenidos con las herramientas que se implementarán en la etapa de desarrollo. Podemos dividir este bloque en los siguientes apartados:
 - a) Desarrollo de la memoria.
 - b) Creación de las figuras y tablas para la memoria.
 - c) Creación de la presentación para la defensa del proyecto.

Podemos ver una representación gráfica de la planificación del proyecto en el diagrama de Gantt de la Figura 4.1:

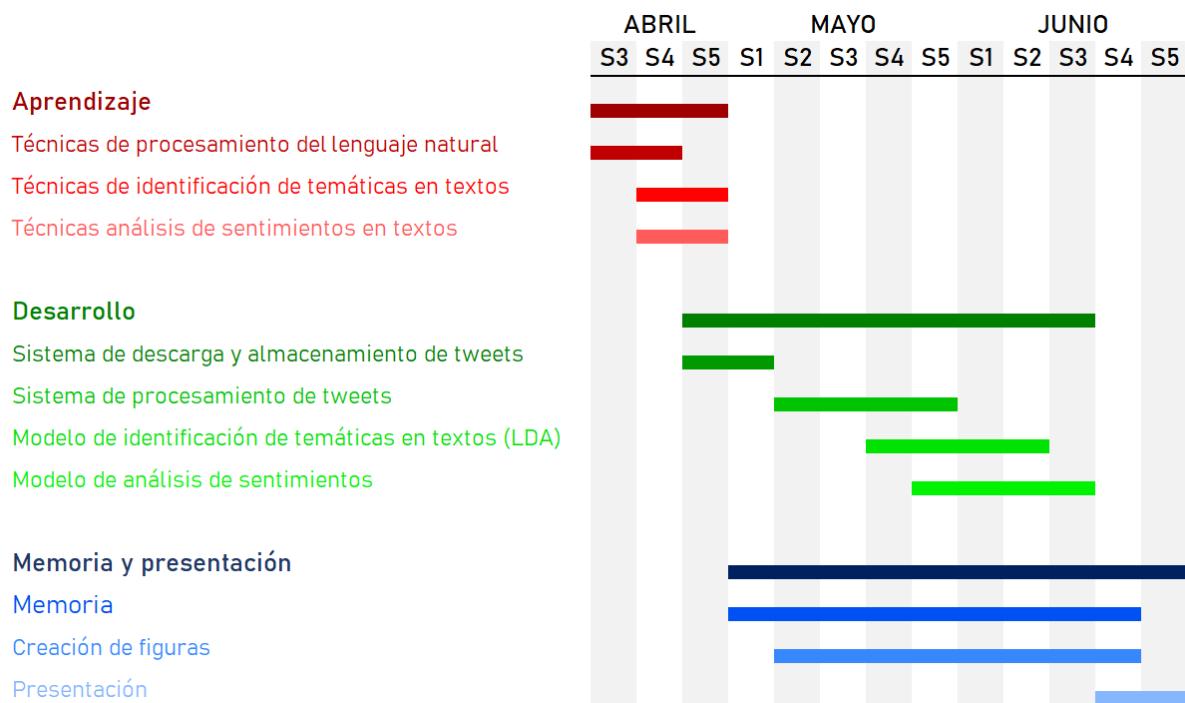


Figura 4.1: Diagrama de Gantt

4.2. Identificación de cuentas

Es importante poder identificar qué políticos/organizaciones políticas tienen cuentas en Twitter que nos puedan ser útiles para generar un *corpus* de calidad para el proyecto. Así, usaremos los tweets publicados durante el año 2020 de las cuentas que se ajusten a los siguientes criterios:

- Cuentas oficiales de todos los ministerios y de sus máximos responsables.
- Cuentas oficiales de todos los gobiernos autonómicos y de sus presidentes.
- De los partidos con diez diputados o más en el Congreso, las cuentas oficiales de los primeros diez candidatos al Congreso de Diputados en la lista electoral de cada uno de esos partidos.

4.3. Descarga de tweets

En este punto, ya tenemos seleccionadas las cuentas de Twitter vamos a usar para obtener los documentos (tweets) que formen nuestro *corpus*. El siguiente paso es hacer un *script* capaz de recuperar y almacenar todos los tweets de cada una de las cuentas de forma automática. Para ello, usaremos *Tweepy*, una librería para Python que nos permitirá usar la API de Twitter.

4.3.1. Autenticación

El primer paso para poder usar *Tweepy* es autenticarnos con las correspondientes claves que nos da Twitter una vez nos concede los permisos de desarrollador. El siguiente fragmento de código en Python ilustra cómo podemos hacerlo:

```
import tweepy

auth = tweepy.OAuthHandler(consumer_token, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)
```

4.3.2. Descarga de tweets

Ya tenemos tanto las cuentas que nos interesan identificadas como el objeto *api* listo para ser usado, por lo que ya podemos pasar a descargar los tweets. En este caso, hemos almacenado información básica sobre las cuentas en un archivo Excel que tiene la estructura que muestra la tabla 4.1.

Nombre del actor político 1	@UsuarioTwitterActorPolitico1	Partido político al que pertenece el actor político 1
Nombre del actor político 2	@UsuarioTwitterActorPolitico2	Partido político al que pertenece el actor político 2
Nombre del actor político 3	@UsuarioTwitterActorPolitico3	Partido político al que pertenece el actor político 3
...
Nombre del actor político N	@UsuarioTwitterActorPoliticoN	Partido político al que pertenece el actor político N

Cuadro 4.1: Estructura del fichero que almacena la información de los actores políticos

Por otro lado, la estructura de ficheros será la siguiente (figura 4.7) :

- Cada partido político que tenga a un miembro en nuestra lista tendrá una carpeta.
- En cada carpeta de cada partido político se guardará un archivo Excel por cada miembro del partido que figure en nuestra lista.
- En cada Excel se creará una hoja Excel por cada mes del año.
- En cada hoja Excel, se guardarán los tweets descargados del mes correspondiente.

Para asegurar que la estructura de ficheros es la correcta, un script se ejecutará antes de comenzar el proceso de descargas, y comprobará que todo esté configurado adecuadamente, o en caso de que no lo estuviese, creará los archivos o ficheros que faltasen.

Una vez verifiquemos que la estructura de ficheros es la esperada, otro script empezará la descarga de tweets. Para ello usamos el objeto *Cursor* que nos ofrece *Tweepy* [22]. Este objeto nos permite navegar y descargarnos tweets de la *timeline* del usuario que indiquemos de una manera muy sencilla.

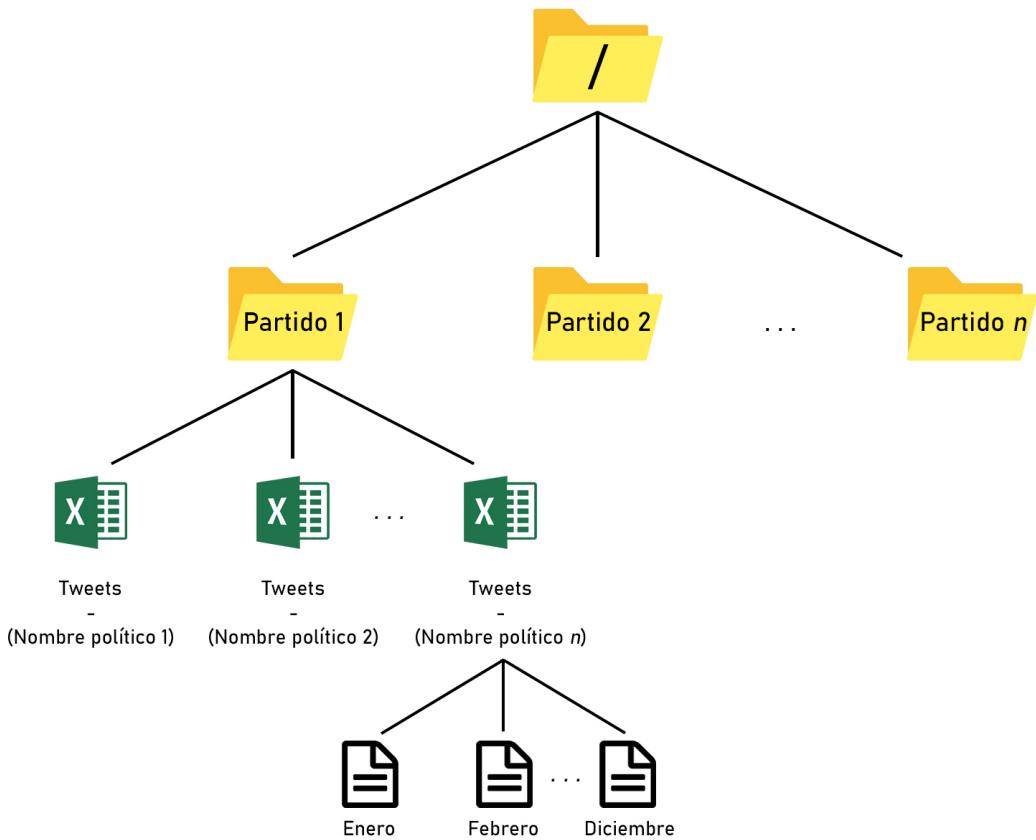


Figura 4.2: Estructura de ficheros para almacenar tweets descargados

Para cada tweet, haremos lo siguiente:

1. Comprobaremos que no sea anterior al año 2020.
2. Comprobaremos que esté escrito en castellano.
3. Comprobaremos que no sea un tweet que ya esté descargado.
4. Comprobaremos si es un *retweet* o un tweet propio.
5. Guardaremos el tweet junto a algunas de sus características (si es o no un *retweet*, el número de *likes* que tiene, el número de *retweets*, la fecha de publicación...) en la hoja Excel de su mes correspondiente, haciendo uso de la librería *OpenPyXL* [21] para el manejo de este tipo de archivos.

El código correspondiente a este proceso es el siguiente:

```

#lista => excel con datos de las cuentas
for politico in lista.iter_rows():

    rutaExcel = './' + politico[2].value + '/' + politico[0].value + '/tweets - ' + politico[0].value + '.xlsx'
    excel = load_workbook(filename=rutaExcel)

    for tweet in tweepy.Cursor(api.user_timeline,
                                id = politico[1].value, tweet_mode= 'extended').items():
        guardarTweet(tweet, excel)

    excel.save(rutaExcel)

def guardarTweet(tweet, excel):
    if (tweet.created_at.year < 2020):
        #Comprobamos que no sea anterior a 2020
        break

    if (tweet.lang != 'es'):
        #Comprobamos que sea en castellano
        continue

    hojaExcel = excel[str(tweet.created_at.month)]

    if ('retweeted_status' in tweet._json): #Es un rt
        text = tweet.retweeted_status.full_text
        rtText = 'Y'
    else:
        text = tweet.full_text
        rtText = 'N'

    if tweet.place is None: #No hay info de ubicacion
        city = ''
        coordinates = ''
    else:
        city = tweet.place.name
        coordinates = str(tweet.place.bounding_box.coordinates)

    hojaExcel.append([text, rtText, tweet.retweet_count, tweet.favorite_count, tweet.created_at.day, tweet.created_at.hour, tweet.created_at.minute, city, coordinates, tweet.id_str])

```

4.4. Preprocesamiento de texto

Antes de empezar con el procesamiento del lenguaje natural, es necesario preprocesar cada documento. Con esto conseguiremos que las mayúsculas, los acentos o símbolos de puntuación, carentes de valor semántico, no influyan en el posterior procesamiento del lenguaje natural; o diferenciar partes especiales del texto, como las URLs o los *hashtags* del resto del documento. Así, cada Tweet que descarguemos pasará por el siguiente preprocesamiento:

1. **Fecha e idioma.** Comprobamos que esté publicado en el año 2020, y que esté escrito en español. Esto lo hacemos usando los atributos *created_at* y *lang* que nos ofrece el objeto Tweet, respectivamente. Si no cumple una de estas dos condiciones, descartamos el Tweet.
2. **URLs.** Twitter tiene su propio servicio de acortamiento de enlaces [11], el cual se hace cargo de que cualquier URL que agreguemos a un Tweet tenga siempre el mismo formato, que es la url del servicio (<https://t.co/>) unido a 10 caracteres alfanuméricos. Aprovechando esto, podemos identificar fácilmente las URLs con la siguiente expresión regular:

$$\text{https : //t\\.co/}\w\{10\}$$

3. **Emojis.** Para diferenciar los emoticonos del resto del texto, usamos la macro *get_emoji_regexp* de la librería *emoji* de Python, que básicamente es una lista de cada código *unicode* de cada emoji [12]
4. **Signos de puntuación.** Eliminaremos todos los signos de puntuación, así como los saltos de línea, ya que carecen de significado.
5. **Menciones y hashtags.** Como vimos anteriormente, podemos mencionar a otro usuario en nuestro Tweet escribiendo una @ antes del nombre de la cuenta de dicho usuario. Además, un nombre de usuario en Twitter ha de cumplir las siguientes condiciones [13] :
 - a) La longitud máxima será de 15 caracteres (sin incluir la @).
 - b) El nombre deberá estar compuesto solo por caracteres alfanuméricos (sin acento gráfico), a excepción del guión bajo (_), que también está permitido.

Luego nuestra expresión regular para detectar menciones será la siguiente:

$$@[\w_]\{1,15\}$$

El caso de los *hashtags* es similar. La @ pasa a ser un #, y las condiciones ha cumplir pasan a ser las siguientes:

- a) La longitud máxima será de 100 caracteres (sin incluir el #).
- b) El nombre deberá estar compuesto solo por caracteres alfanuméricos, a excepción del guión bajo (_), que también está permitido.

Y usaremos la siguiente expresión regular para localizar *hashtags*:

[A-zA-ú0-9_]{1,100}

Nótese que el rango A-ú permite caracteres acentuados.

6. **Mayúsculas.** Todos los caracteres pasarán a ser minúsculas. De esta forma, las palabras que estaban en mayúsculas debido a los signos de puntuación dejarán de ser distinguibles respecto del resto. Hacemos este paso en último lugar para que no afecte a las menciones y a los *hashtags*, donde el uso de mayúsculas si puede ser relevante.

En este punto, ya tenemos el documento listo para aplicarle el procesamiento del lenguaje natural.

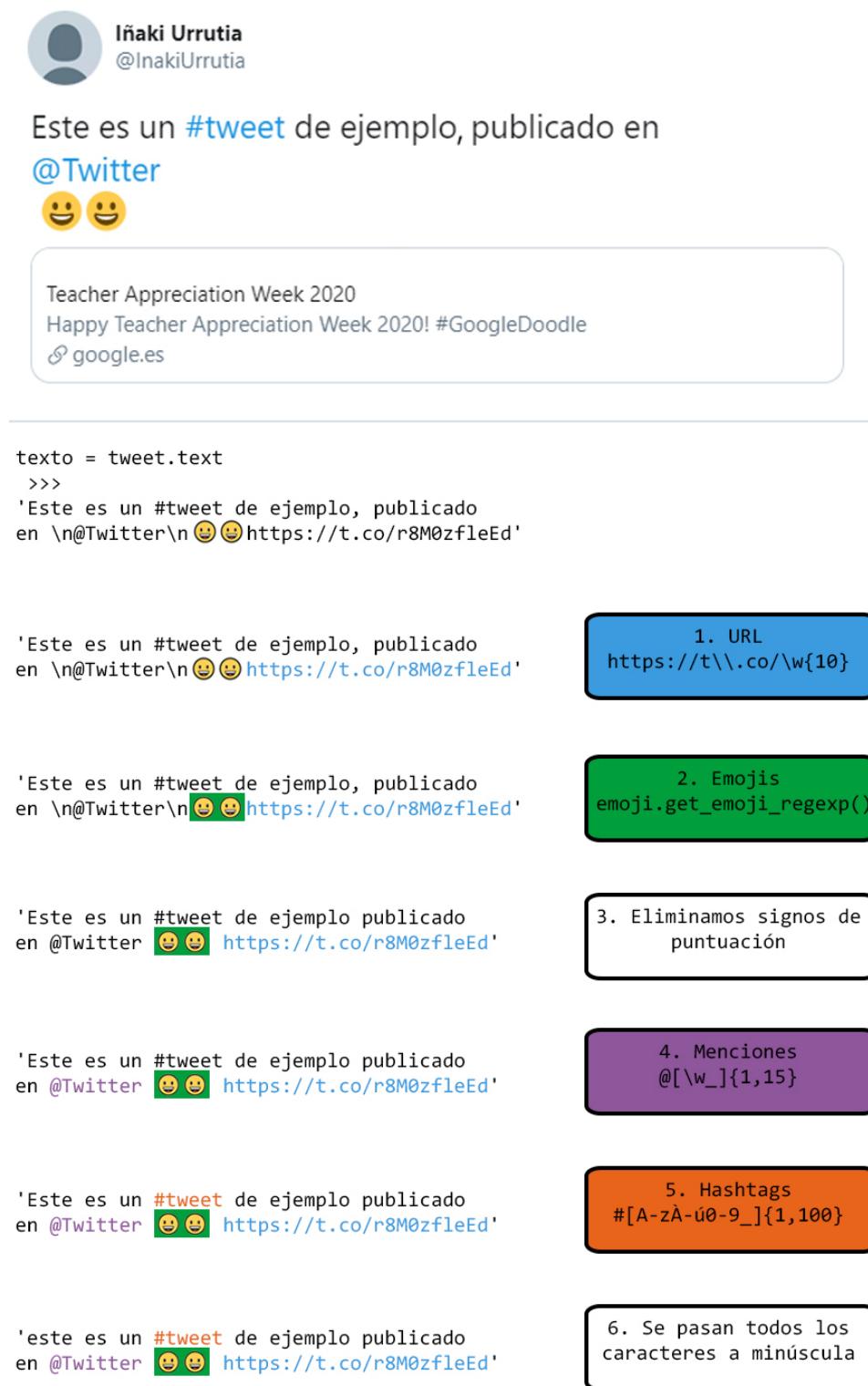


Figura 4.3: Diagrama del preprocesamiento de los Tweets descargados

4.4.1. Tokenización

En el caso de nuestro proyecto, gracias a que antes preprocesamos cada tweet, podemos hacer la *tokenización* de una manera muy sencilla:

1. Twitter sólo usa saltos de líneas y espacios blancos simples en sus mensajes, así que no tenemos que preocuparnos de caracteres como tabulaciones. Además, en el preprocesamiento ya eliminamos los saltos de línea, luego solo tenemos que encargarnos de los espacios. Para esto, debemos asegurarnos que entre palabra y palabra haya únicamente un sólo espacio. Esto podemos hacerlo sustituyendo todas las coincidencias que nos arroje la siguiente expresión regular:

“ {2,}”

por simples caracteres en blanco.

2. En este punto, ya podemos asegurar que todas las palabras están separadas por un único espacio en blanco, luego podemos *tokenizar* la cadena con la función `split()` de Python, que justamente divide una cadena en palabras, basando esta división en caracteres separadores simples.

4.4.2. Stopwords

Conformaremos el conjunto de *stop words* uniendo los conjuntos que nos proporcionan tanto la biblioteca nltk como spacy. Esto nos da como resultado un total de 704 palabras vacías, que eliminaremos de cada documento.

4.4.3. Stemming

Para realizar el *stemming* de los *tokens*, usaremos la implementación del *Snowball stemmer* de la librería nltk, ya que tiene una versión preparada para ser usada en documentos escritos en castellano.

4.5. Análisis de sentimientos

A día de hoy, no existen muchas librerías que ofrezcan modelos de análisis de sentimientos en texto, y la mayoría de las existentes son para trabajar con textos en inglés. Por este motivo, vamos a desarrollar un modelo propio que nos permita predecir si un documento (un tweet) transmite un mensaje positivo o negativo.

4.5.1. Preparación del *dataset*

Un elemento clave en la creación de cualquier modelo basado en Machine Learning es el dataset con el que vamos a entrenar al mismo. Es importante que los datos que lo conforman sean de la mayor calidad posible.

En nuestro caso, podemos obtener un buen dataset usando los documentos del TASS (Taller de Análisis Semántico en la SEPLN) [23]. Cada uno de estos documentos es un archivo XML en el que hay una serie de tweets etiquetados según su polaridad.

```
<tweets>
  <tweet>
    <tweetid>123456789</tweetid>
    <user>987654321</user>
    <content>Hoy me he despertado muy feliz.</content>
    <date>Tue May 26 22:29:21 +0000 2020</date>
    <lang>es</lang>
    <sentiment>
      <polarity><value>P</value></polarity>
    </sentiment>
  </tweet>
  <tweet>
    <tweetid>123456788</tweetid>
    <user>987654322</user>
    <content>Está siendo un mal día, estoy triste y enfadada</content>
    <date>Tue May 26 22:36:22 +0000 2020</date>
    <lang>es</lang>
    <sentiment>
      <polarity><value>N</value></polarity>
    </sentiment>
  </tweet>
```

Figura 4.4: Dos ejemplos de tweets con la estructura básica de los documentos del TASS

En total, tenemos 9 documentos diferentes, con pequeñas diferencias en los esquemas de cada uno de ellos, pero en general con una estructura muy parecida. Para procesar cada XML, haremos uso de *Pandas* [25], que es una librería para Python que cuenta con funcionalidades muy útiles a la hora del manejo de datos.

Por cada documento, obtendremos cada tweet y su polaridad asociada. Aquí se nos presenta un problema, y es que no todos los documentos tienen el mismo rango de polaridades. Para solucionar esto, haremos que la polaridad se transforme en una variable binaria (mensaje positivo o mensaje negativo), ya que todos los documentos manejan como mínimo este rango. Así, nos quedan un total de 47858 tweets, de los cuales 27065 son tweets con mensajes clasificados como positivos y 20793 son negativos.

A continuación, aplicaremos el preprocesamiento explicado anteriormente a cada uno de los tweets, de manera que lo que antes era una cadena de texto con el mensaje original de cada tweet pasará a ser un conjunto de *stems*. Posteriormente, creamos una matriz en la que representaremos cuantas veces aparece cada *stem* en cada documento, lo que más adelante nos valdrá para que nuestro modelo sea capaz de aprender qué palabras influyen en la positividad/negatividad de un mensaje.

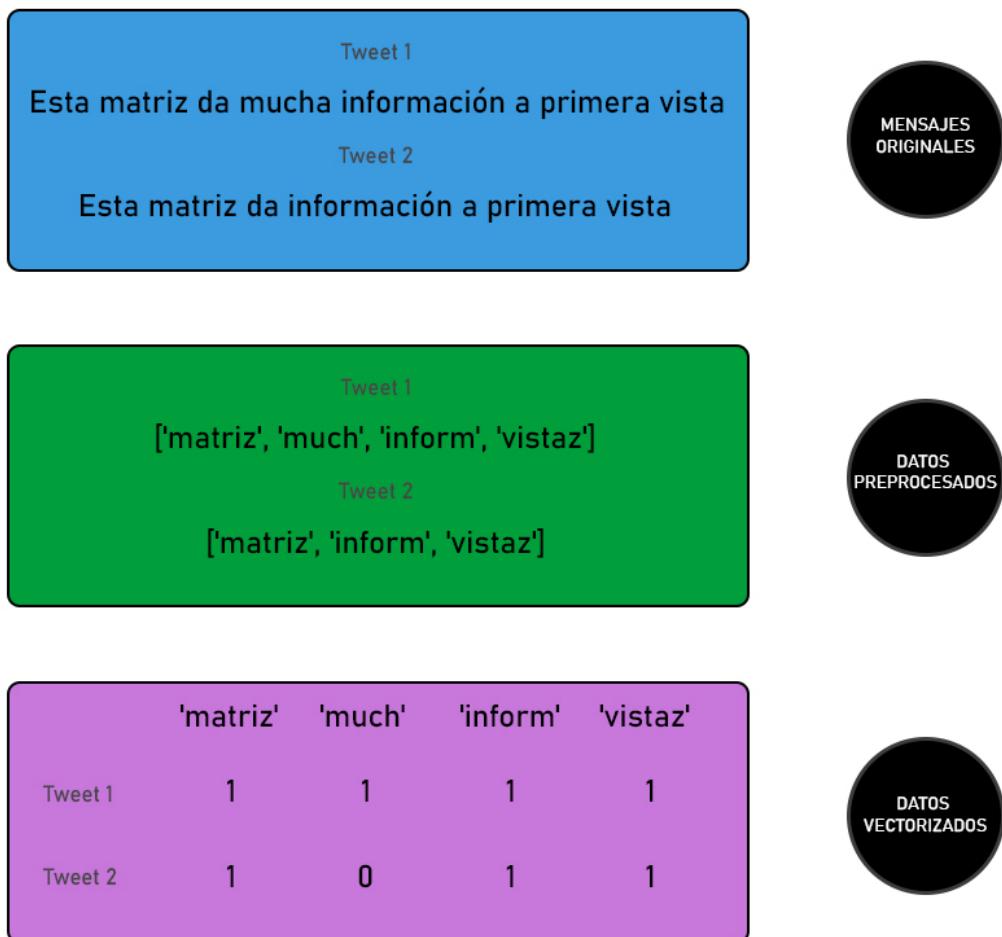


Figura 4.5: Ejemplo de la vectorización de los mensajes

4.5.2. Creación del modelo

En este punto, ya tenemos nuestro dataset preparado para crear y entrenar nuestro modelo de predicción. Para esta tarea nos vamos a apoyar en Scikit-learn [26], una librería enfocada al aprendizaje automático en Python.

Así, nuestro modelo se va a basar en los algoritmos SVC (Support Vector Clasification), que, en resumen, son una familia de algoritmos de aprendizaje supervisado en los que, dado un conjunto de puntos en los que cada punto puede pertenecer a una categoría A o a una categoría B , se construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una u otra categoría.

Ahora tenemos que buscar los mejores hiperparámetros para nuestro modelo. Estos hiperparámetros definen características como por ejemplo el número máximo de iteraciones en el entrenamiento, el umbral mínimo de veces que tiene que salir un *stem* para que sea considerado relevante, etc. Explorar todas estas características tiene un coste computacional elevado, ya que son muchas combinaciones posibles, y en este caso, cada combinación tarda algo más de un minuto, por lo que el proceso total tarda bastantes horas, usando todos los procesadores lógicos de la CPU al 100 %

```
[CV] cls_C=0.2, cls_max_iter=500, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1), score=0.879, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1), score=0.882, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1), score=0.888, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1), score=0.888, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1), score=0.756, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 2)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 1), score=0.879, total= 1.1min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 2)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1), score=0.889, total= 1.2min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 1), score=0.891, total= 1.2min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 2)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 2)
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=20, vect_ngram_range=(1, 2), score=0.891, total= 1.2min
[CV] cls_C=0.2, cls_max_iter=500, vect_max_df=0.5, vect_max_features=500, vect_min_df=10, vect_ngram_range=(1, 2), score=0.877, total= 1.2min
```

Figura 4.6: Output del algoritmo de entrenamiento mientras busca los mejores hiperparámetros

Al final del proceso, seleccionaremos la combinación de hiperparámetros que mejor puntuación nos haya dado. Para calcular esta puntuación usaremos una función de área bajo la curva, ya que es un sistema de evaluación que tiene en cuenta tanto los falsos positivos como los falsos negativos que pueden producirse en un clasificador binario, que es justamente lo que buscamos en nuestro caso.

Además, validaremos el modelo usando el método de validación cruzada con $k = 5$. Esto significa que, para evaluar cada modelo, dividiremos nuestro conjunto de datos en 5 bloques del mismo tamaño. De estos 5 bloques, usaremos 4 para entrenar el modelo, y usamos el modelo resultante para clasificar los datos del bloque restante. En la siguiente iteración, usaremos otra combinación de 4 bloques de datos para entrenar el modelo, y evaluaremos el bloque restante, y así sucesivamente hasta haber probado todas las combinaciones posibles.

De esta manera evitamos que un modelo sea elegido como el mejor sólo por la casualidad de dar un muy buen rendimiento con un conjunto de datos muy específico, ya que lo óptimo es que el rendimiento sea el mejor posible en todas las circunstancias.

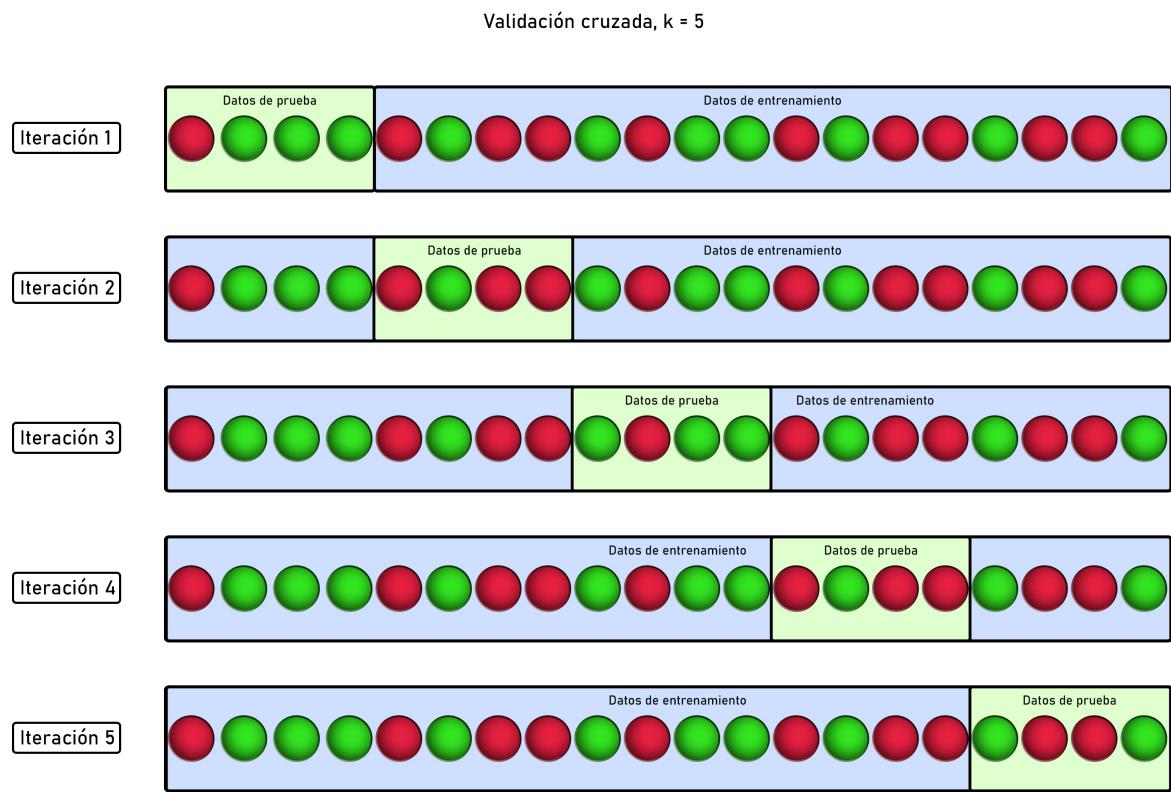


Figura 4.7: Validación cruzada con $K = 5$

4.6. Marco experimental

En esta sección se definen los recursos y herramientas usados para llevar a cabo el proyecto.

4.6.1. Recursos

Para la realización de este proyecto, se ha usado un equipo de sobremesa con las siguientes características:

- Placa base: ASRock B360 Pro4.
- Procesador: Intel Core i7 8700.
- Memoria RAM: 16GB 2667MHz DDR4.
- Tarjeta gráfica: NVIDIA GeForce GTX 1060 6GB.
- Almacenamiento: SSD 110GB.

4.7. Herramientas

Para la realización de este proyecto, se ha hecho uso de las siguientes herramientas:

- Editor de texto: Sublime Text.
- Jupyter Notebook.
- Compilación del documento de la memoria: Latex.
- Creación de figuras: Photoshop, Matplotlib, Inkscape.
- Dependencias adicionales en Python: Tweepy, Gensim, Matplotlib, WordCloud, Numpy, Nltk, Openpyxl, Pandas, Sklearn.

Capítulo 5

Resultados

5.1. Identificación de los principales actores políticos en Twitter

Cabe señalar que la situación política que se va a mostrar a lo largo de esta sección corresponde a abril de 2020, y podría sufrir algunos cambios en un futuro cercano.

5.1.1. Ministerios

Tras el acuerdo de coalición al que PSOE y Unidas Podemos llegaron en enero de 2020, Pedro Sánchez Pérez-Castejón fue investido presidente del Gobierno por el Congreso de los Diputados. Cinco días más tarde Sánchez anunció la composición de su gabinete, el cual quedó conformado por un total de 22 ministerios, con 11 ministras y 11 ministros.

Como podemos ver en la siguiente tabla, todos los ministerios del actual gobierno tienen cuenta oficial en Twitter, todas ellas activas, y, todos los miembros del Consejo de Ministros, a excepción de Margarita Robles y Fernando Grande-Marlaska, también tienen cuenta oficial en Twitter.

NOMBRE	@	PARTIDO
Presidencia del gobierno	desdelamoncloa	PSOE
Pedro Sánchez	sanchezcastejon	PSOE
Ministerio de Agricultura, Pesca y Alimentación	mapagob	PSOE
Luis Planas Puchades	LuisPlanas	PSOE
Ministerio de Asuntos Económicos y Transformación Digital	_minecogob	PSOE
Nadia Calviño	NadiaCalvino	PSOE
Ministerio de Asuntos Exteriores, Unión Europea y Cooperación	MAECgob	PSOE
Arancha González	AranchaGlezLaya	PSOE
Ministerio de Ciencia e Innovación	CienciaGob	PSOE
Pedro Duque	astro_duque	PSOE
Ministerio de Consumo	consumogob	UP
Alberto Carlos Garzón Espinosa	agarzon	UP
Ministerio de Cultura y Deporte	culturagob	PSOE
José Manuel Rodríguez Uribes	jmrdezuribes	PSOE
Ministerio de Defensa	Defensagob	PSOE
Margarita Robles		PSOE
Ministerio de Derechos Sociales y Agenda 2030	VSocialGob	UP
Pablo Iglesias	PabloIglesias	UP
Ministerio de Educación y Formación Profesional	educaciongob	PSOE
Isabel Celaá	CelaaIsabel	PSOE
Ministerio de Hacienda	Haciendagob	PSOE
María Jesús Montero	mjmonteroc	PSOE
Ministerio de Igualdad	IgualdadGob	UP
Irene Montero	IreneMontero	UP
Ministerio de Inclusión, Seguridad Social y Migraciones	inclusiongob	PSOE
José Luis Escrivá	joseluisescriva	PSOE
Ministerio de Industria, Comercio y Turismo	mincoturgob	PSOE
Reyes Maroto	MarotoReyes	PSOE
Ministerio de Justicia	justiciagob	PSOE
Juan Carlos Campo	Jccampm	PSOE
Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática	M_Presidencia	PSOE
Carmen Calvo	carmencalvo_	PSOE
Ministerio de Política Territorial y Función Pública	territorialgob	PSOE
Carolina Darias	CarolinaDarias	PSOE
Ministerio de Sanidad	sanidadgob	PSOE

NOMBRE	@	PARTIDO
Salvador Illa	salvadorilla	PSOE
Ministerio de Trabajo y Economía Social	empleogob	UP
Yolanda Díaz Pérez	Yolanda_Diaz_	UP
Ministerio de Transportes, Movilidad y Agenda Urbana	mitmagob	PSOE
José Luis Ábalos	abalosmeco	PSOE
Ministerio de Universidades	UniversidadGob	UP
Manuel Castells	manuelcastells	UP
Ministerio del Interior	interiorgob	PSOE
Fernando Grande-Marlaska		PSOE
Ministerio para la Transición Ecológica y el Reto Demográfico	mitecogob	PSOE
Teresa Ribera	Teresaribera	PSOE

Cuadro 5.1: Ministerios de España, ministros/as y sus respectivas cuentas en Twitter

5.1.2. Gobiernos autonómicos

A día de hoy, España se compone de 17 comunidades autónomas, todas y cada una de ellas con un gobierno propio. En la siguiente tabla podemos ver todos los gobiernos autonómicos y a sus respectivos presidentes, junto a sus cuentas oficiales en Twitter y partido político.

NOMBRE	@	PARTIDO
Junta de Andalucía	AndaluciaJunta	PP
Juan Manuel Moreno	JuanMa_Moreno	PP
Gobierno de Aragón	GobAragon	PSOE
Javier Lambán	JLambanM	PSOE
Gobierno de Asturias	GobAsturias	PSOE
Adrián Barbón	AdrianBarbon	PSOE
Govern Illes Balears	goib	PSOE
Francina Armengol	F_Armengol	PSOE
Gobierno de Canarias	PresiCan	PSOE
Ángel Víctor Torres	avtorresp	PSOE
Gobierno Cantabria	cantabriaes	PRC
Miguel Ángel Revilla	RevillaMiguelA	PRC
Junta de Castilla y León	j cyl	PP
Alfonso Fernández Mañueco	alferma1	PP
Gobierno de Castilla-La Mancha	gobjccm	PSOE
Pablo Bellido	PabloBellido_Az	PSOE

NOMBRE	@	PARTIDO
Generalitat de Catalunya	govern	JxCat
Quim Torra i Pla	QuimTorraiPla	JxCat
Generalitat Valenciana	generalitat	PSOE
Ximo Puig	ximopuig	PSOE
Junta de Extremadura	Junta_Ex	PSOE
Guillermo Fernández Vara	GFGVara	PSOE
Xunta de Galicia	Xunta_c	PP
Alberto Núñez Feijóo	FeijooGalicia	PP
Comunidad de Madrid	ComunidadMadrid	PP
Isabel Díaz Ayuso	IdiazAyuso	PP
Gobierno de la Región de Murcia	regiondemurcia	PP
Fernando López Miras	LopezMirasF	PP
Gobierno de Navarra	gob_na	PSOE
María Chivite Navascués	mavichina	PSOE
Gobierno de Euskadi	Irekia	PNV
Iñigo Urkullu	iurkullu	PNV
Gobierno de La Rioja	lariojaorg	PSOE
Concha Andreu	ConchaAndreu	PSOE

Cuadro 5.2: Gobiernos autonómicos, presidentes autonómicos y sus respectivas cuentas en Twitter

Podemos ver que en esta ocasión todos los miembros de este grupo tienen cuenta oficial de Twitter. Sin embargo, tenemos otro problema: las cuentas de comunidades autónomas con idiomas cooficiales tweetean o bien sólo en dicho idioma, o bien en español y en el idioma cooficial indistintamente. Este problema será estudiado más tarde, en la fase de procesamiento de los tweets descargados.

5.1.3. Miembros del Congreso de los Diputados

Como dijimos anteriormente, también vamos a usar los tweets de las diez personas que encabezaban las listas electorales de los partidos con diez diputados o más en el Congreso. Estos partidos son el Partido Socialista Obrero Español (PSOE), con 120 escaños; el Partido Popular (PP), con 89 escaños; VOX, con 52 escaños; Unidas Podemos (UP) con 35 escaños; Esquerra Republicana de Catalunya (ERC), con 13 escaños y Ciudadanos (Cs) con 10 escaños.

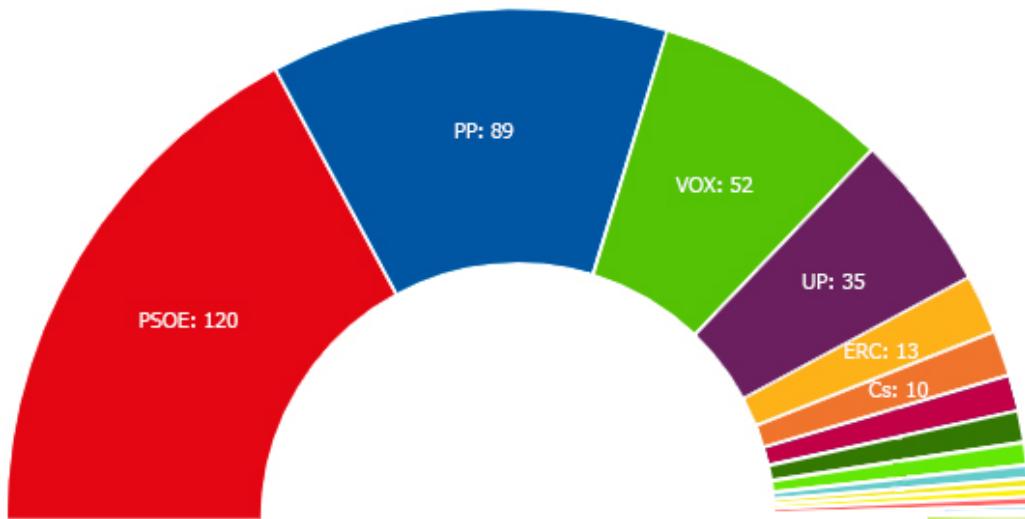


Figura 5.1: Composición actual del Congreso de los Diputados

Utilizando el criterio anteriormente expuesto y esta información, las cuentas seleccionadas de este grupo son las siguientes:

NOMBRE	@	PARTIDO
José Manuel Franco	conJoseMFranco	PSOE
Rafael Simancas	SimancasRafael	PSOE
Beatriz Corredor	BeatrizCorredor	PSOE
Zaida Cantera	ZaidaCantera	PSOE
Daniel Viondi	Viondi	PSOE
Pablo Casado	pablocasado_	PP
Ana Pastor Julián	anapastorjulian	PP
Isabel García Tejerina	tejerinapp	PP
Elvira Rodríguez	erodriguez_2019	PP
Edurne Uriarte	EdurneUriarte	PP
Ana Beltrán	abeltran_ana	PP
Antonio González Terol	Aglezterol	PP
Pilar Marcos Domínguez	pilarmarcosd	PP
María del Carmen Navarro Lacoba	CnLacoba	PP
César Sánchez Pérez	sanchezcesar	PP

NOMBRE	@	PARTIDO
Santiago Abascal	Santi_ABASCAL	VOX
Javier Ortega Smith	Ortega_Smith	VOX
Iván Espinosa de los Monteros	ivanedlm	VOX
María de la Cabeza Ruíz Solás	RuizSolas	VOX
Carla Toscano de Balbín	eledhmel	VOX
Juan Luis Steegmann Olmedillas	jlsteeg	VOX
Mireia Borrás Pabón	_mireiaborras	VOX
Rafa Lomana	RafaLomana	VOX
Manuel Mestre Barea	mestremanuel	VOX
Rocío De Meer	MeerRocio	VOX
Enrique Santiago Romero	ensanro	UP
Gloria Elizo	GloriaElizo	UP
Rafael Mayoral	MayoralRafa	UP
Txema Guijarro García	TxemaGuijarro	UP
Juan López de Uralde	juralde	UP
Sofía Castañón	SofCastanon	UP
Antonia Jover Diaz	Antonia_jover_	UP
Lucía Miriam Muñoz Dalda	luciadalda	UP
Jaume Asens	Jaumeasens	UP
Aina Vidal	AinaVS	UP
Gabriel Rufián	gabrielrufian	ERC
Inés Arrimadas	InesArrimadas	Cs
Marcos de Quinto	MarcosdeQuinto	Cs
Sara Giménez	SaraGimnez	Cs
María Muñoz Vidal	mariadelamiel	Cs
Marta Martín Llaguno	martamartirio	Cs

Cuadro 5.3: Diputados seleccionados para el estudio

Como se puede ver en la tabla, hay algunos partidos que no llegan a los diez miembros. Esto se debe a las siguientes razones:

- PSOE: Muchos de los miembros ya están seleccionados por formar parte del gobierno.
- ERC: A excepción de Gabriel Rufián, el resto de diputados de ERC tweetean exclusivamente en catalán.
- Cs: Solo cinco de los diputados de Ciudadanos tiene cuenta oficial de Twitter.

5.2. Conteo general

En esta sección vamos a desarrollar una visión global de nuestro conjunto de datos, con el fin de extraer información general antes de entrar a analizar rangos de fechas más específicos.

Un buen comienzo es ver como ha ido evolucionando la cantidad media de tweets que las cuentas de las que hacemos el seguimiento publicaban conforme avanzaba el tiempo, tal y como muestra la Figura 5.2.

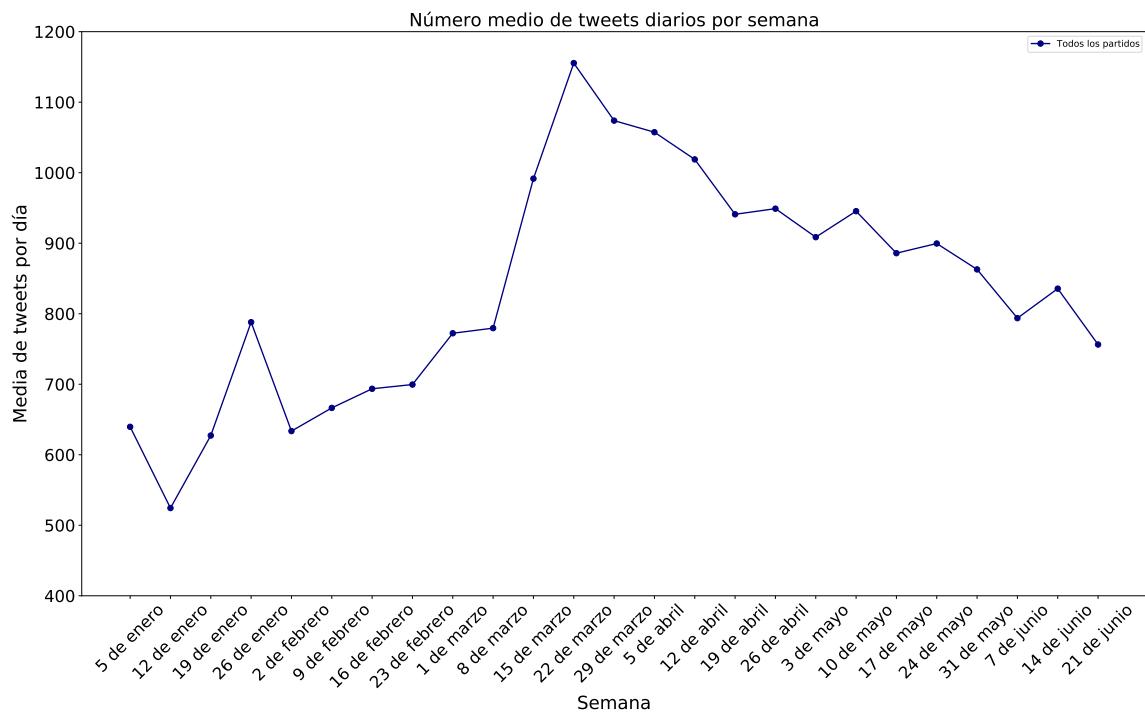


Figura 5.2: Media de tweets diarios, agrupado por semanas

Como podemos apreciar, se nota un claro incremento a partir de la semana del 8 de marzo, semana llena de eventos clave como la declaración del COVID-19 como pandemia por parte de las OMS el día 11, o la declaración de la cuarentena en España el 14, entre otros muchos eventos que desarrollaremos en profundidad más adelante. Concretamente, antes de esta semana la media de tweets publicados al día era de 672, y a partir de esa fecha y hasta la primera semana de junio aumentará un 42,8 % hasta alcanzar la media de 954 tweets por día.

Este incremento que hemos mencionado puede explicarse aún mejor analizando también la Figura 5.3. La gráfica refleja el número medio de tweets diarios que contienen la palabra “coronavirus”, “covid” o “pandemia”. En este caso, el incremento es aún más brusco: pasa de no llegar a los 100 tweets diarios a rozar los 500, lo cual evidencia como la crisis del coronavirus disparó la actividad de los políticos en Twitter y se convirtió indiscutiblemente en el principal tema de sus mensajes, como se evidencia en la Figura 5.4

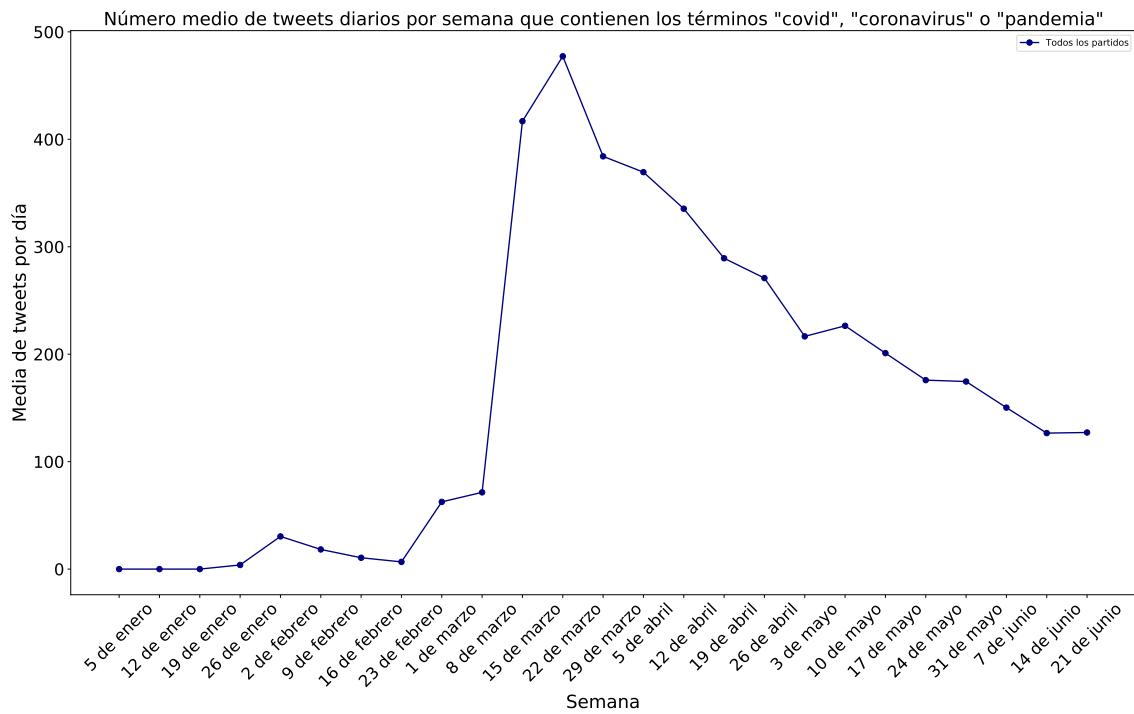


Figura 5.3: Media de tweets diarios que hacen referencia al COVID-19, agrupado por semanas

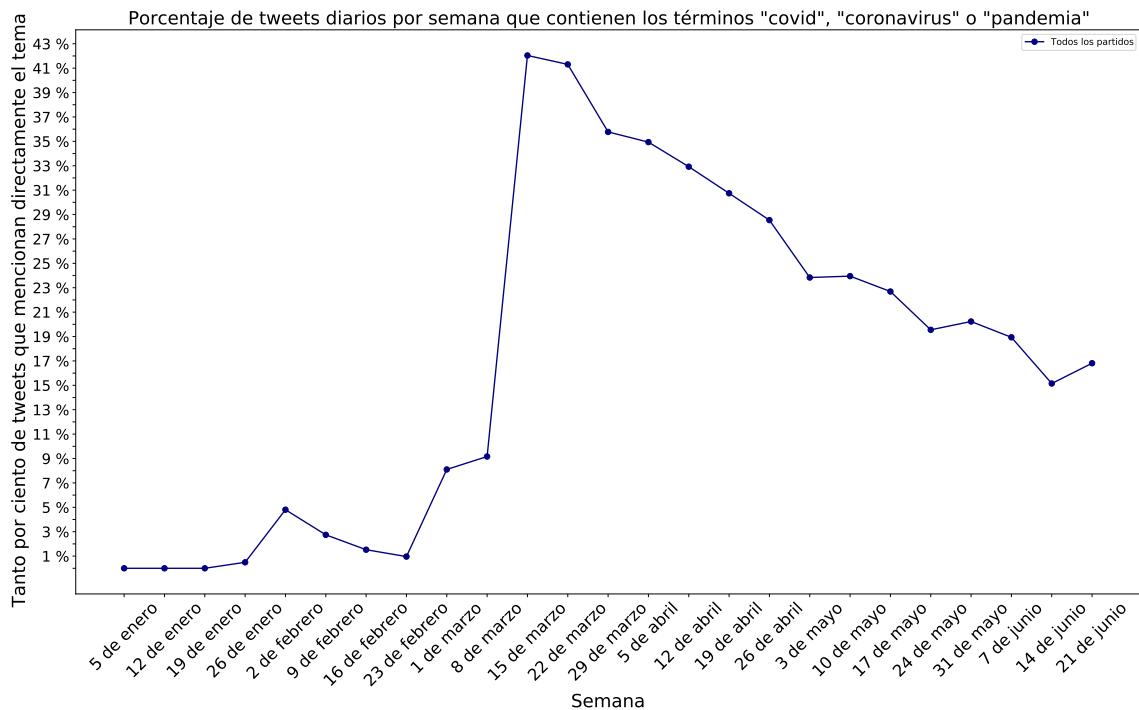


Figura 5.4: Porcentaje de tweets diarios que hacen referencia al COVID-19, agrupado por semanas

Otro análisis que merece la pena hacer es el de uso de hashtags conforme han ido pasando las semanas, ya que nos da de un vistazo información global de cómo han ido evolucionando las temáticas. Para ello, vamos a ver y comentar brevemente los diez hashtags más usados en cada mes.

En enero, como podemos ver en la Figura 5.5 que no hay ningún hashtag entre los diez primeros que haga referencia al coronavirus, sino que se refieren a temas tan ajenos a la futura pandemia como a la Feria Internacional de Turismo en Madrid (#fitur2020) o a la investidura del gobierno (#unspiparaavanzar , #sesióndeinvestidura).

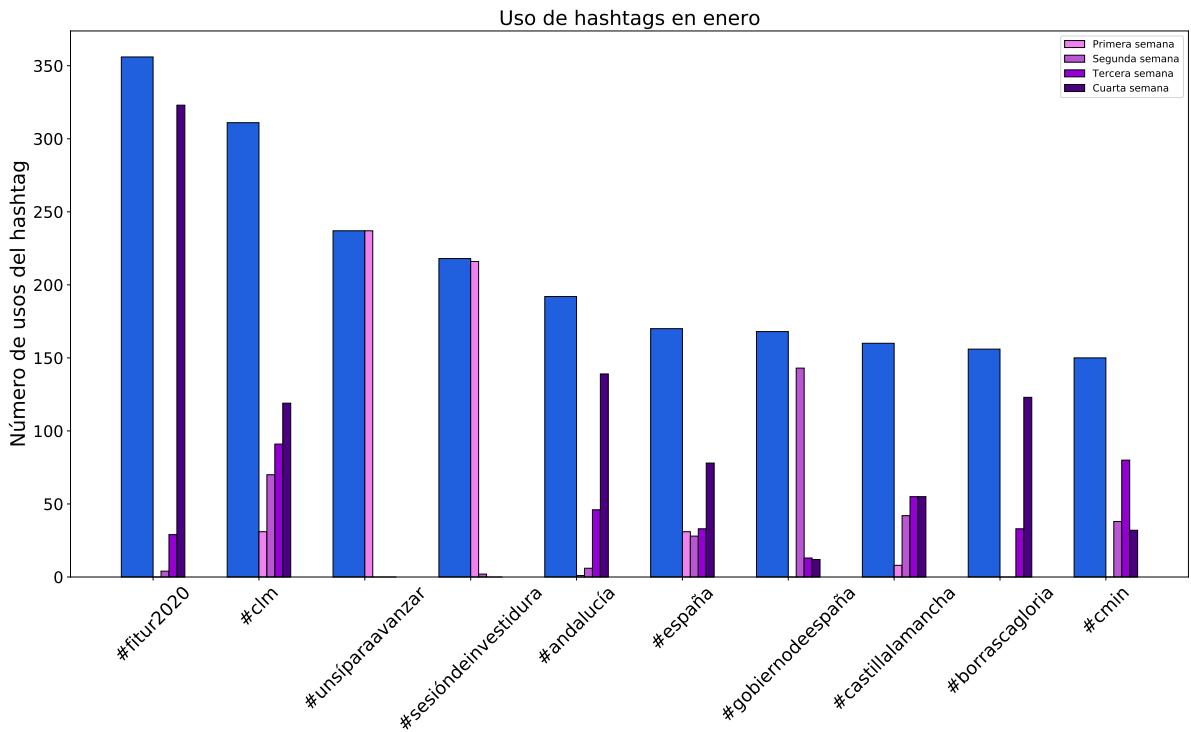


Figura 5.5: Hashtags enero

Es en febrero cuando el primer hashtag relacionado con la pandemia se hace importante, como refleja la Figura 5.6, donde el hashtag más usado es #coronavirus, con un total de 505 mensajes. Otros hashtags importantes en febrero fueron los relacionados con Andalucía (#andalucía, #28f40años, #28f), ya que el 28 de este mes fue el 40 aniversario de la celebración del referéndum por el cual Andalucía pasó a ser una comunidad autónoma.

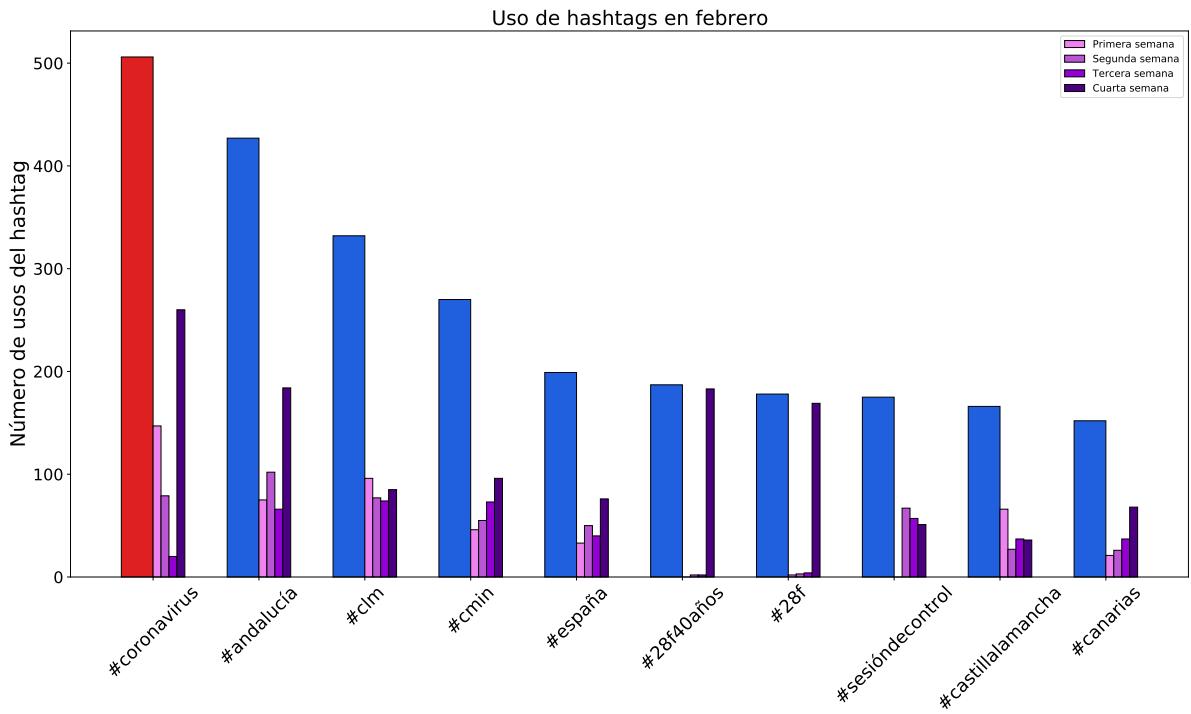


Figura 5.6: Hashtags febrero

A partir de marzo es cuando la crisis del COVID-19 empieza a ser el tema central en el día a día, tal y como tal se refleja en la Figura 5.7. Es destacable que de los diez hashtags más usados de este mes, 4 hablen directamente del coronavirus (#covid19, #coronavirus, #covid_19, #covid) y otros 2 lo hagan del confinamiento y de la lucha contra la pandemia (#estevirusloparamosunidos, #quédateencasa).

Además, no solo hay más hashtags relacionados con la pandemia, sino que también podemos ver como el número de mensajes de cada uno de los hashtags es muy superior al de meses anteriores. Tanto es así que en enero y febrero los hashtags más usados entre los políticos de nuestro estudio solo alcanzaban los 353 y 505 usos respectivamente, mientras que en el caso de marzo esta cifra se dispara hasta los 4711 mensajes.

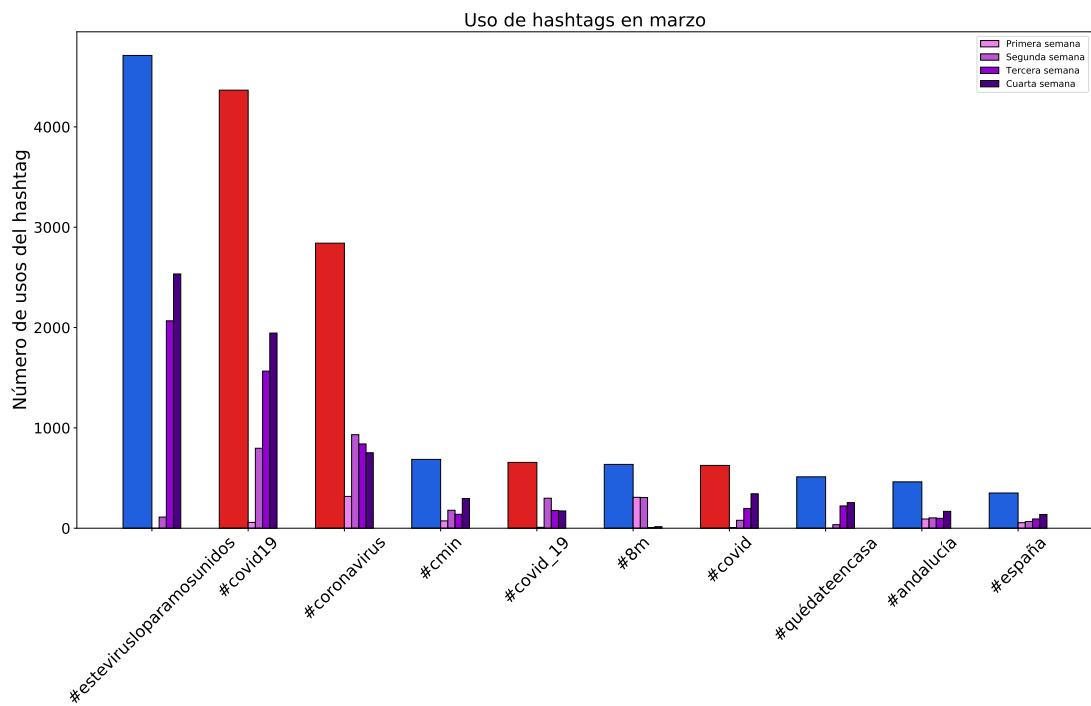


Figura 5.7: Hashtags marzo

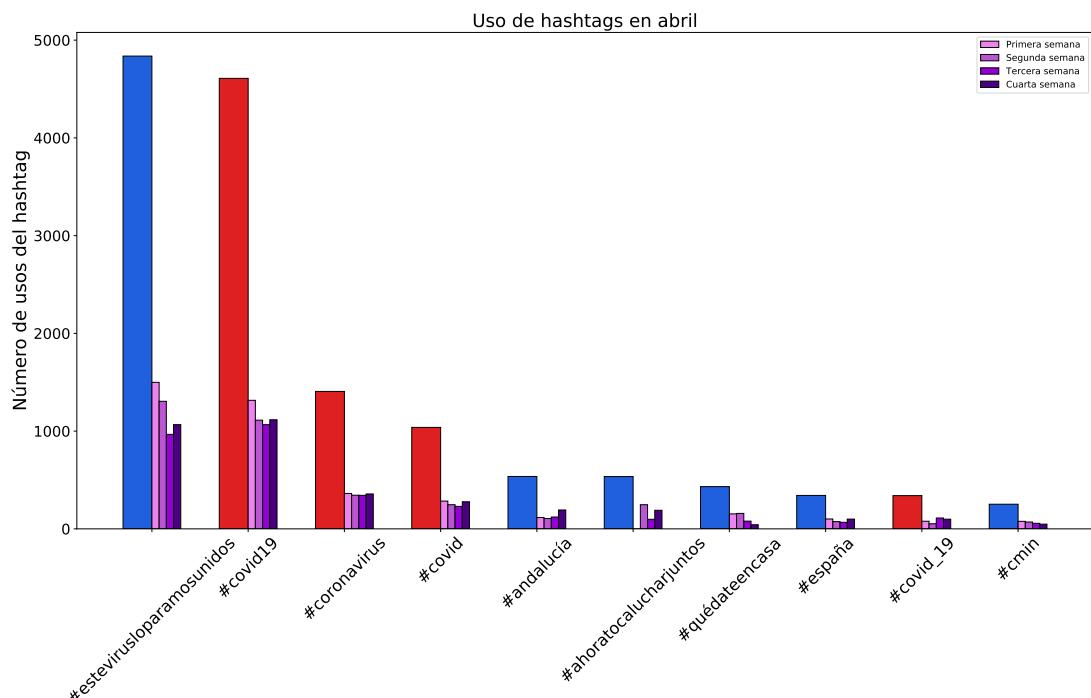


Figura 5.8: Hashtags abril

Abril (Figura 5.8) sigue la misma tónica que marzo, tanto en la temática de los hashtags como en el número de tweets que los usan.

En el caso de mayo (Figura 5.9), vemos como en general, los hashtags dominantes son los mismos que en los dos meses anteriores, pero notamos algunas diferencias, como una bajada importante en cuanto al número de mensajes total, y la aparición de los primeros hashtags que ya hablan de la desescalada (`#desescalada`, `#fase1`).

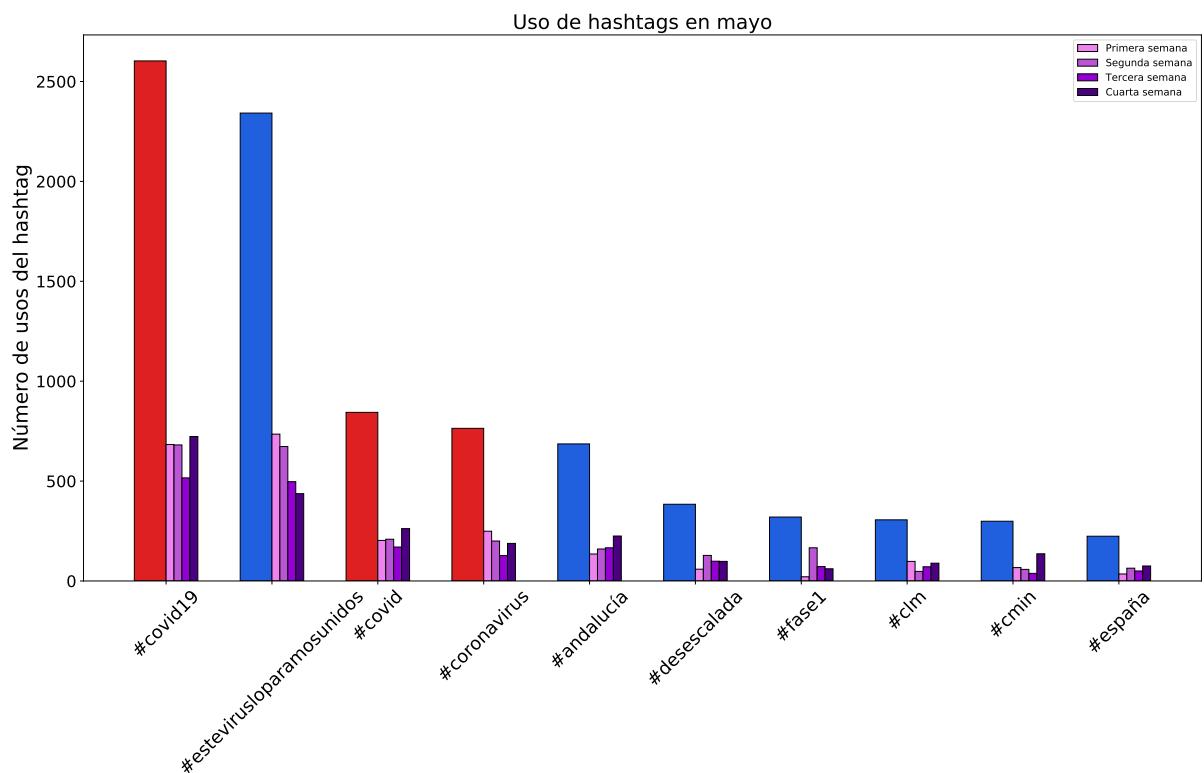


Figura 5.9: Hashtags mayo

5.3. Resultados por períodos temporales

Con el fin de analizar tanto el cambio de temáticas como el grado de positividad de los tweets según cada momento de la crisis del COVID-19, vamos a dividir todo este período temporal en los siguientes intervalos:

1. Desde el 1 de enero hasta el 31 de enero.
2. Desde el 1 de febrero hasta el 16 de febrero.
3. Desde el 17 de febrero hasta el 8 de marzo.
4. Desde el 9 de marzo hasta el 15 de marzo.
5. Desde el 16 de marzo hasta el 26 de marzo.
6. Desde el 27 de marzo hasta el 5 de abril.
7. Desde el 6 de abril hasta el 22 de abril.
8. Desde el 23 de abril hasta el 10 de mayo.
9. Desde el 11 de mayo hasta el 26 de mayo
10. Desde el 27 de mayo hasta el 8 de junio.
11. Desde el 9 de junio hasta el 21 de junio.

De esta manera, para cada intervalo, aplicaremos tanto el modelo LDA como el de predicción de sentimientos, generalmente en los mensajes de los 4 partidos políticos con más peso y representación parlamentaria.

Con el fin de hacer los resultados más claros, organizaremos las temáticas que nos devuelvan en un “nube de palabras”, que es una representación que nos permitirá, de un vistazo, saber qué palabras eran las más relevantes en según qué temática.

Una vez tengamos las temáticas más relevantes identificadas gracias al LDA, procesaremos los tweets que participen en cada una de ellas con nuestro modelo de análisis de sentimiento, con la meta de saber cuál era el grado de positividad/negatividad de cada uno de los principales partidos políticos respecto a los principales temas del momento.

5.4. Periodo 1 (1 de enero - 31 de enero)

En el caso de enero, no hacemos ninguna división temporal, ya que aún se veía al coronavirus como algo muy remoto y lejano. Aún así, es interesante analizar cuales fueron los temas más relevantes por aquel entonces.

Como hemos mencionado anteriormente, en el mes de enero el COVID-19 no se acercaba a ser un tema de actualidad política. Tanto es así que es a partir del día 22 cuando aparecen los primeros tweets mencionando al virus, informando sobre él y hablando de la situación de los españoles en Wuhan. En total, solo 126 de los 20391 (0,61 %) tweets de los que componen nuestro corpus en enero hablan del coronavirus. Es destacable que de esos 126 tweets, 112 (88,89 %) son de cuentas de miembros del PSOE.

Análisis de temáticas (LDA)

Como dijimos anteriormente, para tener una visión más específica sobre qué temáticas eran clave según el espectro político, vamos a dividir el corpus en los cuatro partidos con más peso en el parlamento, y para cada uno de ellos aplicaremos LDA, empezando por el PSOE:

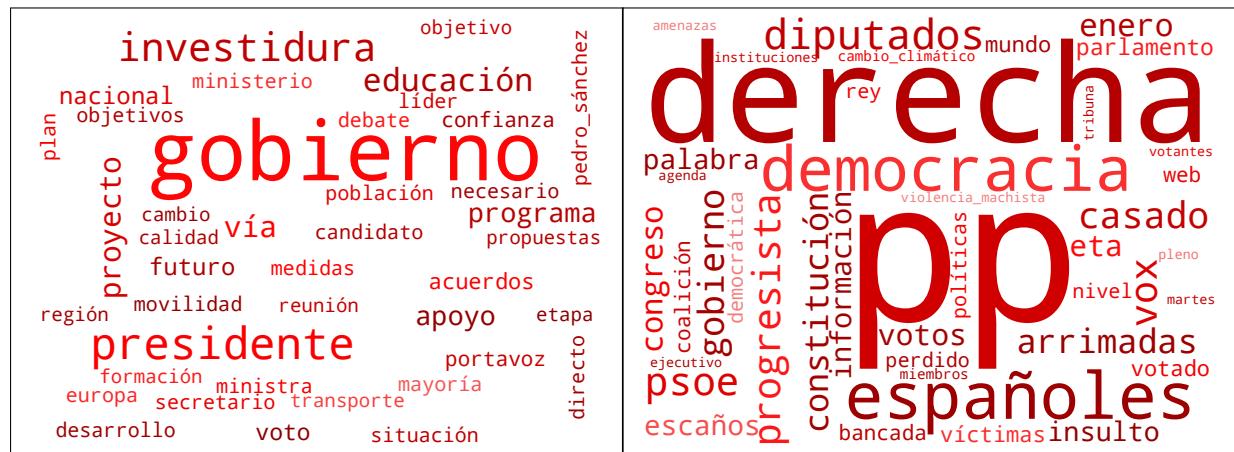


Figura 5.10: Principales temáticas del PSOE en el primer periodo

Lo que podemos ver en los dos grupos de palabras de la Figura 5.10 es una representación de las dos temáticas principales, determinadas por el modelo LDA.

Hemos de recordar que estamos en enero, mes en el que se formó el actual Gobierno de España, así que no es de extrañar que la investidura fuera el principal tema del que hablaron los socialistas durante el mes de enero, como podemos ver reflejado en la temática de la izquierda, con palabras como *investidura, presidente, gobierno*, etc.

En la otra temática se puede leer que en líneas generales el PSOE hacía mucho hincapié en mencionar a la oposición, hablando de derecha y ultraderecha, lo cual es concuerda con la época de crispación política que se vivía por la propia investidura.

En el caso del PP, como podemos ver en la Figura 5.11, las temáticas siguen la misma línea de hablar sobre la investidura. En el caso de la primera temática, podemos apreciar como el mensaje del partido relacionaba fuertemente la investidura de Pedro Sánchez con el independentismo catalán (podemos incluso ver la importancia de nombres propios como Torra o Junqueras en el discurso de los populares).

La segunda temática se centra algo más en lo que sería el gobierno de coalición, y otra vez vemos como relaciona al gobierno de Sánchez con el independentismo, esta vez sin centrarse tanto en el catalán sino que también habla del vasco (aparecen términos como ERC, Bildu, independentistas, etc.).



Figura 5.11: Principales temáticas del PP en el primer periodo

Por su parte, las temáticas más destacadas de VOX son similares a las del PP pero bastante más directas en las formas, tal y como muestra la Figura 5.12.



Figura 5.12: Principales temáticas de VOX en el primer periodo

La primera temática vuelve a incidir en la relación del gobierno (y en especial de la Figura de Pedro Sánchez) con el independentismo catalán, aunque usando formas más agresivas, con términos como *golpistas*, *tragedia*, *separatistas*, etc. En la segunda temática, vemos que VOX hizo mucha alusión directa al PSOE, relacionandolo con términos como, de nuevo, *golpismo*, u otros como *antiespañol*, *enemigos*, etc.

En cuanto a Unidas Podemos, podemos ver en la Figura 5.13 que el mensaje gira en torno al gobierno de coalición que formarían con PSOE, y algunas medidas que querrían llevar al mismo, tal y como pueden reflejar términos como *salario mínimo, pensiones, emergencia climática*. En el caso del segundo grupo de palabras, se aprecia que hay una mezcla de diferentes puntos que forman parte del discurso político del partido, como vuelve a ser el tema del medio ambiente (*contaminación, cambio climático*), el feminismo (*igualdad, mujeres*), etc.



Figura 5.13: Principales temáticas de UP en el primer periodo

Análisis de polaridad

Como hemos podido ver en los resultados que hemos obtenido en el LDA, el tema principal y compartido por todos los partidos políticos en enero fue la investidura de Pedro Sánchez y la formación del gobierno de coalición PSOE-UP. Por ello, es interesante analizar el grado de positividad con el que se abordaba el tema desde los principales puntos de vista políticos.

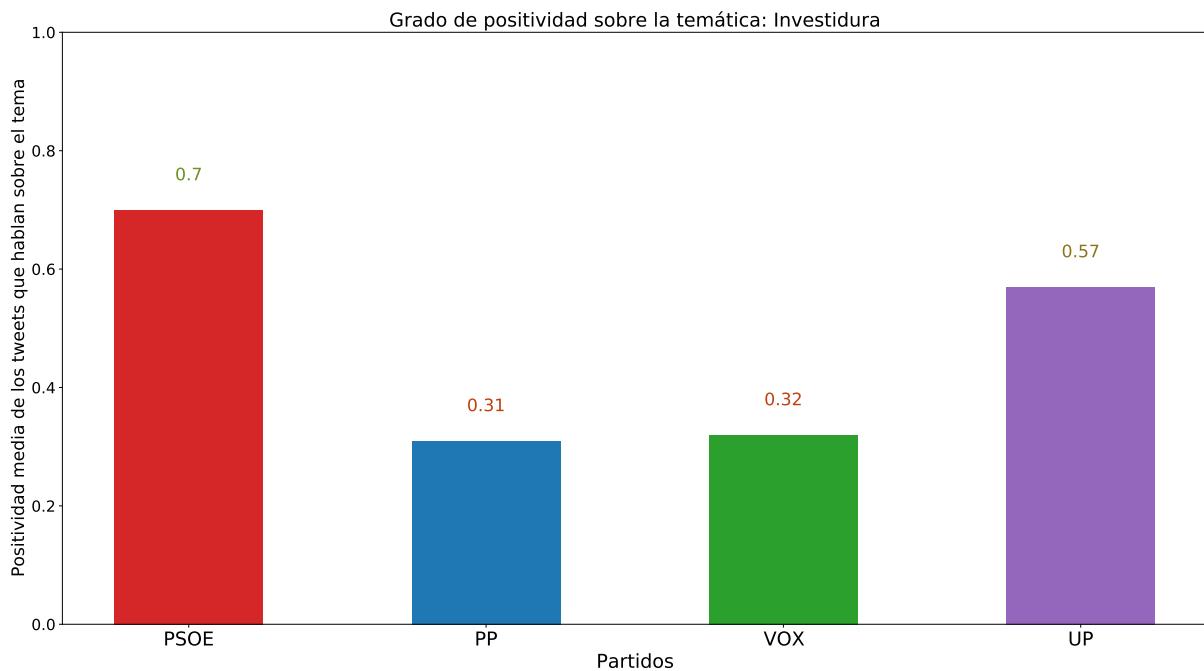


Figura 5.14: Análisis de sentimientos sobre la investidura de enero

Para ello, apoyándonos en los resultados del LDA, seleccionamos el conjunto de tweets que habla del tema dicho anteriormente, y aplicamos nuestro modelo de análisis de sentimientos.

Los resultados son los mostrados en la Figura 5.14, donde el 0 significa negatividad máxima y el 1 el máximo grado de positividad.

Como era de esperar, el PSOE, principal interesado en la formación de gobierno, es el que muestra una media de mensajes más positiva. UP, por su parte, muestra una tendencia también positiva, aunque algo menos, lo que podría deberse a las difíciles negociaciones que mantuvieron con el partido socialista para alcanzar un acuerdo. PP y VOX, como principales opositores a dicho pacto, reflejan una polaridad claramente negativa hacia el pacto de investidura.

5.5. Periodo 2 (1 de febrero - 16 de febrero)

Nos situamos ahora en la primera mitad de febrero, mes en el que la crisis provocada por el COVID-19 en España seguía pareciendo lejana, pero donde ya era un tema más importante. De hecho, este período coincide con los dos primeros casos de contagio en España, un turista alemán en La Gomera (31 de enero), y un británico en Palma de Mallorca (10 de febrero).

Estos hechos se reflejan en el aumento de tweets hablando sobre el virus, que pasan de ser un 0,61% en enero a un 3,67% en este período. Aún así, el coronavirus sigue sin ser un tema con gran presencia en el discurso político.

Análisis de temáticas (LDA)

Por estas fechas, el gobierno de coalición de PSOE-UP lleva ya cerca de un mes en funcionamiento. Esto se refleja en las nuevas temáticas, donde el tema más relevante en enero, la investidura, ha desaparecido por completo.

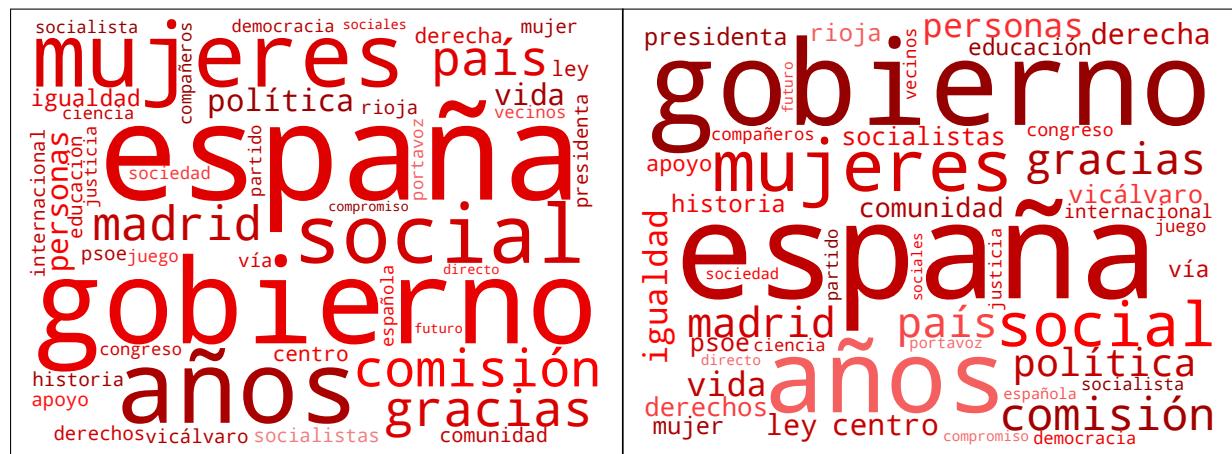


Figura 5.15: Principales temáticas del PSOE en el segundo periodo

Por su parte, el mensaje del PSOE (Figura 5.15) muestra la línea política en la que se basa su idea de gobierno, con términos como *reto demográfico*, *sanidad*, *transición ecológica*, y destacando especialmente el movimiento feminista.



Figura 5.16: Principales temáticas del PP en el segundo periodo

En el caso del PP (Figura 5.16), podemos observar que en su discurso estuvo muy presente la polémica que sufrió el actual ministro de Transportes, Movilidad y Agenda Urbana, José Luis Ábalos, por su reunión con la vicepresidenta de Venezuela, Delcy Rodríguez, persona a la que la portavoz del Partido Popular, Cayetana Álvarez de Toledo, llegó a calificar de “torturadora”.

Otro tema recurrente en los populares fue el de relacionar al gobierno de Sánchez con el independentismo catalán, tema que se repite en el caso de VOX (Figura 5.17). Otro asunto que estuvo muy presente en el mensaje de VOX en esta época fue el de las denuncias falsas, que según ellos, promueve el feminismo.



Figura 5.17: Principales temáticas de VOX en el segundo periodo



Figura 5.18: Principales temáticas de UP en el segundo periodo

El partido liderado por Pablo Iglesias (Figura 5.18), por su parte, tuvo muy presente en su mensaje a la oposición, y, en otra línea más parecida a la de su socio de gobierno, habló también de temas que definen su línea política.

Análisis de polaridad

Hemos visto que un factor común en el mensaje de los cuatro partidos en este período ha sido la mención directa a los partidos de ideología opuesta. Por esto, puede ser interesante aplicar el modelo de predicción de sentimientos para saber la polaridad del mensaje de cada partido respecto a los partidos contrarios, es decir, ver el grado de positividad de los mensajes de PSOE y UP respecto a PP y VOX y viceversa.

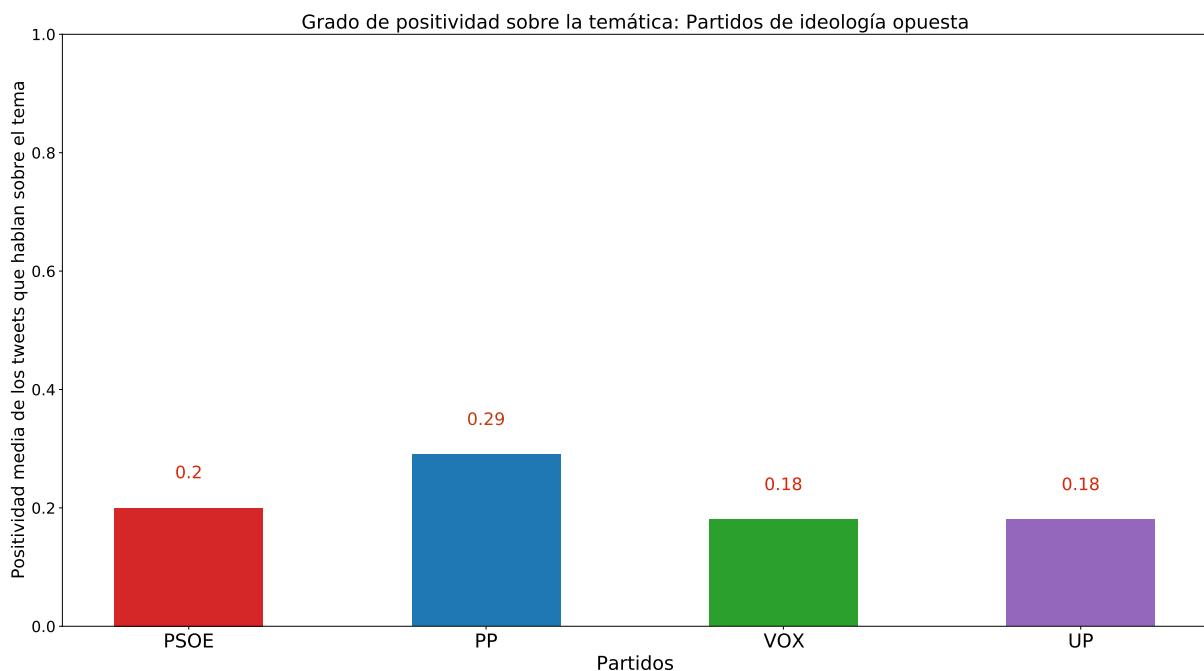


Figura 5.19: Análisis de sentimientos sobre partidos de ideología opuesta

Como muestra la Figura 5.19, la percepción de los partidos respecto a sus rivales políticos es muy mala, lo que no es una sorpresa, dado el clima de crispación política que se respiraba (y se sigue respirando) en ese momento.

5.6. Periodo 3 (17 de febrero - 8 de marzo)

Al inicio de este periodo nos encontramos en el principio de la crisis del COVID-19 en Italia, que fue la primera en Europa, y fue seguida por la crisis de España, cuyo inicio se corresponde al final de este periodo.

Además, este periodo abarca hasta el 8 de marzo, día de importancia mundial en el movimiento feminista pues es la fecha en la que se convocan movilizaciones a favor del feminismo en todas partes del mundo.

Como era de esperar, la presencia del coronavirus en los tweets de nuestro corpus sigue aumentando, alcanzando una presencia del 6,25 %.

Análisis de temáticas (LDA)

En este periodo podemos distinguir dos temáticas claramente dominantes: feminismo y el coronavirus.



Figura 5.20: Principales temáticas del PSOE en el tercer periodo

Por su parte, en el discurso del PSOE (Figura 5.20) se refleja la temática de feminismo con términos como *mujeres*, *igualdad*, *feminismo*, etc.

En cuanto al coronavirus, vemos que aparece en el discurso socialista en términos como *coronavirus*, *sanidad* o *salud*.



Figura 5.21: Principales temáticas del PP en el tercer periodo

El PP (Figura 5.21) también tiene presente al feminismo como temática en su discurso, aunque en menor medida que en el discurso del partido liderado por Pedro Sánchez.

En relación al COVID-19, es remarcable como en el mensaje del Partido Popular se establece una relación entre Italia y el virus, como se puede apreciar en la segunda nube de palabras.



Figura 5.22: Principales temáticas de VOX en el tercer periodo

En el caso de la formación de Santiago Abascal (Figura 5.22), se habló también de feminismo, pero en un tono muy diferente al usado por formaciones como el PSOE o UP (Figura 5.24), como veremos en el análisis de polaridad. En cuanto a Unidas Podemos, podemos ver que el feminismo fue indiscutiblemente el tema principal en su discurso.

En lo relativo al coronavirus, como podemos ver en términos como *entrada*, *italia* o *fronteras*, o en el tweet de la Figura 5.23, VOX empieza a hablar sobre control en las fronteras a causa de la enfermedad.



Figura 5.23: Tweet de Santiago Abascal sobre el control de las fronteras en el inicio de la expansión del coronavirus



Figura 5.24: Principales temáticas de UP en el tercer periodo

Análisis de polaridad

En este rango es interesante analizar la polaridad de los mensajes relacionados con el feminismo, ya que es una temática de relevancia común en los cuatro principales partidos.

En la Figura 5.25 podemos apreciar que el tono general de los partidos políticos sobre este tema es bastante positivo, a excepción de VOX, donde el grado de positividad apenas alcanza un 35 %, hecho que puede deberse a que según esta formación política, el movimiento feminista representado el 8M es un “feminismo supremacista”, tal y como dijo la cuenta oficial del Grupo Parlamentario de VOX en el Congreso el 4 de marzo.

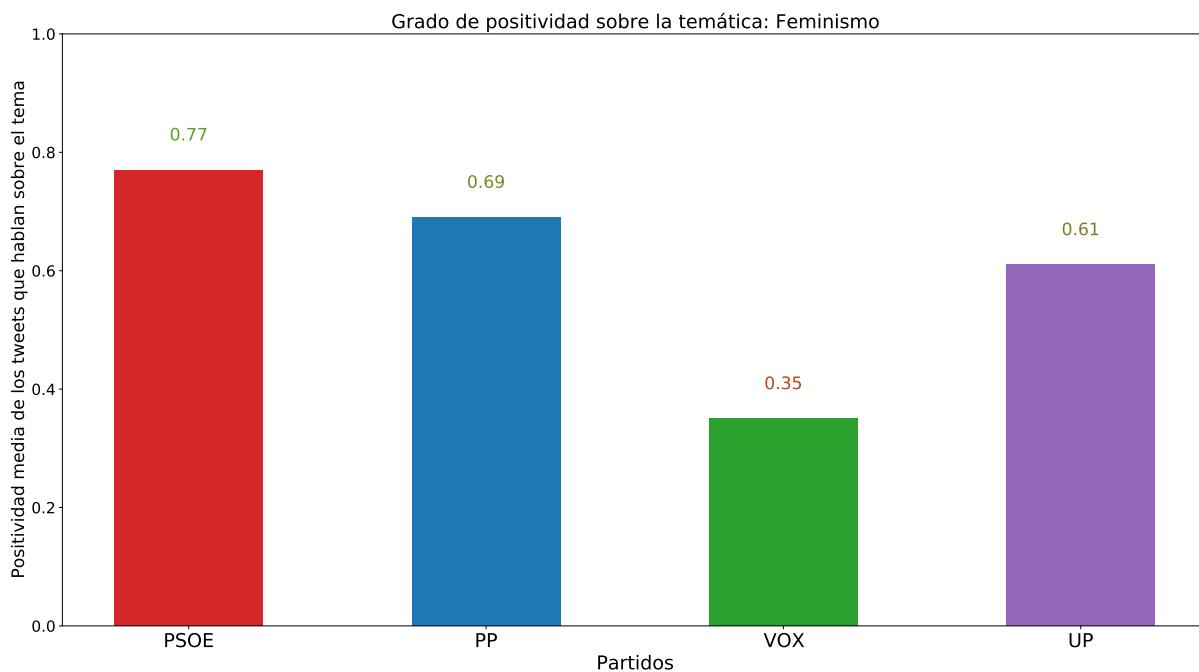


Figura 5.25: Análisis de sentimientos sobre feminismo

5.7. Periodo 4 (9 de marzo - 15 de marzo)

Este periodo cubre desde el 9 de marzo, que fue el inicio de la crisis en España, hasta el 15 de marzo, día posterior al inicio de la cuarentena.

En este rango temporal, el incremento de tweets respecto al coronavirus es enorme, llegando al 42 % de mensajes que lo mencionan directamente.

Análisis de temáticas (LDA)

Como era previsible, todas las temáticas giran de una manera u otra en torno al COVID-19.



Figura 5.26: Principales temáticas del PSOE en el cuarto periodo

Por su parte, el PSOE (Figura 5.26) hace bastantes referencias al confinamiento y al cese de actividad que conlleva, con términos como *actividad*, *casa*, *presencial*, etc.

También hace bastante menciones a las comparecencias telemáticas que se harían tan habituales durante el confinamiento, con términos como *reunión*, *videoconferencia*, *comparecencia*, etc.

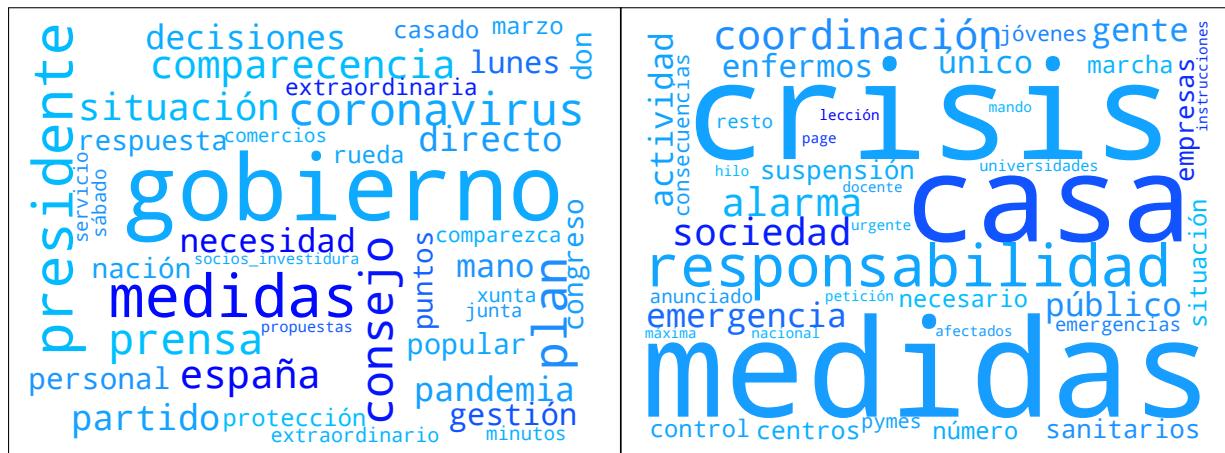


Figura 5.27: Principales temáticas del PP en el cuarto periodo

Sorprende la similitud entre los temas del PP (Figura 5.27) y el PSOE. Como podemos ver, las temáticas que hablan de comparecencias y del confinamiento son comunes en ambas fuerzas políticas, lo cual es más que probable que se deba al hecho de que al ser ambos partidos los que gobiernan en la mayoría de territorios, son también los que más comparecencias han tenido que dar, sobre todo al principio de la crisis, para hablar sobre el confinamiento en cada una de las regiones.

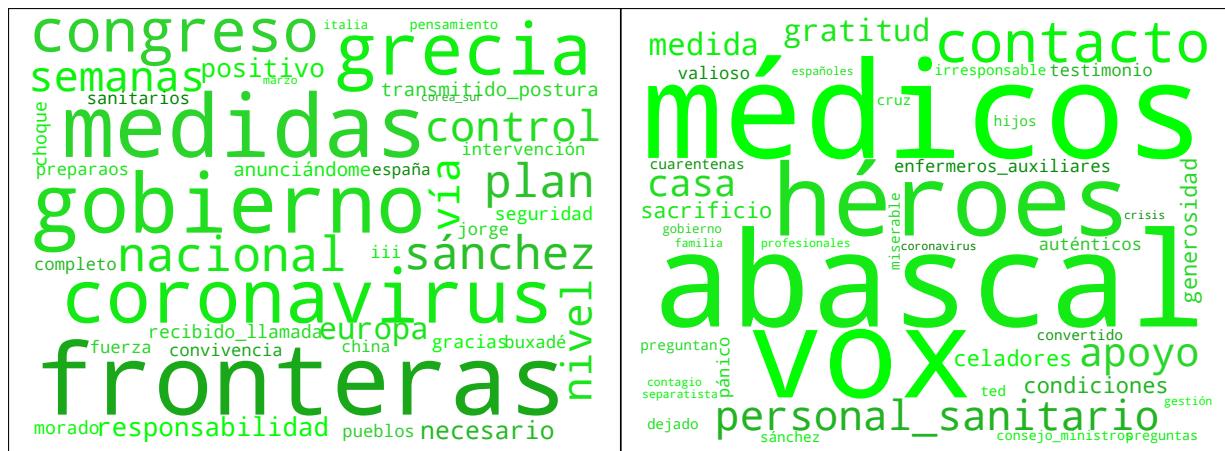


Figura 5.28: Principales temáticas de VOX en el cuarto periodo

VOX (Figura 5.28) habla sobre fronteras, relacionandola con el coronavirus, pero no de manera exclusiva: también habla sobre las fronteras en relación a la crisis migratoria entre Turquía y Grecia (de ahí que aparezca Grecia entre los términos más destacados).

Por otra parte, VOX también habla del personal sanitario en unos términos bastante positivos, como reflejan términos como *héroes* o *gratitud*.



Figura 5.29: Principales temáticas de UP en el cuarto periodo

Unidas Podemos (Figura 5.29) también centra gran parte de su mensaje en agradecer al colectivo sanitario, aunque haciendo hincapié en el carácter público de este. Por otra parte, habla también de las consecuencias sociales y económicas de la crisis del coronavirus en la clase trabajadora (*crisis, medidas, sociales, pymes, autónomos, impacto, empresas, trabajadoras*).

Análisis de polaridad

Como hemos visto, gran parte del mensaje se centra en el personal sanitario. Por eso, puede ser significante medir esta vez la polaridad de los tweets que hablen sobre el sistema sanitario y todo lo que este engloba.

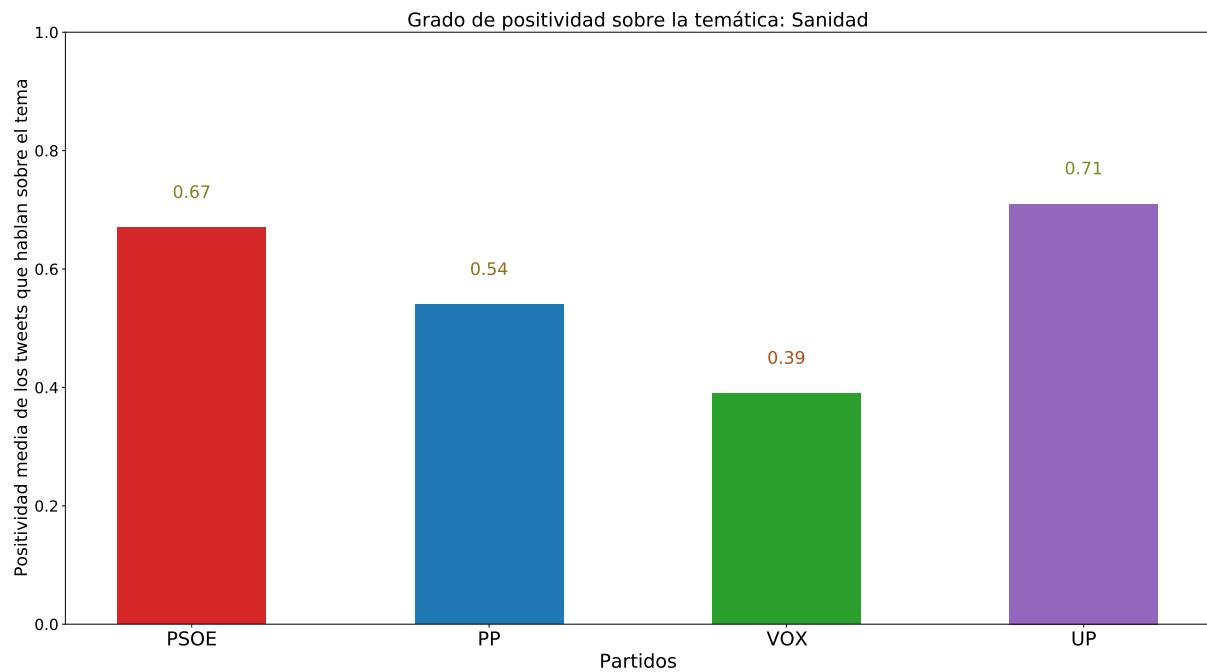


Figura 5.30: Análisis de sentimientos sobre el sistema sanitario

Como refleja la Figura 5.30, los partidos del Gobierno tienen una tendencia más positiva que la oposición, lo cual se debe a que aunque tanto el PP como VOX hablaron de manera positiva sobre el personal sanitario, también criticaron la gestión de los recursos sanitarios, lo cual se refleja en la polaridad media, ya que dichos mensajes también hablan sobre el tema.

5.8. Periodo 5 (16 de marzo - 26 de marzo)

Durante este periodo España pasa uno de los momentos más duros de la crisis, con el miedo al colapso sanitario y la apertura de hospitales de campaña.

En cuanto a los tweets que hablan directamente de la enfermedad, se mantienen en un porcentaje bastante alto, concretamente en un 39,67 %.

Análisis de temáticas (LDA)



Figura 5.31: Principales temáticas del PSOE en el quinto periodo

Podemos ver como gran parte mensaje del PSOE (Figura 5.31) en este periodo llamaba a la unidad ante la crisis, con términos como *coordinación*, *gracias*, *crisis* o *juntos*, hecho que no es de extrañar, ya que pasábamos un momento de especial crudeza e incertidumbre, así que es lógico que el gobierno intentara promover este tipo de mensajes.

Por otro lado, términos como *medidas*, *administraciones*, *recursos* o *gobierno* reflejan que otra parte del mensaje de los socialistas hablaba sobre la gestión de la crisis.



Figura 5.32: Principales temáticas del PP en el quinto periodo

El PP por su parte (Figura 5.32) volvió a un discurso más bronco, en el que señala al Gobierno (y especialmente a Pedro Sánchez), calificándolo de *vergüenza* o *irresponsable*.

Otra parte del mensaje del PP proviene de sus gobiernos autonómicos, donde se habla de la gestión de la crisis a nivel regional, haciendo especial hincapié en las personas mayores y en las residencias de ancianos.



Figura 5.33: Principales temáticas de VOX en el quinto periodo

El traslado al puerto de Almería de 74 personas rescatadas de una patera en el mar de Alborán fue uno de los detonantes para que la inmigración fuera un tema principal en el discurso de VOX (Figura 5.33). La formación dirigida por Santiago Abascal criticaba que los “puertos siguen abiertos a la inmigración ilegal en un momento en el que nuestros recursos, sanitarios y policiales, son más necesarios que nunca”, según palabras textuales de Rubén Pulido, secretario de prensa y comunicación de VOX Andalucía.

La duda sobre la gestión de recursos del Gobierno fue otra temática habitual en el discurso de VOX, tal y como reflejan términos como *gasto superfluo*, *recursos* o *gobierno*.



Figura 5.34: Principales temáticas de UP en el quinto periodo

Unidas Podemos (Figura 5.34), por su parte, sigue con un discurso muy similar al del periodo anterior, con un mensaje que muestra agradecimiento al sector sanitario público. Por otro lado, también afronta el tema de la crisis social provocada por la pandemia (*crisis, social(es), medidas, etc.*)



Figura 5.35: Tweet de la cuenta oficial del Ministerio de Derechos Sociales, agradeciendo a la ciudadanía y al sector sanitario

Análisis de polaridad

Como hemos podido ver en el LDA, la gestión de recursos es un tema muy discutido en este periodo. Por ello, vamos a analizar la polaridad de mensajes que hablen sobre este tema.

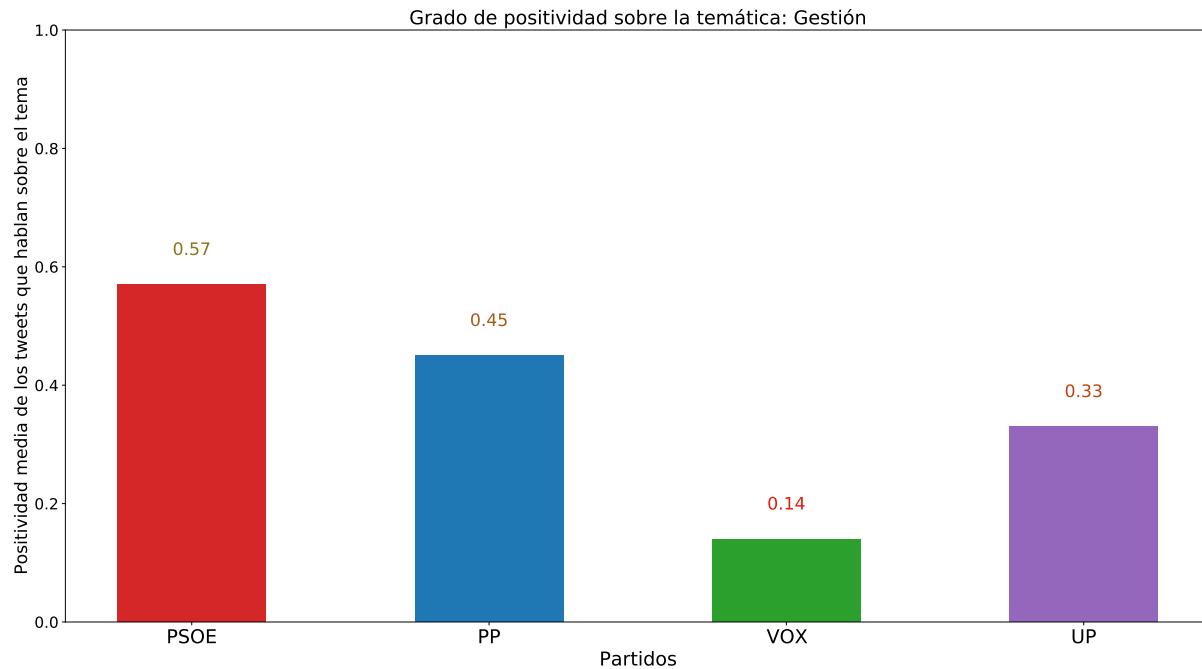


Figura 5.36: Análisis de sentimientos sobre la gestión

Como se ve en la Figura 5.36, el PSOE y el PP muestran un tono relativamente neutro. Por su parte, Unidas Podemos tiene un mensaje negativo, aunque es VOX el que con diferencia tiene el mensaje más crítico.

5.9. Periodo 6 (27 de marzo - 5 de abril)

Durante este periodo se suspendió toda actividad presencial no esencial, y se vivieron algunos de las jornadas más duras de la crisis, como el día en el que se registró el mayor número de infectados o el día en el que se registró el mayor número de fallecidos a causa del virus.

El porcentaje de tweets que mencionan al coronavirus es de un 34,7 %, manteniéndose prácticamente igual que el periodo anterior.

Análisis de temáticas (LDA)

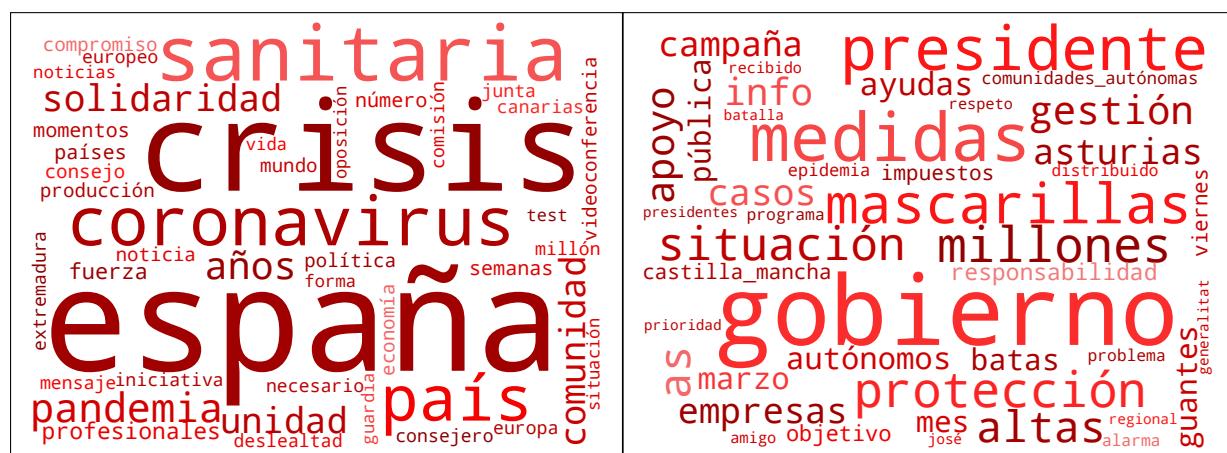


Figura 5.37: Principales temáticas del PSOE en el sexto periodo

El mensaje del PSOE (Figura 5.37) sigue buscando un sentimiento de unidad ante la pandemia, como podemos ver en el primer grupo de palabras, donde se relacionan términos como *coronavirus*, *España* o *crisis* con otros como *fuerza*, *unidad* o solidaridad.

Por otro lado, la formación liderada por Pedro Sánchez también habla mucho sobre material sanitario (*mascarillas, guantes, batas, protección, etc.*), ya que este periodo coincide con el clamor social por la urgente necesidad de este tipo de bienes para el personal sanitario.



Figura 5.38: Principales temáticas del PP en el sexto periodo

El discurso del PP (Figura 5.38) comparte con el del PSOE el tema del material sanitario. Además, también discute sobre las consecuencias económicas de esta crisis, como reflejan términos como *familias*, *empresas*, *autónomos* o *ertos*.



Figura 5.39: Principales temáticas de VOX en el sexto periodo

VOX (Figura 5.39) por su parte mantiene un discurso muy duro y frontal, relacionando al *Gobierno*, a *Sánchez* o a *Iglesias* con términos como *criminal* o *muertos*.



Figura 5.40: Principales temáticas de UP en el sexto periodo

Es Unidas Podemos (Figura 5.40) la formación que más cambia su discurso, pasando de tener un mensaje relativamente neutro a acusar de manera directa a la oposición de estar llevando una campaña de odio mediante la difusión de bulos (*datos, vox, odio, bulos, etc.*).

En otra linea, también hablan sobre medidas para afrontar la crisis sanitaria y social que tendrán que gestionar como parte del Gobierno.

Análisis de polaridad

En este periodo, la Figura 5.41 mide las polaridades de los mensajes que tienen relación con la temática del material sanitario, que es uno de los más importantes de este bloque.

Como podemos ver, VOX es el partido que tiene un mensaje más negativo, mientras que UP, PSOE y PP mantienen un tono relativamente neutro.

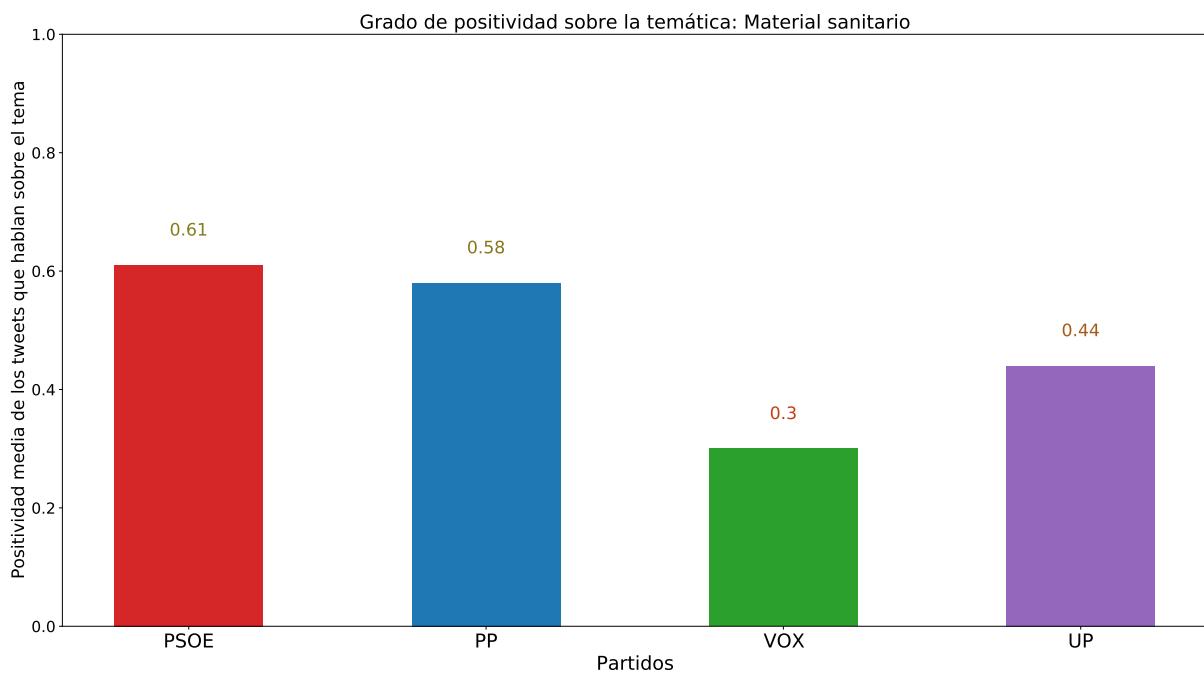


Figura 5.41: Análisis de sentimientos sobre la temática del material sanitario

5.10. Periodo 7 (6 de abril - 22 de abril)

En este periodo entramos ya en la fase de bajada de la famosa curva, y como consecuencia de esto, empiezan a surgir los primeros debates sobre cómo y cuando debe empezar el desconfinamiento.

El 32 % del total de los tweets mencionan de manera directa al COVID-19 en este periodo, porcentaje que se mantiene alto aunque baja gradualmente respecto a periodos anteriores.

Análisis de temáticas (LDA)



Figura 5.42: Principales temáticas del PSOE en el séptimo periodo

En esta ocasión, el discurso del PSOE (Figura 5.42) hace campaña contra la desinformación en torno al coronavirus. Podemos ver esto reflejado en términos como *derecha*, *odio*, *estrategia*, *fake news*, *desinformación* o *extrema derecha*.

Por otro lado, aparecen las primeras palabras relacionadas con el desconfinamiento (*fase*, *desescalada*), acompañada de otros términos que evocan unión y positividad : *responsabilidad*, *esperanza*, *esfuerzo*, etc.



Figura 5.43: Tweet de la cuenta oficial del Ministerio de Asuntos Exteriores hablando de la lucha contra la desinformación

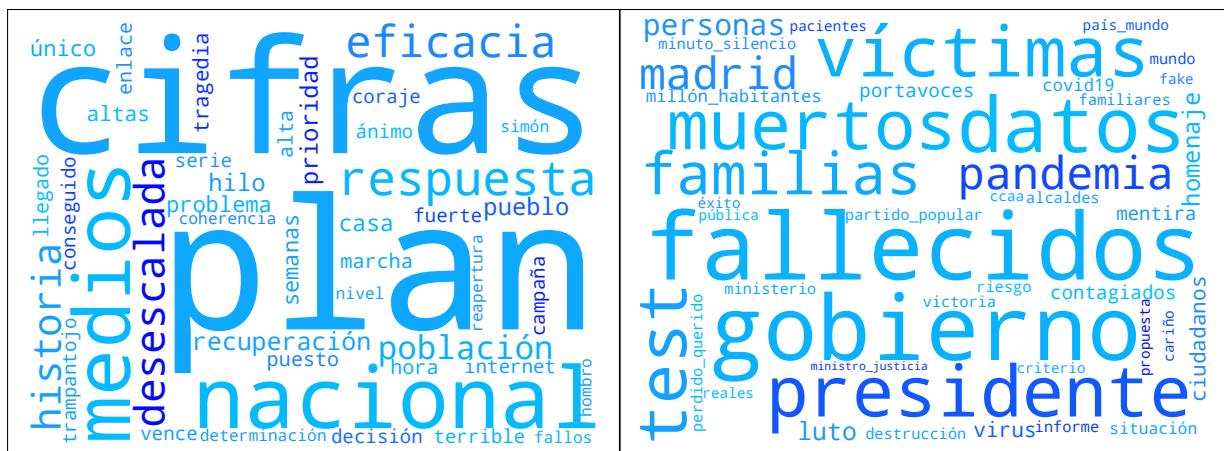


Figura 5.44: Principales temáticas del PP en el séptimo periodo

En el caso de la formación dirigida por Pablo Casado (Figura 5.44), vemos que también empiezan a aparecer palabras que hablan de la desescalada (*desescalada, plan, recuperación*)).

También podemos ver que el discurso del PP se vuelve cada vez más crispado contra el Gobierno, como refleja el tweet de Pablo Casado de la Figura 5.45.



Figura 5.45: Tweet de Pablo Casado criticando las intervenciones de Pedro Sánchez



Figura 5.46: Principales temáticas de VOX en el séptimo periodo

VOX por su parte (Figura 5.46) sigue dándole mucha prioridad al debate de la inmigración ilegal, como reflejan términos como *inmigrantes ilegales*, *llegados*, o *almería*, ya que el día 11 de abril siete inmigrantes dieron positivo por coronavirus en el centro de acogida de Almería.



Figura 5.47: Principales temáticas de UP en el séptimo periodo

En cuanto a UP (Figura 5.47), podemos ver términos positivos como *apoyo* y *protección*. También aparece por primera vez en su discurso el concepto de *ingreso mínimo*, medida que sería aprobada por el Gobierno el 29 de mayo.

Análisis de polaridad

El confinamiento y la desescalada son temas importantes en este periodo, por lo que va a ser la temática a la que realizarle la predicción de polaridad. Los resultados en general son relativamente neutros, siendo el PSOE la formación que mayor positividad refleja y VOX el que mayor negatividad, como refleja la Figura 5.48.

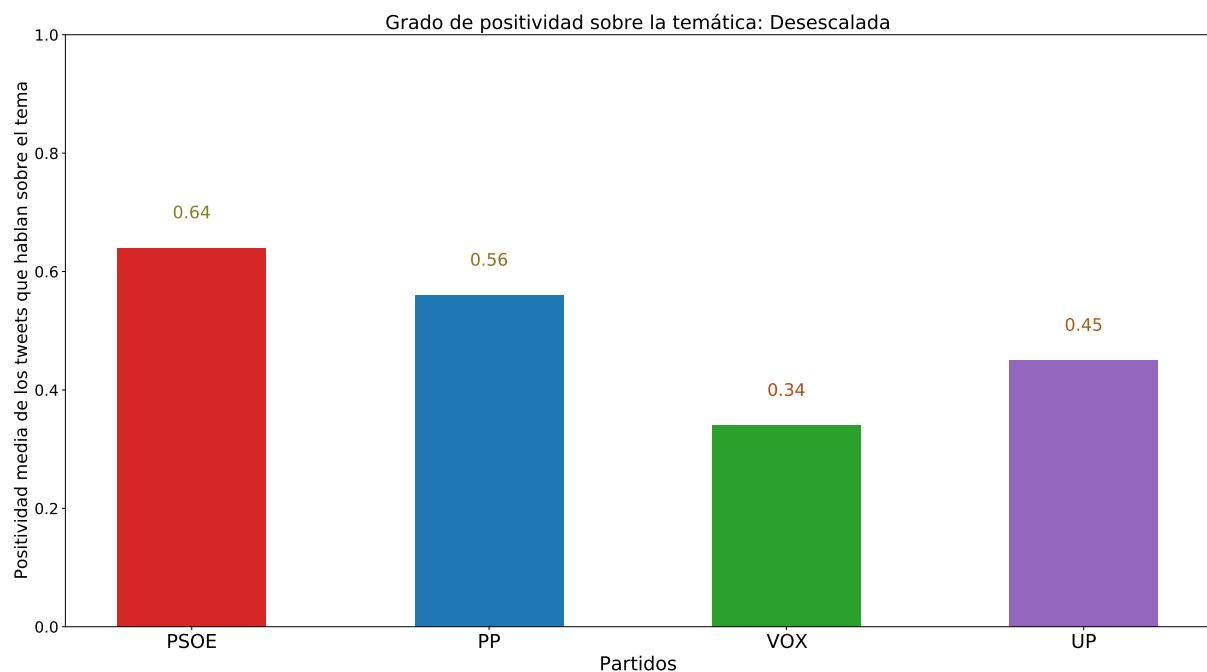


Figura 5.48: Análisis de sentimientos sobre la desescalada

5.11. Periodo 8 (23 de abril - 10 de mayo)

Este es el segundo periodo consecutivo donde tanto el número medio de fallecidos diarios como el número medio de contagiados diario disminuye. En esta ocasión, el porcentaje de tweets que mencionan al coronavirus es del 24,7 %

Ya se empezaba a pensar en la vida después del confinamiento, y la desescalada ya empezaba a ser una realidad, siendo la fase 0 que entró en vigor el 4 de mayo el punto de partida de la misma. Este hecho se refleja en que un 7,3 % de los tweets hablan directamente de la desescalada, frente al 1,1 % del periodo anterior.

Análisis de temáticas (LDA)



Figura 5.49: Principales temáticas del PSOE en el octavo periodo

Podemos ver como las temáticas del discurso del PSOE (Figura 5.49) empiezan a enfocarse más en la desescalada, como reflejan términos como *reconstrucción, plan, desescalada* o *normalidad, mascarillas, distancia*, etc.

El PP por su parte (Figura 5.51) vuelve señalar directamente al Gobierno (*Gobierno, presidente, Sánchez*), criticando el plan de desescalada y exigiendo más material sanitario para una “desescalada sin riesgos” (Figura 5.50).



Figura 5.50: Tweet de Pablo Casado exigiendo tests masivos y mascarillas



Figura 5.51: Principales temáticas del PP en el octavo periodo



Figura 5.52: Principales temáticas de VOX en el octavo periodo

El discurso de VOX (Figura 5.52) sigue siendo irritado y directo contra el Gobierno, e incluso empiezan a culparlo del confinamiento, como se ve en un tweet del líder de la formación en la Figura 5.53.

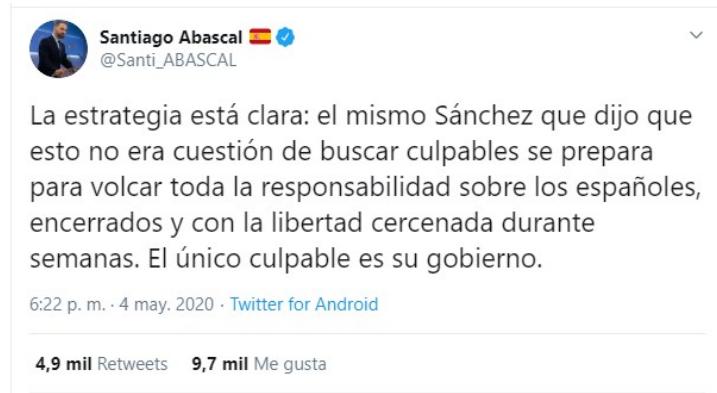


Figura 5.53: Tweet de Santiago Abascal culpando al Gobierno del confinamiento



Figura 5.54: Principales temáticas de UP en el octavo periodo

Unidas Podemos por su parte habló bastante sobre medidas sociales contra la crisis provocada por el coronavirus (*medidas, crisis, derechos, ertes*).

También estuvo muy presente en el discurso de UP (5.54) la figura de Billy el Niño, policía durante la época franquista española conocido por ser un cruel torturador, que murió el 7 de mayo por coronavirus sin que se le retiraran sus reconocimientos y medallas, y sin ser juzgado ni investigado.

Análisis de polaridad

Como hemos podido observar, el ambiente político en este periodo es crispado y lleno de menciones directas contra el Gobierno por parte de la oposición. Por ello, es interesante ver el grado de positividad de los mensajes que hablan sobre el Gobierno.

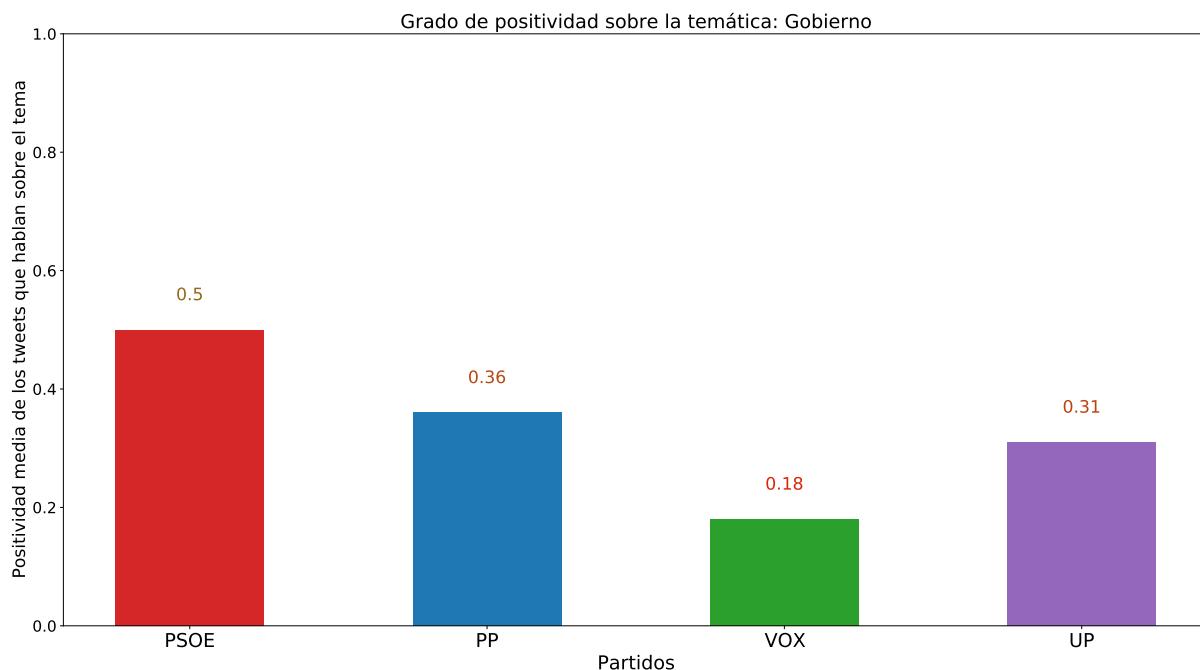


Figura 5.55: Análisis de sentimientos sobre el Gobierno

Los resultados son los mostrados en la Figura 5.55. Al principio puede sorprender el grado de negatividad de, por ejemplo, Unidas Podemos, pero investigando un poco se puede intuir que se debe a mensajes como este (Figura 5.56) de Juan López de Uralde, diputado de Unidas Podemos, en los que se critica la actitud de la oposición pero hablando también del Gobierno.



Figura 5.56: Tweet de Juan López de Uralde criticando a la oposición

5.12. Periodo 9 (11 de mayo - 26 de mayo)

El inicio de este periodo coincide con el día en el que comienza la fase 1 de la desescalada, y dura hasta un día después del inicio de la fase 2.

Se mantiene la tendencia a la baja en cuanto al porcentaje de tweets que hablan del coronavirus, siendo esta vez un 21 %. En cuanto a la presencia de tweets que hablan directamente de la desescalada, se mantiene prácticamente igual, en un 6,3 %.

Análisis de temáticas (LDA)

La gravedad de la pandemia en España va remitiendo, lo cual se refleja en las temáticas, que sin dejar el virus de totalmente de lado, empiezan a hablar de otras cosas.



Figura 5.57: Principales temáticas del PSOE en el noveno periodo

En el caso del PSOE (5.57) vemos que un tema clave es el plan de desconfinamiento, como podemos ver en términos como *medidas*, *desescalada*, *plan*, *seguridad*, etc.

También entra en escena el debate de si Madrid puede o no avanzar a la fase 1, evento que se refleja en el segundo grupo de palabras de las temáticas del PSOE.



Figura 5.58: Principales temáticas del PP en el noveno periodo

El PP por su parte tiene bastante presente en su discurso a la Guardia Civil, hecho provocado por la polémica destitución de Pérez de los Cobo, jefe de la Comandancia de la Guardia Civil de Madrid, que investigaba la manifestación del 8M. El propio Pablo Casado decía que es “un insulto a la Guardia Civil y al Estado de Derecho que Sánchez tape el cese del responsable de la investigación sobre los presuntos delitos el 8-M, con la equiparación salarial que aprobó el PP hace dos años.”



Figura 5.59: Principales temáticas de VOX en el noveno periodo

La temática de la Guardia Civil sería común en VOX, como podemos ver en la Figura 5.59.



Figura 5.60: Principales temáticas de UP en el noveno periodo

Mientras tanto, en Unidas Podemos (5.60), podemos ver que aunque siguen siendo importantes términos como *crisis* o *sanitaria*, empiezan a aparecer nuevos temas en palabras como *cambio climático* o *reforma laboral*.

Análisis de polaridad

En este periodo, Madrid concentró varios eventos importantes cargados de la polémica: por un lado, se debatía si la ciudad debía o no pasar a la Fase 1, y por otro lado, las caceroladas alentadas por VOX y respaldadas por el PP. Por ello, vamos a aplicar la predicción de polaridad a Madrid y a estos eventos que la rodearon. Los resultados podemos verlos en la Figura 5.61

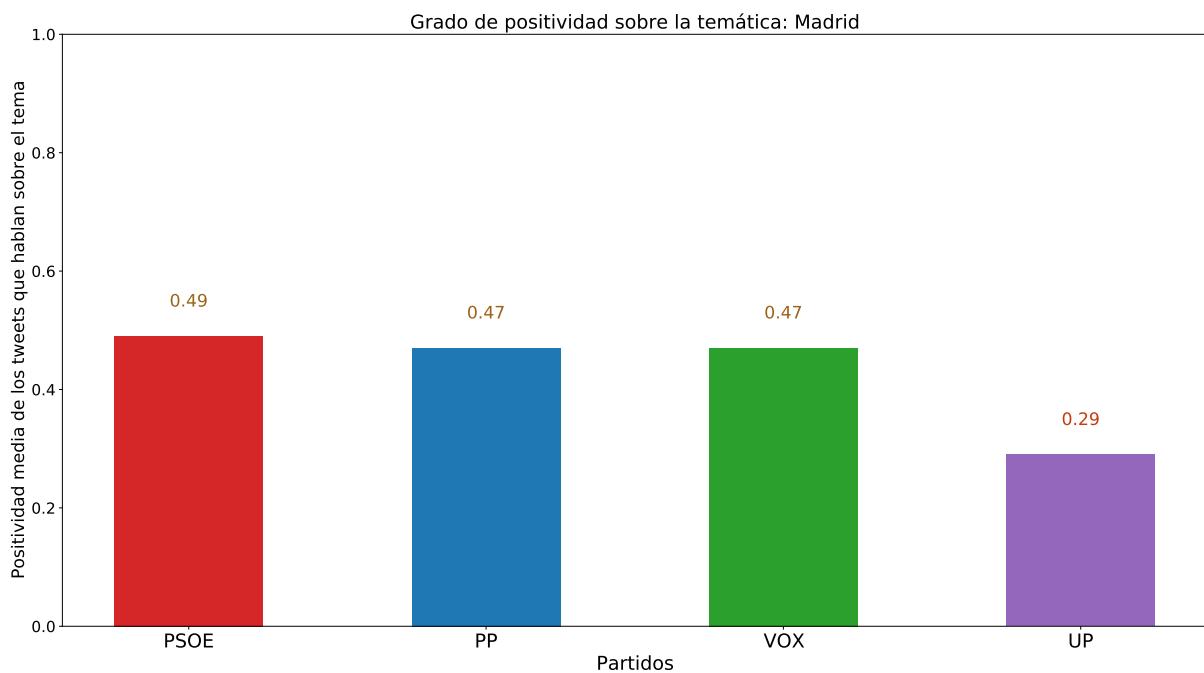


Figura 5.61: Análisis de sentimientos sobre Madrid

5.13. Periodo 10 (27 de mayo - 8 de junio)

Este penúltimo intervalo empieza apenas dos días después de la entrada de la Fase 2 de desescalada, extendiéndose hasta el 8 de junio, día en el que precisamente gran parte del territorio español entra en Fase 3.

El porcentaje de tweets que hablan del coronavirus es de un 19,5 %, manteniendo la tendencia a la baja de periodos anteriores. También es destacable la bajada en el porcentaje de tweets que hablan sobre el desconfinamiento, situándose en apenas un 3,1 %, lo que supone una bajada de más de 3 puntos respecto al periodo anterior.

Análisis de temáticas (LDA)



Figura 5.62: Principales temáticas del PSOE en el décimo periodo

Entre las principales temáticas del PSOE (Figura 5.62) vuelve a destacar *Madrid*, que una vez más, fue el centro de varias polémicas. Por un lado, fue una de las regiones que más tardó en avanzar de fase, y sin duda una de las más polémicas acarreó por las discrepancias entre el gobierno central y el autonómico.

Por otro lado, en este periodo explotó el escándalo provocado por supuestas órdenes del gobierno de Madrid de no hospitalizar a ancianos que vivían en residencias enfermos del COVID-19 con patologías graves o edades avanzadas [27].

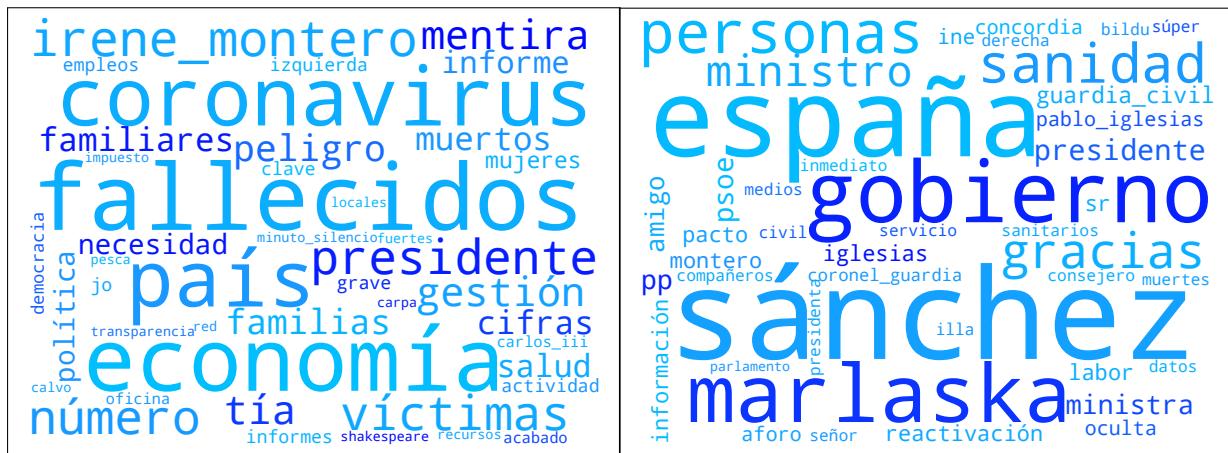


Figura 5.63: Principales temáticas del PP en el décimo periodo

En el PP (Figura 5.63) podemos ver muchas menciones directas a nombres propios: a Pedro Sánchez, Irene Montero o Fernando Grande-Marlaska. Las menciones a Pedro Sánchez han sido habituales durante todo el estudio, pero las otras dos si responden a eventos concretos:

El caso de Fernando Grande-Marlaska tiene respuesta en las polémicas destituciones y dimisiones en parte de la cúpula de la Guardia Civil (Figura 5.64).

Irene Montero fue centro de la polémica que trajo la filtración de un video grabado el 9 de marzo, en el que hablaba sobre como el miedo al coronavirus había repercutido en el número de asistentes a la polémica manifestación del 8-m.



Figura 5.64: Tweet de Pablo Casado pidiendo el cese de Marlaska

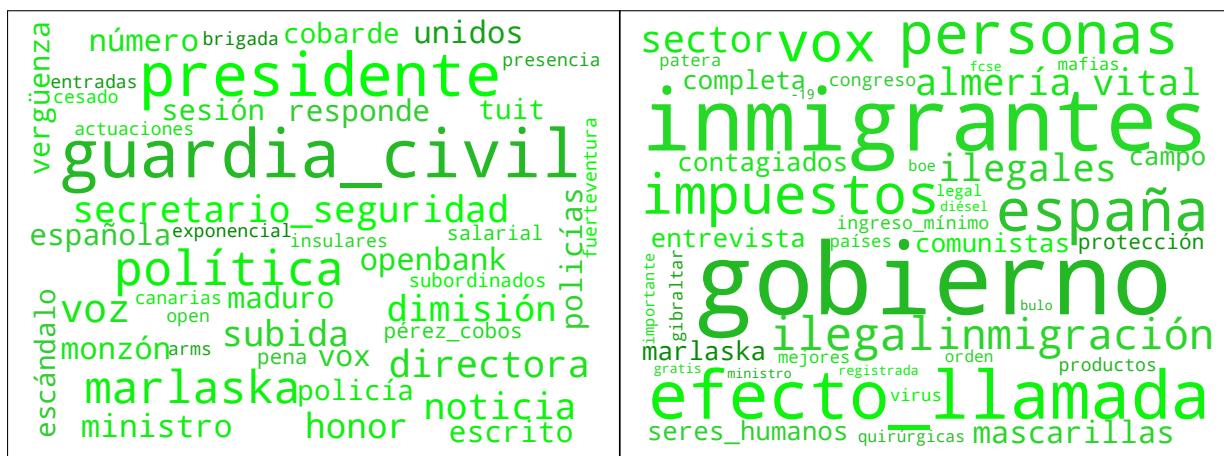


Figura 5.65: Principales temáticas de VOX en el décimo periodo

La polémica entre el actual ministro del Interior y parte de la cúpula de la Guardia Civil también fue un tema muy presente en el discurso de VOX (Figura 5.65).

Otro tema (de nuevo) muy presente en el mensaje de este partido fue el tema de la inmigración (*inmigrantes*, *inmigración*, *efecto llamada*, etc.). Esto en parte se debe a que, según VOX, la implementación del ingreso mínimo vital iba a desembocar en un efecto llamada de inmigrantes ilegales.



Figura 5.66: Retweet de Javier Ortega Smith en el que relaciona el efecto llamada con el impuesto mínimo vital

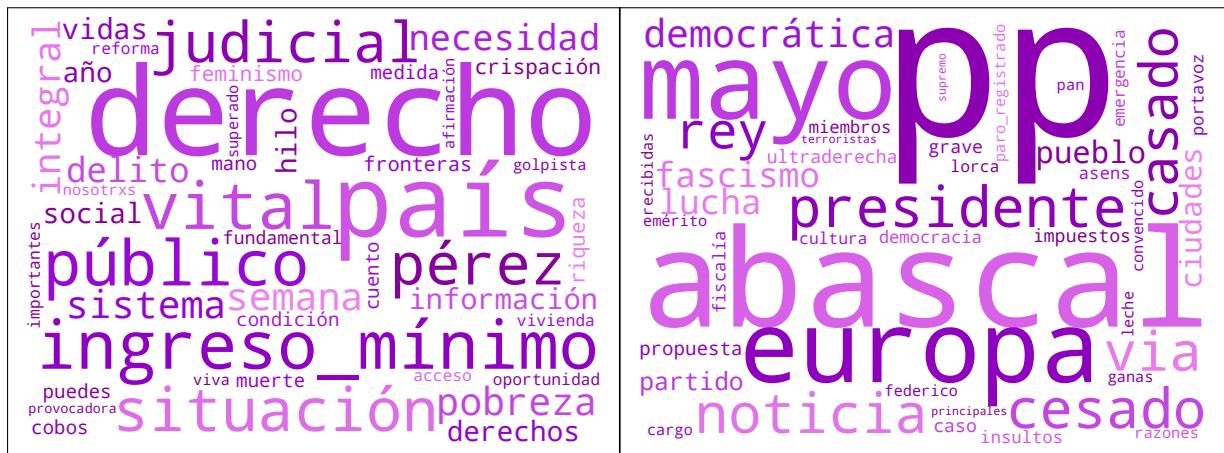


Figura 5.67: Principales temáticas de UP en el décimo periodo

El ingreso mínimo vital también tuvo mucha relevancia en el mensaje de Unidas Podemos (*ingreso mínimo, vital, derecho, pobreza, necesidad*).

También tuvieron muy presentes a la oposición, sobre todo al PP. Esto responde a la polémica intervención de su portavoz, Cayetana Álvarez de Toledo, en la que llamaba “terrorista” al padre del vicepresidente del Gobierno y líder de la formación morada, Pablo Iglesias.



Figura 5.68: Tweet de Pablo Iglesias criticando las declaraciones de Cayetana Álvarez de Toledo sobre su padre

Análisis de polaridad

En este periodo, la aprobación del ingreso mínimo vital ha sido uno de los temas más discutidos. Por ello, vamos a aplicar la predicción de polaridad a los mensajes que hablan de esto.

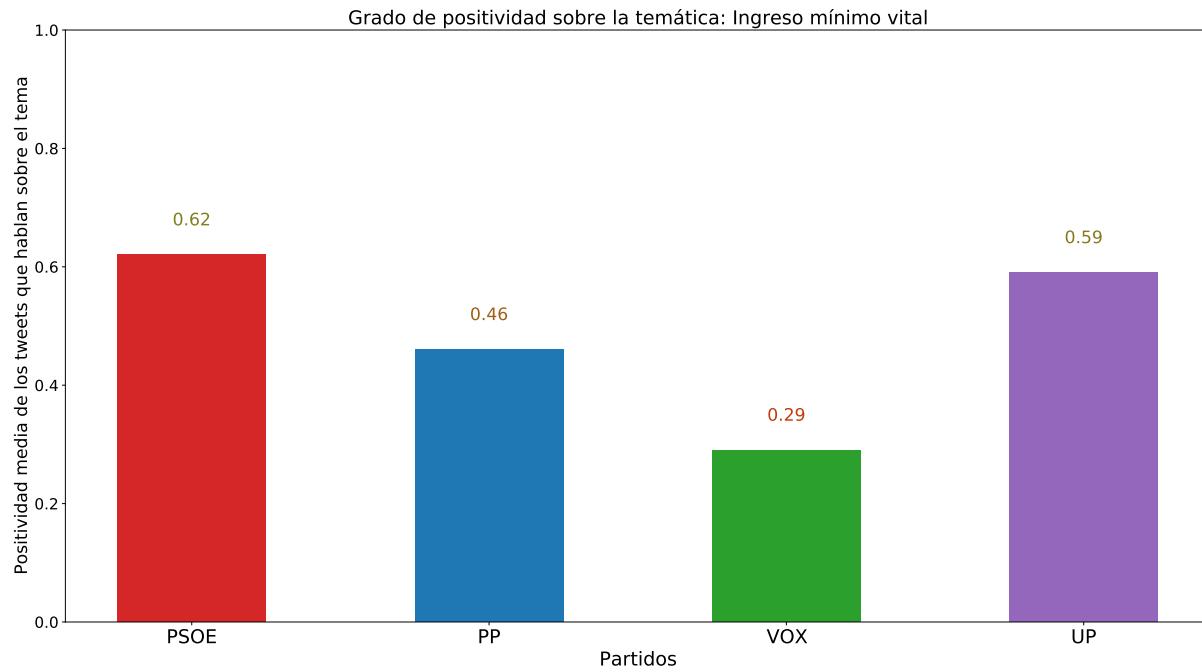


Figura 5.69: Análisis de sentimientos sobre el IMV

Como podemos ver en la Figura 5.76, los partidos del Gobierno son los que tienen un mensaje más positivo respecto al tema. El PP tiene un tono que se acerca a la neutralidad, lo cual tiene sentido, ya que aunque cuestionaron la medida, acabaron por apoyarla. VOX es el único partido que claramente tiene un tono negativo respecto al ingreso mínimo vital, hecho que se corresponde con lo que pudimos ver en el análisis de temáticas.

5.14. Periodo 11 (9 de junio - 21 de junio)

Entramos en el último periodo que vamos a analizar, que empieza el 9 de junio, justo un día después de la entrada de la Fase 3 en gran parte del país, y dura hasta el 21 de junio, día en el que finaliza el estado de alarma después de 99 días, para dar paso a la tan sonada “nueva normalidad”.

En esta ocasión, los mensajes que hablan directamente de la pandemia ocupan el 17% del corpus total. En cuanto a los mensajes que hablan directamente de la desescalada, bajan su presencia hasta situarse en un escaso 1,2%.

Análisis de temáticas (LDA)

Como se puede intuir por los números de tweets que hemos analizado arriba, este periodo tiene las temáticas en las que menos se habla del coronavirus desde el inicio de la pandemia.



Figura 5.70: Principales temáticas del PSOE en el undécimo periodo

En el PSOE (Figura 5.70) se distingue un discurso que habla de recuperación económica (*recuperación, consejo, económica, comisión reconstrucción*) tras la crisis del coronavirus, como refleja el hashtag *#covid19*, que fue el hashtag más utilizado en mensajes de esta temática.

También se mantiene la polémica de las residencias en Madrid, hecho que se refleja con términos como *residencias, investigación, protocolos, mentiras o PP*.



Figura 5.71: Principales temáticas del PP en el undécimo periodo

En el conjunto de las palabras que más destacan del discurso del Partido Popular (figura 5.71) vemos que hay bastantes menciones directas al gobierno de Sánchez, estableciendo una relación con conceptos como *fallecidos*, *pandemia* o *responsable*, entre otros.

Esta relación y el uso político que el PP ha dado de ella se sintetiza bastante bien en un tweet de Pablo Casado, que podemos ver en la Figura 5.72.



Figura 5.72: Tweet de Pablo Casado criticando la gestión de la crisis por parte del Gobierno



Figura 5.73: Principales temáticas de VOX en el undécimo periodo

En el caso de VOX (Figura 5.73) vemos que no hay muchos cambios respecto al periodo igual. Seguimos viendo muchas menciones directas al Gobierno y a miembros de este, y, por otro lado, sigue destacando en su mensaje el tema de la inmigración ilegal.



Figura 5.74: Principales temáticas de UP en el undécimo periodo

En el discurso de Unidas Podemos (Figura 5.74) sigue destacando el ingreso mínimo vital, y, por otra parte, la polémica de las residencias de ancianos en Madrid, tal y como podemos ver en la relación entre los términos *mentira* y residencias, y como se puede exemplificar en el Tweet de la figura 5.75.



Figura 5.75: Tweet de Pablo Iglesias sobre la oposición

Análisis de polaridad

En este periodo, la controversia sobre la gestión de las residencias en mayores en la Comunidad de Madrid ha sido uno de los temas más polémicos, y, por ello, va a ser el tema al que apliquemos nuestro modelo de predicción de polaridad.

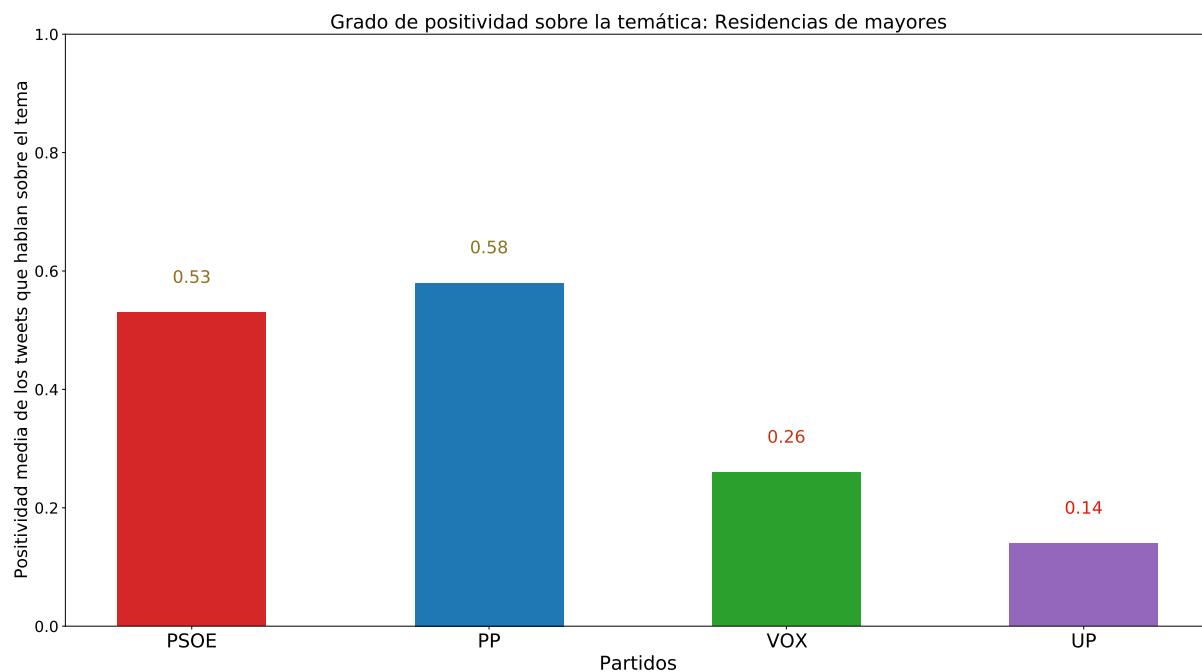


Figura 5.76: Análisis de sentimientos sobre la gestión de las residencias de ancianos en Madrid

Como podemos ver desde el análisis de temáticas, Unidas Podemos ha sido el partido con diferencia más crítico en este tema, seguido de Vox. El PSOE muestra un tono medio de neutralidad, mientras que el Partido Popular, formación que gobierna en Madrid, ha sido el partido que ha mantenido un mensaje más positivo.

Capítulo 6

Conclusiones

En este proyecto se ha realizado un análisis sobre cuales han sido las principales temáticas y el tono respecto a estas en los discursos de los principales actores políticos del país durante el transcurso de la crisis social, sanitaria y económica provocada por el COVID-19, utilizando Twitter para acceder a dichos discursos. Las principales conclusiones son los siguientes:

- La pandemia ha sido el protagonista absoluto en el discurso político en Twitter, y además ha provocado un incremento en el número total de mensajes en las cuentas de la que hemos hecho el seguimiento.

En enero y febrero, que son los meses antes de la pandemia, nuestro corpus tiene una media de 657,8 y 694 mensajes al día, respectivamente. Marzo y abril, que son los meses en los que encontramos un mayor porcentaje de tweets dedicados al COVID-19, tenemos una media de 994,9 y 978,8 mensajes, respectivamente. A partir de ahí el porcentaje de mensajes sobre el coronavirus va decayendo, tendencia que sigue el número de tweets diarios, bajando hasta los 890,1 mensajes diarios en mayo, y a 724,6 a día 22 de junio.

- Generalmente, hemos tenido un clima político muy irritado. Como podemos ver en la inmensa mayoría de periodos que hemos analizado, no han faltado las menciones directas entre los principales partidos políticos y sus líderes.

Un dato que refuerza esta conclusión es la polaridad de los mensajes de cada partido político respecto a sus rivales. En el caso del Partido Popular, la polaridad media de los mensajes respecto a los partidos que conforman el Gobierno es de 0,28 sobre 1, valor que en el caso de VOX baja hasta el 0,19. En el caso de los partidos del Gobierno se repite la tendencia, siendo la polaridad media de los mensajes sobre la oposición de 0,23 sobre 1 en el caso del PSOE, y 0,22 en el caso de UP.

- El Partido Socialista es la formación con un mensaje medio más positivo, con una polaridad media de 0,65 sobre 1. A la formación liderada por Pedro Sánchez le siguen el Partido Popular (0,59), Ciudadanos (0,5) y Unidas Podemos (0,49), mientras que VOX tiene el valor más bajo, con un 0,37 sobre 1.

Bibliografía

- [1] John H. Parmelee; Shannon L. Bichard (2013). Politics and the Twitter Revolution: How Tweets Influence The Relationship Between Political Leaders And The Public.
- [2] Pérez-Dasilva, Jesús-Ángel; Meso-Ayerdi, Koldobika; Mendiguren-Galdospín, Terese (2020). “Fake news y coronavirus: detección de los principales actores y tendencias a través del análisis de las conversaciones en Twitter”. *El profesional de la información*, v. 29, n. 3, e290308.
- [3] Bakal, Gotkhan; Kavuluru, Ramakanth (2017). “On quantifying diffusion of health information on Twitter”. *2017 IEEE EMBS International conference on biomedical & health informatics (BHI)*, pp. 485-488.
- [4] Fox, Susannah (2011). “The social life of health information, 2011”. Pew Research Center: Internet, science & tech, 12 May.
- [5] Estela Pato, *Estadísticas de redes sociales 2020 en España - Concepto05*. 10 de marzo de 2020
- [6] Pablo Martínez, *Twitter es la red donde la información política tiene mayor relevancia* - Blog de Twitter. 9 de abril 2019
- [7] Twitter, *Developer*
- [8] Twitter, *Get Tweet timelines*
- [9] Twitter, *Tweet objects*
- [10] Dipanjan Sarkar, *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. 2019.
- [11] Help Twitter, Acerca del servicio de enlace de Twitter (<http://t.co>)
- [12] Github. Emoji
- [13] Help Twitter. Ayuda con el registro del nombre de usuario

- [14] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi e Iti Mathur, *Natural Language Processing: Python and NLTK. Tokenization.* Noviembre de 2016.
- [15] NLTK. Accessing Text Corpora and Lexical Resources
- [16] Snowball. Defining R1 and R2
- [17] Siddhartha Chatterjee y Michal Krystyanczuk, *Python social media analytics.* 2018.
- [18] Thushan Ganegedara, *Intuitive Guide to Latent Dirichlet Allocation - Towards Data Science.* 23 de agosto de 2018
- [19] Wikipedia. Dirichlet distribution.
- [20] Tellez, Eric and Miranda-Jiménez, Sabino and Graff, Mario and Moctezuma, Daniela and S. Siordia, Oscar and Villaseñor García, Elio *A case study of Spanish text transformations for twitter sentiment analysis.* Septiembre de 2017
- [21] OpenPyXL
- [22] Tweepy. Cursor
- [23] TASS: Workshop on Semantic Analysis at SEPLN
- [24] Pandas - Python Data Analysis Library
- [25] Pandas - Python Data Analysis Library
- [26] Scikit-learn. Machine Learning in Python
- [27] Eldiario. Madrid indicó a los médicos de Primaria evitar el traslado al hospital de mayores con COVID-19 y patologías graves