COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

TEXT MINING Y EL CICLO ELECTORAL ESPAÑOL DE 2023: UN ANÁLISIS DE LOS PRINCIPALES DIARIOS

Alumno: Luis Villanueva Ribes
Clave: 201901368
Director: Alejandro Rodríguez Gallego

MADRID | Junio 2024

RESUMEN

Desde el inicio de la sociedad de la información, la influencia de los medios de comunicación de masas en la opinión pública ha sido un asunto de largo y tenso debate, especialmente en lo referido al ámbito político. En este trabajo se pretende estudiar esa influencia a través del análisis de cinco de los principales medios de comunicación españoles: El Mundo, El País, El Confidencial, ABC y El Español. Este análisis se centra en el ciclo electoral de 2023, entendido como todo el periodo comprendido desde el inicio de la campaña electoral de las elecciones autonómicas y municipales del 28 de mayo hasta la investidura de Pedro Sánchez Pérez-Castejón como Presidente del Gobierno. Para ello se ha conformado un *corpus* a través de la noticia principal de cada periódico en cada uno de los días del periodo. Este *corpus* ha sido posteriormente sometido a un *topic analysis*, cuyos resultados han sido debidamente analizados. Por último, se ha realizado un análisis más individualizado, centrado en cuál ha sido la forma en la que cada medio ha presentado y tratado los diferentes *topics*. De esta manera, este trabajo pretende extraer cuáles han sido los temas más importantes para la prensa española durante el ciclo electoral de 2023.

PALABRAS CLAVE

Medio de comunicación, ciclo electoral, coalición, investidura, Congreso de los Diputados, encuesta electoral, análisis de tópicos, portadas, negociaciones

ABSTRACT

Since the beginning of the society of information, the influence of mass media in public opinion has been a topic subject to a long and tense debate, especially when related to the political sphere. In this paper we plan to study the aforementioned influence through the analysis of five of the most widespread Spanish newspapers: *El Mundo*, *El País*, *El Confidencial*, *ABC* and *El Español*. This analysis focuses on the 2023 electoral cycle, understood as the period that spans from the beginning of the regional and local elections held on the 28th of May to the appointment of *Pedro Sánchez Pérez-Castejón* as Prime Minister. In order to do so, a *corpus* has been ensembled with the main news item of each newspaper on each day of the period. This *corpus* has subsequently been subject to topic analysis, whose results have been properly interpreted. Finally, a more individualized analysis has been carried out, focusing on what has been the way in which each newspaper has presented and dealt with the different topics. In this manner, this study pretends to extract the most important topics for the Spanish press during the 2023 electoral cycle.

KEY WORDS

Media, electoral cycle, coalition, appointment, Congress of Deputies, electoral poll, topic analysis, newspaper cover, negotiations

ÍNDICE:

CAPÍTULO I: CONTEXTO E INTRODUCCIÓN AL TRABAJO

1. INTRODUCCIÓN

Tras las elecciones generales de 1993, el entonces Presidente del Gobierno español, Felipe González, popularizó la frase por la que afirmó que no era lo mismo la opinión pública que la opinión publicada. Esta frase, que hacía referencia a su victoria inesperada y contraria a las expectativas generadas por todos los medios de comunicación, ha servido para ilustrar la constante pugna entre lo que la gente cree y lo que los medios difunden.

En las sociedades globalizadas, el papel de los medios de comunicación como vías para informar acerca de la realidad social ha sido desde hace cierto tiempo cuestionado y criticado por algunos sectores. Donde más intensidad ha cobrado esta crítica ha sido en el sector de la información política, que ha experimentado una significativa polarización en el último ciclo electoral. Por ello, el presente trabajo busca a través de *text mining* realizar un análisis de los principales temas presentes en los medios de comunicación durante el ciclo electoral de 2023, además de su evolución temporal.

El ciclo electoral de 2023 comenzó el 3 de abril, con la publicación del Real Decreto (en adelante, RD) 207/2023, de 3 de abril, por el que se convocan elecciones locales y a las Asambleas de Ceuta y Melilla para el 28 de mayo de 2023. El RD, en su artículo segundo, recoge que la campaña electoral comenzará a las 00:00 horas del viernes 12 de mayo, día de inicio del presente trabajo y día de referencia para el inicio del estudio de campo. De la misma manera, dicho ciclo finaliza con la publicación del RD 828/2023, de 16 de noviembre, por el que se nombra Presidente del Gobierno a don Pedro Sánchez Pérez-Castejón. La fecha de publicación de este RD es la tomada como referencia para la finalización del estudio de campo y supone el final de su extensión.

Como todo ciclo electoral, suscitó un enorme interés en los medios de comunicación, que siguieron de forma diaria la evolución de la campaña o la selección de candidatos. En este sentido, lo que busca este trabajo es verificar en primer lugar, cuál es el grado de impacto del ciclo electoral en los contenidos presentados por los medios: qué se menciona con mayor frecuencia. En segundo término, la intensidad con la que lo hacen: cuál es la evolución temporal de lo que más se menciona.

Para ello, se va a llevar a cabo un proceso de *text mining* sobre las noticias de cinco diarios que reflejan posiciones políticas diversas: El Mundo, El País, El Confidencial, ABC y El Español. El estudio se ha estructurado en varias etapas diferenciadas:

extracción del *corpus*, adaptación y análisis de este, preprocesamiento de la información y conclusiones iniciales, *topic analysis* con sus correspondientes conclusiones posteriores y análisis particular de los diferentes medios. Mediante el *topic analysis* lo que se busca es obtener un reflejo de qué temas son los que más interesan a los medios, es decir, de qué se ha hablado más a lo largo de los meses. El análisis de los distintos diarios nos permitirá ver cómo han contribuido las noticias de cada uno a la creación del paisaje global derivado del *topic analysis*.

Todo este análisis se realizará de la misma manera: seleccionando la noticia de portada de los diarios desde el día 12 de mayo hasta el día 16 de noviembre. Esta portada será siempre la de la versión matutina, pues si se seleccionasen noticias posteriores se podría perder la capacidad comparativa entre los diferentes medios. Por añadidura, se trata de la primera información disponible para la ciudadanía, por lo que es la noticia que los medios entienden como más importante o reseñable del día.

A través de esta investigación, se pretende ver cómo la forma que tienen los medios de comunicación de informar termina generando un estado de opinión que nace de posiciones diversas, fruto de la propia línea editorial de cada medio y que, como avisó el ahora expresidente González, en muchas ocasiones no se corresponde en absoluto con la realidad.

2. OBJETIVOS

Los objetivos principales del presente trabajo son los siguientes:

Primero: Verificar el grado de atención que los diferentes medios de comunicación seleccionados han prestado al ciclo electoral de 2023. Para ello, se elegirá la noticia de portada de cada uno de los cinco medios elegidos. La muestra abarcará título, subtítulo y parte del cuerpo. Esta selección busca dos cosas: analizar el nivel de interés que se presta a la información política y el nivel de desinterés que pasan a tener otras informaciones de distinta índole, como la información económica, social, internacional o cultural, entre otras.

Segundo: Comprobar los temas más relevantes de la información escogida, tanto en términos generales como temporales. Los ciclos electorales son procesos complejos, que abarcan muchos elementos: precampañas, campañas, resultados o pactos, entre otros. De la misma manera, nuestro mundo globalizado se encuentra siempre en constante cambio, con temas nuevos apareciendo y en muchos casos desapareciendo en cuestión de días. Mediante este análisis no sólo nos queremos circunscribir a los temas que más se mencionan, sino que buscamos expandirlo para ver cuál es su intensidad a lo largo del periodo estudiado.

Tercero: Examinar cómo los diferentes medios de comunicación han tratado los distintos asuntos. Si bien es cierto que del análisis global resultarán una serie de temas, buscamos también descomponer dicho análisis global en función de los diferentes medios de comunicación objeto de estudio. Con esta descomposición, el objetivo fundamental es ver tanto los *topics* más frecuentes en cada medio, como la aportación en términos proporcionales que realiza cada medio a cada *topic*, lo que nos puede dar pistas acerca de su línea editorial o forma de presentar la información. La vía escogida para cumplir con este objetivo será la de agrupar los resultados del análisis en función del medio, ilustrando las diferencias a través de gráficos.

Cuarto: Analizar la evolución que determinados términos han tenido a lo largo del tiempo. Las preocupaciones o implicaciones de los ciclos electorales son diversas en función del proceso electoral que se trate, especialmente fruto de la distribución competencial de nuestro modelo territorial. A la hora de poner propuestas sobre la mesa, no se tiene la misma capacidad si eres una corporación municipal que si eres el Gobierno de la Nación o el Consejo de Gobierno de una comunidad autónoma. De la misma manera, los candidatos que concurren a cada proceso electoral son, como es lógico, también

diferentes. Para poder ilustrar esta diferencia, analizaremos el desarrollo de los términos más relevantes a lo largo del tiempo y lo contextualizaremos en función del proceso electoral en el que se enmarcan.

3. METODOLOGÍA

Lo primero que se va a hacer en este trabajo es detallar el marco teórico en el que nos basamos. Para ello, se expondrán sucintamente tanto la historia y línea editorial de los diferentes medios como el ciclo electoral de 2023: procesos electorales, acontecimientos relevantes a lo largo de su desarrollo y conclusiones lógicas de los mismos. En la elaboración de este marco, nos apoyaremos en gráficos y en esquemas que permitan una mejor comprensión.

Una vez delimitado dicho marco teórico, procederemos a realizar el análisis de los datos mediante *text mining*. Estos datos provienen de diversas fuentes pero que pueden reconducirse a la misma: la hemeroteca de cada uno de los cinco periódicos analizados. Otra fuente empleada ha sido el espacio: "portadas del día", disponible en el periódico Europa Press. En este espacio se muestran las portadas de cada diario en función del día, lo que ha resultado especialmente útil para poder seleccionar las noticias de forma neutral.

Este análisis de *text mining* será eminentemente inductivo, de tal forma que a partir de la información disponible en el *corpus* se extraerán conclusiones lógicas, rigurosas y claras. Con carácter posterior, estas se compararán con la narración de los hechos que acontecieron durante el ciclo electoral. Esta metodología también buscará verificar, a través de la evolución de los *topics* en cada medio, su relación con la línea editorial y el tratamiento que cada uno de ellos realiza de los *topics* en términos cualitativos y cuantitativos. Este extremo puede definirse como análisis deductivo, en tanto que busca a través del análisis exhaustivo de los distintos medios verificar si en efecto, guarda relación con la línea editorial que la opinión pública le ha asignado.

Nuestro análisis se estructurará en varias etapas: en primer lugar, recopilación de todos los datos a partir de las fuentes de información detalladas, tomando como referencia el espacio portadas. Una vez realizado esto, convertiremos el texto bruto y no estructurado en un *dataset* coherente, integral y que permita el tratamiento del texto contenido. Posteriormente, configuraremos el modelo de *text mining*, indicando los parámetros requeridos para el análisis y realizando algunas visualizaciones previas.

Cuando hayamos configurado el modelo, pasaremos al preprocesamiento de los datos. Se trata de una tarea esencial para que nuestro análisis arroje conclusiones claras y

coherentes. Para ello, dividiremos todos los documentos[1] en *tokens* a los que se les asignará una categoría gramatical. Posteriormente, eliminaremos aquellos *tokens* que se encuentren entre las *stopwords* o que no respondan a las categorías gramaticales que nos interesa examinar[2] y crearemos la *document term matrix*, realizando diversos análisis de los términos y bigramas que integran la matriz.

Posteriormente, realizaremos el *topic analysis*, para el que valoraremos la complejidad y perplejidad de las diferentes opciones de *topics*, seleccionando un número de temas que, además de ser lógico o guardar relación con la narrativa, cumpla con las condiciones de complejidad y perplejidad. Tras realizar este estudio, podremos ver los términos más presentes en cada uno de los temas. A partir de estos términos, podremos ver la idea principal que busca arrojar cada tema en concreto.

Una vez hecho esto interpretaremos los resultados y las ideas que arroja cada uno de los temas, junto con sus pesos en cada una de las noticias. Así, veremos qué noticias se refieren más a cada tema en concreto y si los diferentes temas se relacionan con la narrativa expuesta en el marco teórico. Por añadidura, haremos lo mismo con cada medio de comunicación, observando la evolución de los *topics* en cada uno de ellos y las aportaciones que realizan a la conformación de estos. Mediante este estudio podremos sacar conclusiones sólidas acerca de la línea editorial, viendo si en efecto guarda relación con la percepción generalizada que se tiene sobre "la cuerda política" de cada uno de ellos.

[1] Véase que durante toda la realización del trabajo se hará referencia a la palabra documento o noticia de manera indistinta.
[2] Sin perjuicio de otros filtros de menor relevancia que serán posteriormente examinados de forma exhaustiva.

4. ESTADO DE LA CUESTIÓN

España es el cuarto país del mundo con más alarmas instaladas según datos del Observatorio Sectorial DBK, sólo detrás de China, Estados Unidos y Japón. De escuchar a nuestros medios y políticos, parecería que lo que se busca es poner fin a lo que se dibuja como un grave problema: la "ocupación" y los robos violentos. No obstante, al contrario de lo que pueda parecer, los datos muestran otra cosa: mientras que la ocupación, los robos y la criminalidad disminuyen, los españoles no dejamos de instalar cada vez más alarmas.

Desde el origen de la sociedad de la información, autores como Sartori (1999) vienen alertando de la paulatina transformación de la democracia a la telecracia con frases como la siguiente: "la televisión condiciona fuertemente el proceso electoral, ya sea en la elección de los candidatos, bien en su modo de plantear la batalla electoral o en la forma de ayudar al vencedor" (p. 66).

Estos condicionamientos e interrelaciones que sufren los medios de comunicación y los procesos electorales forman una pregunta que aunque parezca fácil de responder, esconde dentro de sí una realidad muy compleja: ¿de qué manera influyen estos medios de comunicación en nuestro estado de opinión? O, mejor dicho, ¿cómo logra la prensa persuadirnos para modificar algo tan personal e intransferible como nuestro voto?

Muchos autores han tratado de exponer, con poco consenso, las diferentes teorías por las que los medios de comunicación influyen en la construcción de la opinión pública. En este sentido, Lippmann (2003) hace un paralelismo con la ficción, en tanto que los medios de comunicación simplemente generan representaciones del entorno que en mayor o menor grado son obra de un individuo (p. 33). Por otro lado, Sartori (2005), defendió que en la creación de la opinión pública existen influyentes e influenciados, de manera que los procesos de opinión son una traslación de información de los primeros a los segundos, en tanto que la opinión se genera en pequeños núcleos de información (p. 176).

No obstante, de todas estas teorías la que más éxito ha tenido es la teoría de la fijación de la agenda, comúnmente conocida como *agenda setting*. Ideada a partir de un estudio de McCombs y Shaw (1972) se ha resumido y explicado en múltiples artículos. Parafraseando la obra de uno de sus ideólogos (McCombs, 2006), los medios seleccionan un conjunto de asuntos (agenda) hacia los que dirigen la atención de los espectadores. Esta agenda, que no es más que un conjunto de temas o atributos, es posteriormente jerarquizada en función de lo que los medios quieren que sea percibido con mayor o

menor importancia. Los temas más importantes son así repetidos de manera continua, en detrimento de los menores, que se mencionan menos o no se nombran en absoluto (p. 31-32).

A partir de ahí, los medios de comunicación consiguen homogeneizar las preocupaciones de todos los ciudadanos, con independencia de la tradicional división demográfica o socioeconómica. En definitiva y como indican Shaw y Martin (1992), lo que la agenda busca fundamentalmente es incrementar el consenso grupal en detrimento del consenso tradicionalmente asociado a cada grupo en función de su sexo, raza, nivel educativo o clase social, entre otras variables (p. 902).

Esta teoría no puede entenderse sin la cooperación necesaria de la teoría del *framing* o encuadre. El concepto de encuadre o enmarque fue definido, entre otros, por Entman (1993) como: "enmarcar es seleccionar algunos aspectos de la realidad percibida y resaltarlos en un texto comunicativo de modo que promueva: a) una definición concreta del problema; b) una interpretación causal; c) un juicio moral; y, d) una recomendación de tratamiento" (p. 52). Es decir, que el *framing* no es más que la manera en la que se describe un tema o acontecimiento, así como el esquema que este acaba generando en la mente del lector o espectador (Ardèvol-Abreu, 2015, p. 427).

Por concluir, mientras que la agenda es un conjunto de temas que los medios jerarquizan para presentar a sus lectores e influir en ellos, el encuadre es la manera en la que los mismos medios presentan los temas a sus lectores, homogeneizando las preocupaciones de todos ellos y logrando un consenso en algunos temas concretos.

Una vez respondida la pregunta de cómo consiguen los medios tener un efecto en la realidad (lo que indefectiblemente incluye la realidad política) y tomando como referencia las dos últimas teorías mencionadas llegamos a la pregunta capital de este trabajo: ¿cuál ha sido la agenda y el encuadre de la prensa española durante el ciclo electoral del año 2023?

El último barómetro del Centro de Investigaciones Sociológicas (en adelante, CIS) de octubre – noviembre 2023 tuvo como objeto la percepción ciudadana de los medios de comunicación. En una de las preguntas, los ciudadanos les asignaron una ideología concreta. Así, siendo uno la extrema izquierda y diez la extrema derecha, la media de los periódicos de nuestra muestra fue de: El Mundo 6.92, El País 4.16, El Confidencial 5.96, ABC 7.79 y El Español 6.41.

Nota / Medio	1	2	3	4	5	6	7	8	9	10	Media	Desviación típica
El Mundo	1.7	1.0	2.0	3.5	11.3	11.6	16.1	15.8	7.8	10.2	6.92	2.07
El País	14.0	7.1	14.1	13.2	13.2	6.5	5.8	4.3	1.1	2.8	4.16	2.35
El Confidencial	0.6	-	0.8	6.3	20.8	41.9	21.2	6.7	-	-	5.96	1.09
ABC	1.6	0.9	1.4	2.0	6.4	5.7	10.1	16.9	14.6	20.1	7.79	2.12
El Español	-	-	-	-	18.0	32.9	34.2	9.5	1.3	-	6.41	0.96

Tabla 1: Percepción ciudadana de la ideología de los periódicos objeto de estudio

Fuente: Elaboración propia a partir de datos del CIS (octubre-noviembre 2023)

Hasta ahí llega la percepción de los ciudadanos. No obstante, ¿se cumple esto realmente cuando analizamos las noticias? En particular, ¿se cumplen estas ideologías en época electoral, que es cuando la importancia de la agenda y el encuadre alcanza su punto más álgido?

A estas preguntas es a las que vamos a tratar, desde el estudio empírico y el rigor científico, dar respuesta a través del presente trabajo. Es esta la pregunta que a nuestro juicio queda por responder de forma directa en un estado de la cuestión que en muchos casos ha sido más estático que dinámico. En definitiva, buscamos examinar y verificar la agenda y el encuadre de los medios yendo más allá de lo que los ciudadanos creen, comprobándolo a través de los correspondientes modelos matemáticos.

CAPÍTULO II: MARCO TEÓRICO

Antes de proceder al análisis exploratorio, conviene dotar al estudio de un contenido teórico que permita entender sus conclusiones e implicaciones. Este se compone de dos partes fundamentales: a) una descripción del ciclo electoral de 2023; y, b) una definición de los periódicos objeto de estudio.

1. CICLO ELECTORAL 2023

Van der Eijk (1987), haciendo referencia al caso concreto de Países Bajos, define los ciclos electorales como fluctuaciones en los votos en distintos procesos electorales con una única variable independiente: su duración (p. 254 y 260). Estos ciclos están integrados por procesos, que para el diccionario panhispánico del español jurídico (s.f.) son: el "conjunto de actos concatenados y regulados por la legislación en materia de elecciones que, con intervención decisiva de los ciudadanos con derecho a voto, son realizados por la Administración especial en materia de sufragio (juntas, mesas) … desde la convocatoria hasta la resolución de las reclamaciones contra las actas de proclamación de los resultados".

En España, los procedimientos electorales se regulan en la Ley 5/1985, de 19 de junio, de Régimen Electoral General (en adelante, LOREG). Esto, sin perjuicio de posible regulación de las elecciones autonómicas por la ley electoral de cada comunidad autónoma[3].

El artículo 42.3 de la LOREG prevé que las elecciones locales y a las Asambleas Legislativas de Comunidades Autónomas cuyos Presidentes de Consejo de Gobierno no tengan expresamente atribuida por el ordenamiento jurídico la facultad de disolución anticipada se celebran el cuarto domingo de mayo del año que corresponda y el decreto de convocatoria se publica el quincuagésimo quinto día antes. Si bien es cierto que la mayoría de Presidentes poseen la facultad de disolución anticipada, hacen coincidir sus respectivas elecciones autonómicas con las locales[4], en lo que ya conocemos en la práctica como "elecciones autonómicas y municipales".

[3] En la práctica, todas excepto Cataluña cuentan con una ley electoral propia.
[4] Las excepciones a esta norma para el ciclo electoral de 2023 son: Andalucía, Castilla y León, Cataluña, Galicia y el País Vasco.

De esta manera, el 3 de abril se publicó el decreto de convocatoria de elecciones a todas las corporaciones locales y Asambleas Legislativas de las ciudades autónomas de Ceuta y Melilla, así como los decretos de convocatoria de elecciones a las Asambleas Legislativas de las Comunidades Autónomas de Aragón, Asturias, Cantabria, Castilla-La Mancha, Comunidad Valenciana, Comunidad de Madrid, Extremadura, Islas Baleares, Islas Canarias, La Rioja, Navarra y la Región de Murcia (Menéndez, 2023a).

El ciclo electoral se iniciaba con un mapa de poder dominado fundamentalmente por el Partido Socialista Obrero Español (en adelante, PSOE), que gobernaba ocho de las doce comunidades y tenía 22,329 concejales de los 67,515. Por otro lado, el Partido Popular (en adelante, PP) gobernaba tres de ellas, contando con 20,325 concejales. Asimismo, el PSOE ostentaba el Gobierno de la Nación bajo la Presidencia de Pedro Sánchez Pérez-Castejón.

Estos partidos gobernaban de diversas formas: bien en solitario o bien en coalición. Estas coaliciones eran diversas: en el caso de las del PSOE eran con partidos a su izquierda, singularmente Unidas Podemos (en adelante, UP) y en el caso del PP era con partidos a su izquierda y a su derecha, entre los que podemos destacar a Ciudadanos (en adelante, CS) a su izquierda y VOX a su derecha. El Partido Regionalista Cántabro (en adelante, PRC) gobernaba la comunidad que da origen a su nombre.

El mapa político previo puede resumirse[5] con las siguientes figuras:

[5] Nótese que por mayor claridad visual hemos decidido incluir un mapa de todas las comunidades autónomas en vez de limitarnos sólo a las que celebraban elecciones ese día.

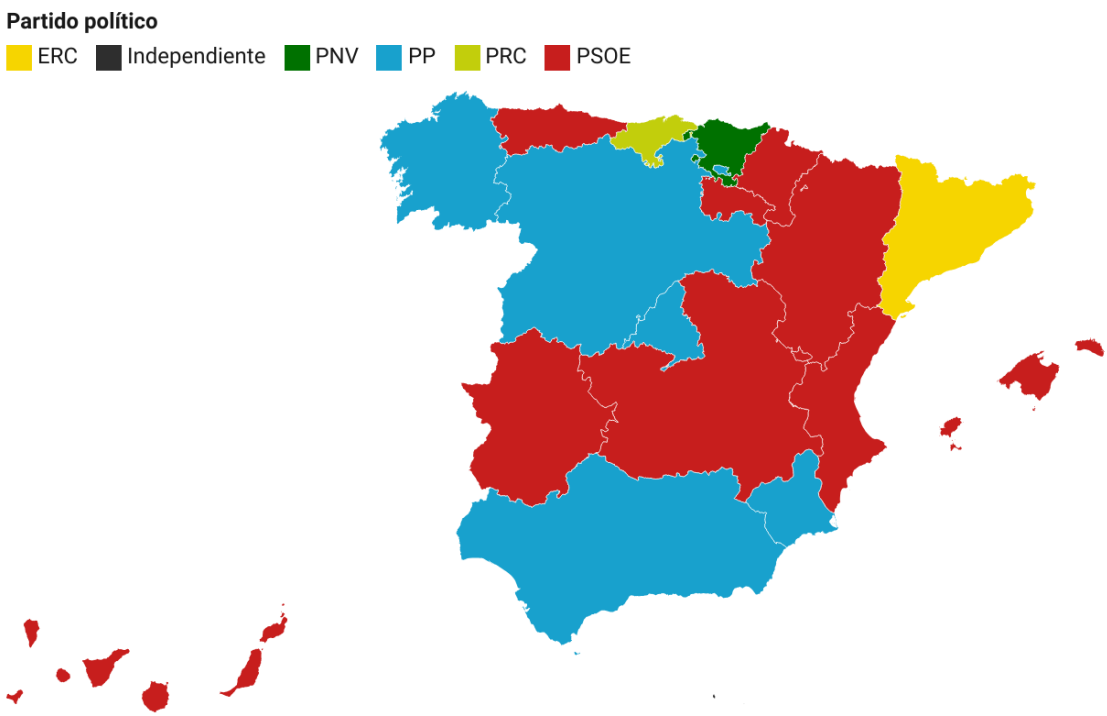


Figura 2: Mapa de los Gobiernos de las Comunidades antes del 28-M

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

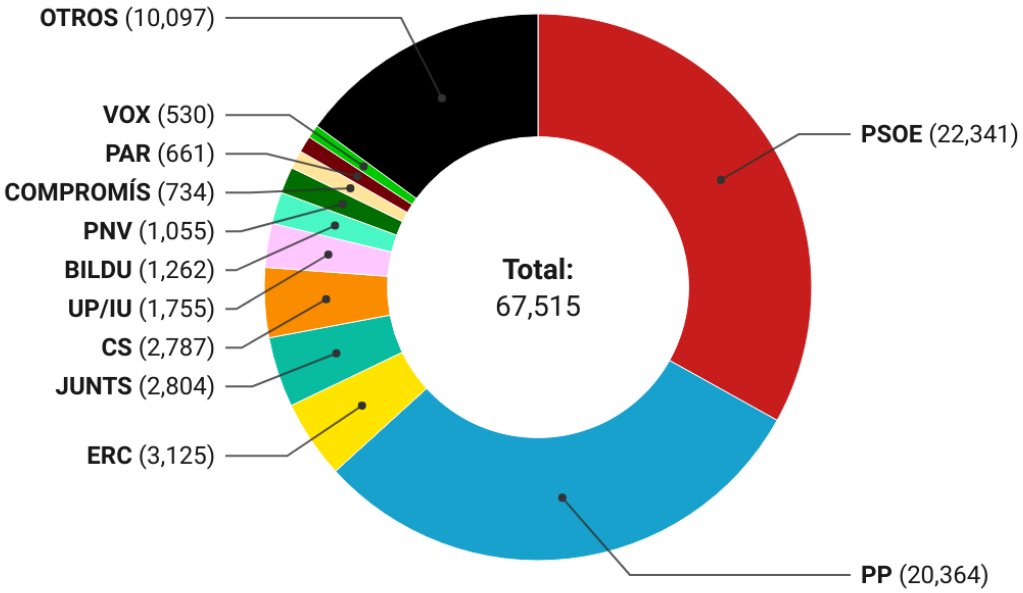


Figura 3: Concejales en las corporaciones municipales antes del 28-M

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

Pues bien, este fue el reparto de poder con el que dio comienzo el ciclo. La campaña estuvo marcada por un fuerte discurso en clave nacional y por determinados escándalos: condenados por delitos de sangre en listas electorales, presunto fraude electoral en Melilla, presunta compra de votos en Mojácar o incluso el presunto secuestro de una concejal en Maracena, entre otros (León, 2023). Con este telón de fondo, el 28-M más de 35 millones de ciudadanos fueron llamados a las urnas.

Los resultados de las elecciones fueron extremadamente positivos para el PP, que alcanzó su pico de poder a nivel territorial: pasaba a tener posibilidades de gobernar en Aragón, Cantabria, Comunidad Valenciana, Extremadura, Islas Baleares, La Rioja, Melilla y conservaba los gobiernos de la Comunidad de Madrid y la Región de Murcia. Por añadidura, podía facilitar el gobierno de las Islas Canarias. Municipalmente sus concejales se incrementaron en un 15%, pasando a superar los 23,000 y colocándose como primera fuerza. VOX, fuerza a la derecha del PP, incrementó notablemente su presencia, por lo que pasó a condicionar los futuros gobiernos de muchas de las citadas comunidades. Por otro lado CS, partido con el que el PP había cogobernado algunas comunidades, no obtuvo representación en casi ninguna de las circunscripciones a las que concurrió por no haber superado el umbral mínimo de entrada[6] (Menéndez, 2023b).

El PSOE y las fuerzas con las que tradicionalmente había venido pactando sufrieron, en términos generales, un fuerte retroceso. Este partido pasó a retener sólo tres comunidades: Asturias, Castilla-La Mancha y Navarra. Singularmente UP, fuerza con la que el PSOE compartía el Gobierno de la Nación, quedó fuera de casi todos los Parlamentos y Ayuntamientos en los que venía también cogobernando con el PSOE, por no superar tampoco el umbral mínimo de entrada (Menéndez, 2023b). Municipalmente, el PSOE perdió más de 1,500 concejales, así como el poder de una gran cantidad de corporaciones municipales (ej. Sevilla, Palma de Mallorca o Gijón, entre otros). No obstante, logró alzarse con el bastón de mando de la segunda ciudad más poblada de España: Barcelona.

[6] El umbral mínimo de entrada es el porcentaje mínimo para que los votos de una formación se traduzcan en escaños en una corporación municipal, Asamblea Legislativa de una Comunidad o ciudad autónoma o las Cortes Generales.

El nuevo mapa político puede resumirse con las siguientes figuras:

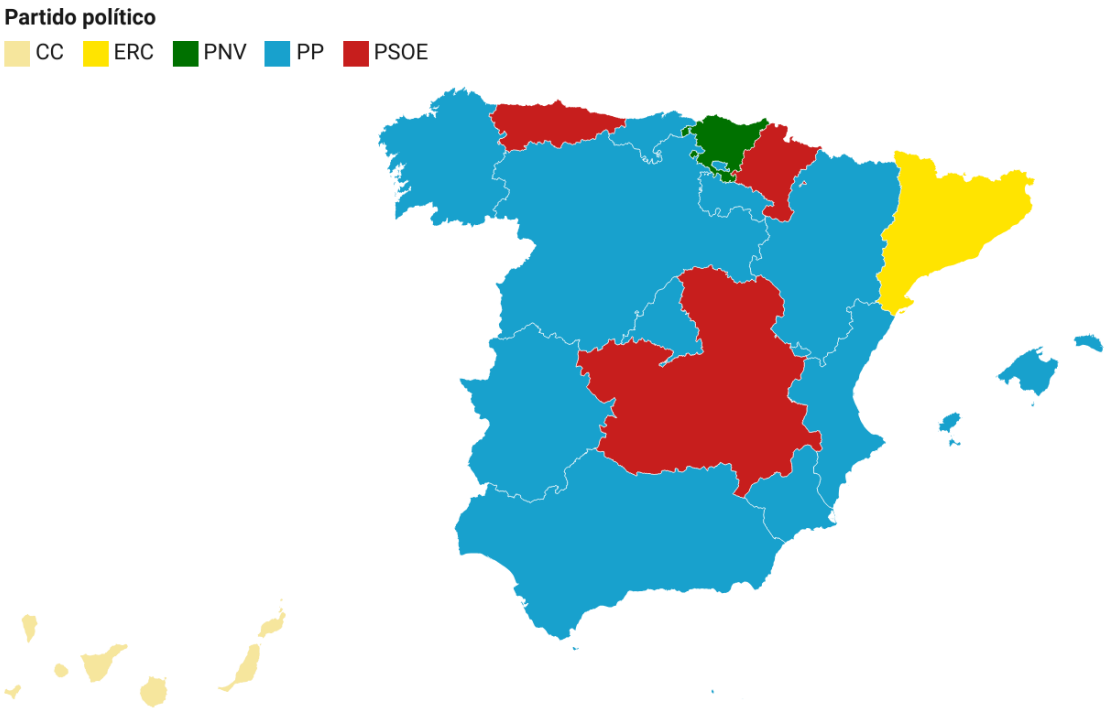


Figura 4: Mapa de los gobiernos de las Comunidades después del 28-M

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

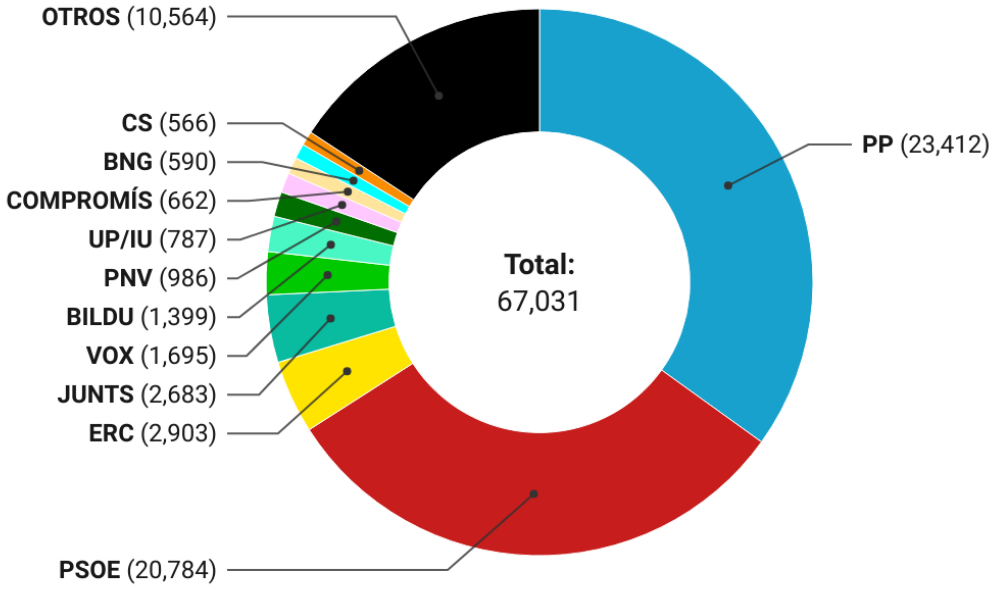


Figura 5: Concejales en las corporaciones municipales después del 28-M

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

La mañana después de conocer estos resultados (y de, en principio, dar por cerrado el ciclo hasta diciembre), el Presidente del Gobierno, Pedro Sánchez Pérez-Castejón, sorprendió a toda la ciudadanía con el siguiente anuncio: la disolución de las Cortes Generales y la correspondiente convocatoria de elecciones generales adelantadas para el 23 de julio de 2023 (EFE, 2023). Estas elecciones fueron convocadas por RD 400/2023, de 29 de mayo, de disolución del Congreso de los Diputados y del Senado y de convocatoria de elecciones.

A pesar de que el artículo cuatro del Decreto fijaba el inicio de la campaña el viernes 7 de julio, la realidad es que desde ese mismo momento la maquinaria de los partidos empezó a funcionar. La novedad de esta campaña fue que se desarrollaron tres procesos simultáneos: conformación de corporaciones municipales, conformación de gobiernos autonómicos y campaña de las elecciones generales.

Las principales cuestiones para resaltar de estas elecciones generales fueron las siguientes:

En primer lugar, los obstáculos para llegar a acuerdos entre el PP y VOX en las diferentes comunidades, que el primero trató de vender al electorado como geometría variable. Así, mientras acordaban con VOX el gobierno de algunas comunidades, también pactaban con el PRC en Cantabria o con Coalición Canaria (en adelante, CC) en la citada comunidad (Carreño, 2023). Algunos pactos reseñables fueron la entrega del bastón de mando del Ayuntamiento de Barcelona al PSOE, el "acuerdo exprés" con VOX en la Comunidad Valenciana o, en particular, el caso de María Guardiola, quien tras anunciar un veto a VOX acabó cediendo y dejándoles entrar en el Consejo de Gobierno de Extremadura, con las correspondientes consecuencias en términos de credibilidad.

En segundo término, el surgimiento del conocido como "frente amplio" de izquierdas, que cristalizó en una candidatura unitaria liderada por la Vicepresidenta Segunda Yolanda Díaz. La llamada "izquierda a la izquierda del PSOE", fulminantemente derrotada tras el 28-M, diagnosticó que su principal error había sido acudir a las elecciones por separado. En consecuencia y para tratar de reeditar el Gobierno de la Nación existente, se unieron en una lista única: Sumar. De los muchos partidos que confluyeron, el caso más llamativo fue el de Podemos, anterior "líder" del citado espacio, quien terminó por incorporarse a la coalición, si bien recriminando a Yolanda Díaz la exclusión de Irene Montero, en ese momento Ministra de Igualdad, de las listas electorales (Santana, 2023).

En tercera instancia, se plantearon problemas derivados de la fecha de la celebración de las elecciones. En concreto, al ser las primeras elecciones celebradas en verano, la oposición argumentó que podían existir problemas con el voto por correo o la conformación de las mesas electorales[7], debido a que muchos ciudadanos se encontraban de vacaciones. Estas reclamaciones fueron apoyadas por algunos sindicatos a causa de las condiciones laborales de los trabajadores de Correos. El Gobierno, como respuesta, acusó a la oposición y sindicatos de sembrar dudas acerca del correcto desarrollo del proceso electoral (Triviño, 2023). Finalmente, no hubo problemas ni en la conformación de mesas ni en el voto por correo, que alcanzó en estos comicios su máximo histórico.

En cuarto lugar, el controvertido papel de las encuestas durante la campaña electoral. España batió récord histórico de encuestas no sólo a nivel nacional sino también europeo: hasta 105 encuestas en las dos semanas previas a las elecciones, llegando a contabilizarse hasta siete *trackings* diarios (Vega, 2023). Estas encuestas, que en su mayoría fallaron en los resultados, condicionaron la campaña de forma clara, pues abocaban a un resultado que parecía inevitable: mayoría simple pero holgada del PP que, de coaligarse con VOX de la misma forma que ya habían hecho en Ayuntamientos y comunidades, resultaría en una holgada mayoría absoluta. Entre los factores que luego se infirieron como causantes del fallo demoscópico generalizado estuvieron la sobrerrepresentación del PP e infrarrepresentación de la izquierda.

Por último, el rearme del PSOE durante el proceso electoral. El PSOE, enormemente derrotado tras las elecciones de mayo y con un clima tanto demoscópico como social en contra de su secretario general y Presidente del Gobierno, sobre el que sobrevolaban acusaciones constantes de mentiras, se rearmó de cara a estos comicios. Este rearme incluyó, entre otras cosas: entrevistas en medios de comunicación de ideología diversa, apelaciones constantes a la posibilidad real de una remontada, campaña en favor de los cambios de opinión frente a las acusaciones de mentira y, especialmente, la necesidad de celebrar debates electorales (Monrosi, 2023). En cuanto a este último extremo, se celebraron dos debates: a) un "cara a cara" entre el PP y el PSOE, del que el consenso infirió que el PP salió victorioso; y, b) un debate entre el PSOE, Sumar y VOX, con la llamativa ausencia del PP. También aprovechó el Presidente la Presidencia española del

[7] La Junta Electoral Central acabó eximiendo de la obligación de acudir a las mesas electorales si se justificaba la ausencia del domicilio por causas como viajes o vacaciones siempre que se hubieran contratado antes del 29 de mayo.

Consejo de la Unión Europea, celebrada desde el uno de julio hasta el 31 de diciembre del mismo año.

Con todas estas líneas maestras expuestas, de nuevo más de 35 millones de españoles fueron llamados a las urnas en un caluroso domingo de julio. Si bien es cierto que el recuento del voto exterior alteró un escaño en el Congreso de los Diputados (del PSOE pasó al PP), el resultado fue demoledor para el que hasta entonces había venido siendo investido Presidente por todas las empresas demoscópicas: el PP. Obtuvo 137 escaños y mayoría absoluta en el Senado que, si bien fue una victoria, no fue suficiente para alcanzar la mayoría absoluta con VOX, quien obtuvo 33 escaños. Esta potencial coalición sólo sumaba 170 escaños, seis por debajo de la citada mayoría que se encuentra en 176.

Por otro lado, el PSOE se alzó como el victorioso de la jornada electoral: se incrementaban sus votos, se mantenían sus escaños pero, sobre todo, era el único capacitado para articular una mayoría favorable a su investidura con los resultados conocidos. El único problema: necesitaba del sí de *Junts per Catalunya*, formación política contraria a su Gobierno y que exigía la amnistía como condición para otorgar su voto para investir a Pedro Sánchez.

La evolución del mapa político[8] en España fue la siguiente:

[8] Excluimos de aquí lo relativo al Senado, pues es una cámara no relevante a efectos de lo que nos queda por examinar del ciclo: investidura, pactos… No obstante, es conveniente resaltar que el PP obtuvo 120 de los 208 senadores que se repartían esa noche. El PSOE recibió 72.

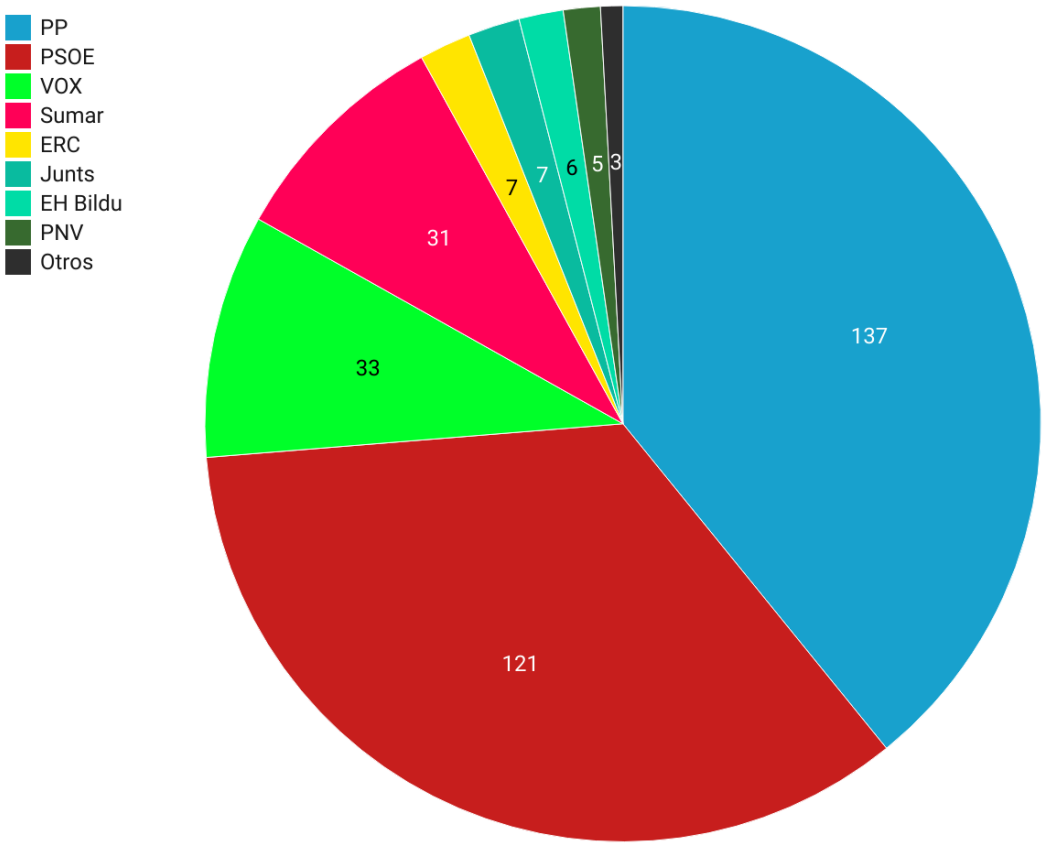


Figura 6: Resultado de las elecciones generales del 23J al Congreso

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

	Escaños 2019	Escaños 2023	Variación
PP	89	137	+48
PSOE	120	121	+1
VOX	52	33	-19
Sumar / UP *	35	31	-4
CS	10	**	
ERC	13	7	-6
Junts	8	7	-1
EH Bildu	5	6	+1
PNV	6	5	-1
BNG	1	1	=
CC	2	1	-1
UPN	2	1	-1
Más País	3	***	-
Teruel Existe	1	0	-1
CUP-PR	2	0	-2
PRC	1	**	-

Tabla 2: Resultados de las elecciones generales del 23J al Congreso y su comparación con 2019

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

* Sumar es el partido que en 2023 vino a ocupar el espacio que desde 2015 venía representando UP.

** Renunció a concurrir a estas elecciones.

*** Concurrió a estas elecciones con Sumar.

Con estas nuevas cartas, comenzó la tercera y última fase del ciclo electoral: los pactos y las negociaciones, con tres fases definitorias: la elección de la Mesa del Congreso, el intento de investidura de Alberto Núñez Feijóo y la investidura de Pedro Sánchez.

La primera etapa abarcó hasta el día 17 de agosto de 2023. El RD de convocatoria electoral, en su artículo quinto, estipulaba que el 17 de agosto se reunirían las Cámaras Legislativas en sesión constitutiva. Por lo tanto, ese fue el día de elección de la Mesa del Congreso.

Tras unas negociaciones discretas, finalmente el PSOE logró alcanzar la Presidencia del Congreso con los votos de la posteriormente bautizada como "mayoría de investidura"[9]. La nueva Presidenta pasó a ser Francina Armengol (Valero, 2023) y se abrió la segunda etapa de esta parte del ciclo: la investidura de Alberto Núñez Feijóo, candidato del PP.

El Rey, de acuerdo con el artículo 99 de la Constitución y tras la preceptiva consulta a todos los grupos parlamentarios, propuso a Alberto Núñez Feijóo como candidato a la investidura el día 22 de agosto. La nueva Presidenta del Congreso, de común acuerdo con el candidato, fijó el debate de investidura los días 26 y 27 de septiembre de 2023 (Marchante, 2023).

Durante este mes, sucedieron los siguientes acontecimientos reseñables:

En primer lugar, la caída de Luis Rubiales, presidente de la Real Federación Española de Fútbol. Tras la impecable victoria de la selección española de fútbol femenino en el mundial a finales del mes de agosto, Luis Rubiales dio un beso (que posteriormente se revelaría no consentido) a Jennifer Hermoso, jugadora de la selección. Este beso provocó una respuesta inmediata por parte de la sociedad que terminó con su dimisión, no sin antes haber tratado de defender su inocencia (Domènech, 2023).

En segundo término, la aprobación de la modificación del Reglamento del Congreso de los Diputados por la que se permite hablar en las lenguas cuyo carácter sea oficial en otras comunidades o partes del territorio nacional, singularmente catalán, euskera y gallego. El compromiso alcanzado por el PSOE con otras formaciones para lograr la presidencia de la Mesa del Congreso hizo imperativa una modificación del Reglamento para que pudiesen utilizarse las lenguas cooficiales en las cámaras, ocasionando quejas por parte del PP y de VOX (Esteve, 2023).

En tercera instancia, la entrada en Telefónica de la empresa saudí STC. En un aleatorio día de septiembre, la prensa abrió con el anuncio de que la empresa saudí STC se hacía

[9] Por mayoría de investidura entendemos el grupo de partidos formado por: PSOE, Sumar, ERC, Junts, EH Bildu, PNV y BNG, que como puede comprobarse en la Tabla 2, suman 178 de los 350 escaños del Congreso.

con el 9.9% de Telefónica, pasando a convertirse en el mayor accionista de la compañía. La reacción de los políticos fue inmediata, con algunos exigiendo la entrada del Estado en Telefónica como consecuencia de la ausencia de mecanismos de control directos de los que disponía el Gobierno para frenar una entrada como la de STC (Díaz, 2023).

Por último, el inicio de las movilizaciones del PP y VOX contra la amnistía. A medida que avanzaban los días y especialmente tras la constitución de la Mesa del Congreso, cada vez quedaba más clara una cosa: se iba a conceder la amnistía a aquellos que la solicitaban, a pesar de que el PSOE mantenía el más absoluto silencio al respecto (Soto y García, 2023). Ante la falta de mayoría parlamentaria y con el supuesto respaldo de la mayoría social, el PP y VOX iniciaron movilizaciones en contra de la amnistía dos días antes de la investidura de Feijóo, que se mantuvieron en el tiempo durante el resto del ciclo.

Como ya venía advirtiéndose, Feijóo no obtuvo la confianza de la Cámara en los dos intentos de investidura. Por ello y tras una nueva ronda de contactos, el Rey propuso a Sánchez como candidato (Lardiez, 2023), abriendo la última etapa del ciclo electoral: la investidura de Pedro Sánchez Pérez-Castejón.

	PP	**PSOE**	**VOX**	**Sumar**	**ERC**	***Junts***	**EH Bildu**	**PNV**	**BNG**	**CC**	**UPN**	**Total**
Si	137		33							1	1	172
No		121		31	7	7	6	5	1			178
Abstención												0

Tabla 3: Votos emitidos en la investidura de Alberto Núñez Feijóo[10]

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

Esta última etapa estuvo marcada por las siguientes cuestiones:

En primer lugar, la amnistía y los acuerdos. Finalmente el PSOE confirmó lo que en anteriores etapas hemos venido introduciendo: la amnistía. El Presidente del Gobierno,

[10] En el segundo intento se repitió este mismo resultado, con la salvedad de que un voto de Junts fue declarado nulo, por lo que habría 172 síes, 177 noes y un voto nulo, así como cero abstenciones.

en un mitin, afirmó que, en efecto, se iba a aprobar una amnistía como consecuencia de los acuerdos con las fuerzas nacionalistas para lograr conformar un gobierno. Esto provocó un aumento de las movilizaciones, especialmente a medida que los acuerdos con las diferentes fuerzas iban anunciándose, culminando con movilizaciones ante la sede del PSOE en el mes de noviembre (Huesca, 2023).

Por último, el ataque terrorista de Hamás en Israel y la posterior reacción israelí. El 7 de octubre, la banda terrorista Hamás lanzó un ataque a gran escala contra Israel, desatando un conflicto que acaparó toda la atención de la prensa nacional e internacional. El país hebreo, como respuesta, lanzó una invasión a la franja de Gaza, acaparando también buena parte de la atención de la prensa de todos los países y dividiendo las opiniones entre los que lo creían con derecho a la legítima defensa y los que abogaban por un alto al fuego (CNN Español, 2023).

En medio de un complejo contexto social y político, el 16 de noviembre de 2023, Pedro Sánchez Pérez-Castejón fue investido de nuevo Presidente del Gobierno, al obtener la confianza de la mayoría absoluta del Congreso de los Diputados.

	PP	**PSOE**	**VOX**	**Sumar**	**ERC**	***Junts***	**EH Bildu**	**PNV**	**BNG**	**CC**	**UPN**	**Total**
Si		121		31	7	7	6	5	1	1		179
No	137		33								1	171
Abstención												0

Tabla 4: Votos emitidos en la investidura de Pedro Sánchez Pérez-Castejón

Fuente: Elaboración propia a partir de datos del Ministerio del Interior

Como último colofón a este ciclo electoral, cabe destacar la poca presencia de noticias económicas. Salvo contadas excepciones como Telefónica, el pago de peajes o los indicadores económicos que se emiten con independencia de si hay o no elecciones (por ejemplo, los datos de paro al inicio de cada mes), la economía pasó muy desapercibida durante la mayoría del ciclo.

2. PERIÓDICOS OBJETO DE ESTUDIO

En este último apartado se pretende exponer sucintamente tanto la historia como la línea editorial de cada uno de los medios objeto de estudio:

Diario El Mundo: Inicialmente denominado "El Mundo del Siglo XXI", fue fundado en 1989 por Alfonso de Salas, Pedro J. Ramírez, Balbino Fraga y Juan González. Su consolidación como diario vino tras revelar una serie de casos de corrupción relacionados con el PSOE en la década de los 90. Perteneciente a Unidad Editorial S. A., es el segundo periódico más leído en España. Su línea editorial fue definida por su director actual, Joaquín Manso, como comprometida con la sociedad abierta; por lo tanto, comprometida con la pluralidad, tolerancia, moderación y compromiso institucional (Doménech, 2023). Se centran mucho en la transparencia y separación de poderes. En la escala ideológica, su propio director lo reconoce como un periódico con vocación de centralidad pero cierta tendencia hacia la derecha, con algunas voces (aunque pocas) de la órbita de la izquierda.

Diario El País: Este diario, que es el más leído de toda España, vio la luz en 1976, medio año después de la muerte de Franco y cuatro años después de que José Ortega Spottorno lo fundara junto con Jesús de Polanco y Juan Luis Cebrián (Sabés Turmo, 2023, p. 20). Perteneciente al grupo PRISA, ha sido definido desde sus inicios como un periódico político pero independiente, alineado con valores socialistas entendidos como valores socialdemócratas europeos. Se define como un periódico liberal, independiente, socialmente solidario, europeo y latinoamericano. En el Anexo disponible en su Estatuto de Redacción puede verse claramente como aboga por: "el desarrollo, perfeccionamiento y buena administración de la Seguridad Social y el reparto justo de la riqueza a través del juego acertado de los impuestos" (El País, 1977). En los últimos años ha sufrido acusaciones de periódico "oficialista"[11].

El Confidencial: Fundado en el año 2001 por José Antonio Sánchez, Jesús Cacho y Antonio Casado, entre otros, desde sus inicios ha sido un periódico totalmente digital, no habiendo contado nunca con versión en papel (Sánchez, 2020). Se destaca por estar comprometido con los valores democráticos, entre los que priman la independencia y la

[11] La RAE define oficialismo como: "aceptación por principio de lo que establece el poder oficial" (RAE, s.f.), entendiendo este como el Gobierno.

veracidad. Está especializado en información relativa al mundo de la empresa y la economía, habiendo informado de varias exclusivas como la quiebra y puesta en venta del Banco Popular. Califican su línea editorial como independiente y plural, tratando de dar cabida a opiniones diversas (Herrero-Beumont, 2024). Su principal foco es, además de la citada información económica, el periodismo de investigación o de exclusiva, con menos foco en la actualidad política entendida como el "día a día" de los políticos. Actualmente es de los diarios digitales de mayor difusión.

Diario ABC: Fundado en 1903 por Torcuato Luca de Tena, se trata de uno de los periódicos más antiguos. Fue pionero en incorporar las crónicas telegráficas, que vinieron incorporándose desde 1905 y fue el primer diario nacional en dar el salto a internet el 20 de septiembre de 1995 (Ayala Sörenssen, 2023). Pertenece en la actualidad al Grupo Vocento. Es de tradición conservadora, fiel a la monarquía, al Magisterio de la Iglesia y a la unidad nacional. Ha sido defensor de la Inmaculada, la Navidad o la Semana Santa (Arroyo Cabello, 2006, pp. 21 y 22) y destaca por una fuerte oposición a cualquier tipo de nacionalismo o postulado de izquierdas, llegando a ser percibido por los ciudadanos como el periódico de difusión nacional más "de derechas". En el último Estudio General de Medios fue el séptimo diario más leído, detrás de "La Vanguardia" y delante de "Mundo Deportivo" (Asociación para la Investigación de Medios de Comunicación, 2024).

El Español: Fundado el 14 de octubre de 2015 por quien fuera fundador de "El Mundo" Pedro J. Ramírez, es el más joven de todos los periódicos analizados (Del Arco Bravo et al., 2016, pp. 528 y 529). Nació a través de las redes sociales y desde sus orígenes ha seguido una línea editorial pública y transparente, concretada en sus treinta valores (o como lo denominan los redactores, sus treinta "obsesiones"). Entre estas destacan algunas asociadas al conservadurismo como la limitación de competencias autonómicas, el adelgazamiento de la Administración o la bajada de impuestos. También están presentes otras más cercanas al progresismo entre las que se encuentran la racionalización de horarios, lucha contra la violencia en el ámbito doméstico o igualdad salarial. Por último, también muestra compromiso con valores "despolitizados": jueces independientes, reducción del paro o defensa de los valores cívicos y la lengua española (El Español, 2016). Su difusión es totalmente digital.

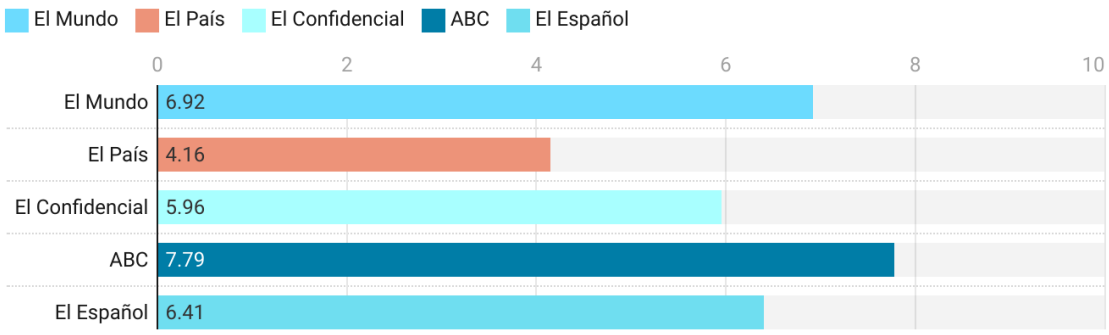


Figura 7: Percepción ciudadana de la ideología de los periódicos objeto de estudio

Fuente: Elaboración propia a partir de datos del CIS (octubre-noviembre 2023)

Escala: del 0 (más a la izquierda) al 10 (más a la derecha)

CAPÍTULO III: ANÁLISIS EXPLORATORIO DE LOS TEMAS MÁS PRESENTES DURANTE EL CICLO ELECTORAL 2023

En este capítulo, se va a realizar el análisis exploratorio de las noticias durante el ciclo electoral de 2023. De esta manera, veremos cuál es el paisaje global que los medios dibujan durante el período electoral y cuál es la agenda de cada uno de ellos en particular.

1. INTRODUCCIÓN AL *TEXT MINING*

Como ya se ha venido mencionando en reiteradas ocasiones a lo largo de este trabajo, el análisis exploratorio que se realiza es un análisis de *text mining*.

El *text mining* es una variación del campo de la minería de datos que intenta encontrar patrones interesantes o relevantes en bases de datos grandes, permitiendo el manejo de documentos no estructurados cuya estructura inicial no está definida, como sucede en nuestro caso con las noticias. El objetivo de esta disciplina es la de descubrir información desconocida y no plasmada previamente (Vijay Gaikwad et al., 2014, p. 42).

Por las dificultades que ofrece el análisis de textos, los procesos de *text mining* han de seguir una serie de pasos predeterminados y claros, de manera que el análisis pueda extraer conclusiones rigurosas e interpretables. En este sentido, los principales pasos que deben tenerse en cuenta son (Vijay Gaikwad et al., 2014, pp. 43 y 44; Vijayarani et al., 2015, pp. 8-11):

Localización y manejo de la información: Paso capital para poder trabajar con esta metodología. Conocido en inglés como *information retrieval*, es la asociación y recuperación de información de un número amplio de documentos de texto. El objetivo fundamental es la construcción del *corpus*. El *corpus* es el conjunto de documentos de texto, noticias o, en definitiva, datos no estructurados cuya finalidad es la de ser utilizados en el modelo de *text mining*. Puede entenderse como la base de datos del estudio. No obstante, para poder trabajar con dicha base esta debe ser manejable y coherente, de manera que así puedan aplicarse modelos de procesamiento de lenguaje natural u otros, garantizando la fiabilidad e interpretabilidad de los resultados.

Análisis preliminar: Este paso, conocido en inglés como *information extraction*, permite solventar el problema de transformar un *corpus* de texto en una base de datos estructurada. Generalmente se trabaja con visualizaciones previas de los datos en bruto y simplifica

mucho el trabajo del usuario, pues infiere relaciones entre los datos que permiten un mejor manejo. Algunos autores incluyen como paso dentro de esto la *tokenización,* que es la separación de filas de texto en palabras o conjuntos de palabras individuales, aunque para nosotros esto se encuadra más en la esfera del preprocesamiento de texto.

Preprocesamiento del texto: Al ser el *text mining* una disciplina que estudia fundamentalmente datos no estructurados, el preprocesamiento de texto es un paso muy importante y el primero en el análisis de *text mining* propiamente dicho. Dentro de este paso, se incluyen tareas como: *tokenización*, transformaciones del texto, segmentación de frases, asignación de *parts-of-speech*, eliminación de *stopwords* o palabras que aportan poco o ningún contenido informativo y lematización de las palabras. Es, en resumen, el proceso de preparar tu texto para ser analizado.

***Topic analysis*:** El *topic analysis*, a diferencia de la categorización de documentos, se emplea para buscar agrupamientos de documentos con contenidos similares de forma natural, sin prefijar el contenido de los diferentes *topics* o temas. Se le conoce también como *clustering*. Los documentos de cada *topic* tienden a ser homogéneos dentro del mismo *cluster* y heterogéneos con respecto a los documentos de los diferentes *clusters*. Así, pueden estudiarse de forma separada los contenidos del *corpus*, extrayendo conclusiones y patrones que, de haberse realizado correctamente los pasos anteriores, serán interpretables y coherentes.

Visualización: Las técnicas de visualización, que son independientes y pueden realizarse en cualquier momento de la investigación, permiten simplificar el descubrimiento de información relevante. Para ello se emplean colores, jerarquías visuales o mecanismos de interacción para que el usuario pueda trabajar con el documento a través de técnicas como el *zoom*.

Por consiguiente, estos son los pasos que se van a seguir en el estudio del ciclo electoral de 2023. El primero de ellos comprende la selección de noticias. Para ello, se ha seleccionado la portada de la edición matutina de los cinco diarios durante todos los días desde el 12 de mayo de 2023 al 16 de noviembre de 2023, lo que arroja un resultado total de 945 noticias. Posteriormente, se ha trabajado con la muestra para lograr una base de datos (*corpus*) con estructura lógica e interpretable, lo que ha permitido un análisis preliminar sólido y coherente. Seguidamente, se ha procesado el texto con las siguientes técnicas: separación de frases, *tokenización*, eliminación de *stopwords*, lematización y

asignación de *parts-of-speech*. Por último, se ha realizado el *topic analysis* tanto en términos generales, como focalizado en cada uno de los medios en particular.

2. LOCALIZACIÓN Y MANEJO DE LA INFORMACIÓN

Como ya se ha expuesto, el *corpus* del presente trabajo está compuesto por 945 noticias recogidas durante el periodo del 12 de mayo al 16 de noviembre de 2023. Estas noticias provienen de cinco diarios diferentes: El Mundo, El País, El Confidencial, ABC y El Español. Para llevar a cabo esta recopilación se ha seleccionado la portada de la mañana de cada uno de estos diarios. Por los evidentes cambios que sufre la página web de un periódico a lo largo de la mañana, nos hemos apoyado en Europa Press, que ofrece una vista de las portadas de los diarios para cada día.

El análisis se ha realizado en Python. Si bien es cierto que durante los estudios de grado la metodología de *text mining* se imparte en R, el lenguaje finalmente escogido ha sido Python. La razón es la siguiente: si se comparan ambos lenguajes, Python tiene una capacidad de ejecución más rápida, sumada a una mejor comprensión de datos brutos y mejores clarificaciones visuales (Bhanot et al., 2019). La principal librería será "*spaCy*", al ser una librería preparada para el procesamiento de lenguaje natural que sirve de forma adecuada a los objetivos del trabajo.

Para poder realizar la recogida de la muestra, el primer problema con el que nos topamos fueron los muros de pago. Entendemos por muro de pago (Ionos, 2022): "Una barrera de pago digital configurada por los editores para cierto tipo de ofertas digitales. Los usuarios solo pueden acceder al contenido que hay detrás tras pagar una cuota o contratar una suscripción". Si bien es cierto que existen distintos tipos de muros, el denominador común era claro: todas las portadas estaban sometidas a una suscripción previa. A la hora de enfocar este problema, hemos optado por contratar la suscripción de los distintos medios y realizar el volcado de las noticias durante el tiempo en que la misma fue contratada (en este caso, un mes).

Una vez suscritos a los cinco periódicos, comenzamos con el volcado a través de sus correspondientes hemerotecas. Para ello, hemos seleccionado o insertado las siguientes cuestiones relativas a cada noticia, que luego acabarán conformando las columnas de la base de datos o *corpus*:

Nombre del medio: En función del periódico, hemos insertado de forma manual su nombre.

Fecha de la noticia: De nuevo para cada día del ciclo, hemos añadido la fecha en una ocasión, después de la que se han insertado las cinco noticias del día, una para cada medio.

Título de la noticia: El título de cada noticia ha sido también insertado en el documento. Por título hemos entendido aquella frase o conjunto de frases que rotulan la noticia.

Subtítulo o subtítulos de la noticia: Todos los medios seleccionados contaban con un subtítulo para cada noticia, llegando a ser dos o incluso tres en el caso del Diario ABC. Por subtítulo entendemos aquella frase o conjunto de frases que acompañan al título, generalmente con una fuente más reducida y que sirven para aclarar o enunciar cuestiones del cuerpo. En este sentido, hemos seleccionado todos los subtítulos disponibles en cada noticia.

Cuerpo de la noticia: El cuerpo de la noticia también ha sido seleccionado. Sin embargo y por la extensión de algunas noticias, se ha optado por limitarlo a una extensión razonable en función del medio, pues algunos tienen párrafos más largos que otros. En general, el cuerpo ha consistido en los dos primeros párrafos en el caso de El Mundo y ABC, uno o dos párrafos en las noticias de El País y El Confidencial y los tres primeros párrafos de las noticias de El Español.

Toda esta información fue almacenada en un documento Word, que contaba con la siguiente estructura:

12/05/2023

El Mundo: **Sánchez y Feijóo confrontan dos modelos de país en su primer duelo electoral**

Preludio del asalto final que se librará en diciembre, las municipales y autonómicas del 28-M abren camino. Los españoles empiezan a optar: continuismo o cambio de ciclo.

Las elecciones municipales y autonómicas del 28-M serán determinantes para proyectar el futuro color político del país. Son el preludio del combate entre dos modelos: el de la izquierda, encabezado por un PSOE de la mano de populistas y secesionistas, y el de la derecha, liderado por un PP encaminado a compartir poder con el extremismo de Vox. Las posibilidades de que uno u otro, sin aditivos, se hagan con el triunfo se limita a muy pocos territorios y ciudades. PP y PSOE batallan por defender los feudos que ocupan al tiempo que intentan arrebatar al rival el poder en bastiones clave, aunque sea por la mínima y con pactos a veces muy poco deseados.

El País: **Empieza la campaña electoral de la incertidumbre**

El CIS revela que la clave local tendrá mucho peso el 28-M, pese a que la lectura

Figura 8: Ejemplo de estructura original del *corpus*

Fuente: Elaboración propia a partir del estudio de campo

De este trabajo resultó un documento de 517 páginas que contaba con las 945 noticias. Sin embargo, el documento no contaba con una estructura que permitiese su manipulación, por lo que procedimos a transformarlo a un formato ".xlsx". Para ello, empleamos expresiones regulares y saltos de línea. Por ejemplo, para la fecha utilizamos la expresión r'\d{2}/\d{2}/\d{4}' que nos devolvería una fecha del tipo: día/mes/año. No obstante, los resultados no fueron satisfactorios, seguramente por la enorme variedad de caracteres que impedían una correcta separación y la estructura manual fruto del volcado manual de noticias.

Por ejemplo, había veces que en vez de aparecer el nombre del medio en la columna específicamente designada para ello, aparecía el titular de la noticia, lo que impedía conocer su medio (ver Figura 9). Asimismo, en otros casos, el nombre del medio y el titular aparecían de forma conjunta, sin separación y en una columna diferente a la que le correspondería a cualquiera de los dos (ver Figura 10, donde aparecen en la columna subtítulo).

20/05/2023	En el consistorio valenciano se sientan 33 concejales. La mayoría a
21/05/2023	El Mundo
21/05/2023	«Es un hombre bonachón que ha saneado las cuentas y que se ha
21/05/2023	Un puñado de votos decidirán si el próximo gobierno de la Genera

El País: Sánchez decidió el adelanto electoral de las generales la noche del domingo y la comunicó a unos pocos fieles
El Confidencial: El PSOE se revuelve contra el adelanto electoral de Sánchez: "Ha vuelto a secuestrarnos"
ABC: Los barones echan la culpa al presidente: «Pedro nos ha hundido»

Figuras 9 y 10: Ejemplos de errores en el volcado de datos

Fuente: Elaboración propia a partir del estudio de campo

Tras intentar subsanar tanto los errores citados como otros, sin ningún éxito, especialmente por la ya mencionada falta de estructura interna derivada del volcado de noticias, se hizo imperativo tener que realizar el cambio de Word a Excel de modo manual. En este sentido, partimos del volcado con errores, pues algunas noticias si se encontraban bien estructuradas. A partir de ahí, se realizó un volcado manual que derivó en un archivo ".xlsx". Este archivo contenía los 945 registros y las siguientes columnas: *Date, Newspaper, Title, Subtitle y Body*. A partir de este *corpus*, puede iniciarse un análisis coherente.

Date	Newspaper	
12/5/23	El Mundo	Sánchez y Feijóo confrontan dos modelos de país en su primer duelo electoral
12/5/23	El País	Empieza la campaña electoral de la incertidumbre
12/5/23	El Confidencial	El 28-M prueba el aguante de Sánchez frente a un Feijóo que necesita imponer el "cambio de ciclo"
12/5/23	ABC	La Ley de Partidos impide candidaturas con condenados por terrorismo
12/5/23	El Español	Feijóo arranca la campaña ligando al PSOE con Bildu, Sánchez exhibe gestión y agenda en Washington
13/5/23	El Mundo	Sánchez, en la Casa Blanca sobre las listas de Bildu: "Hay cosas que pueden ser legales pero no son decentes"

Figura 11: Primeras líneas del *corpus*

Fuente: Elaboración propia a partir del estudio de campo

3. ANÁLISIS PRELIMINAR

Antes de comenzar con el preprocesamiento del texto, conviene realizar una serie de visualizaciones previas acerca del mismo. Por tanto, hemos cargado el *corpus* en Google *collab* para realizar la programación en su totalidad, empezando por este análisis preliminar. A continuación, hemos creado una nueva columna llamada "*All*", en la que se ha añadido la suma del título, subtítulo y cuerpo de la noticia (es decir, toda la noticia objeto de análisis):

	Date	Newspaper	Title	Subtitle	Body	All
0	2023-05-12	El Mundo	Sánchez y Feijóo confrontan dos modelos de paí...	Preludio del asalto final que se librará en di...	Las elecciones municipales y autonómicas del 2...	Sánchez y Feijóo confrontan dos modelos de paí...
1	2023-05-12	El País	Empieza la campaña electoral de la incertidumbre	El CIS revela que la clave local tendrá mucho ...	La política española se ha sumergido en un mar...	Empieza la campaña electoral de la incertidumb...
2	2023-05-12	El Confidencial	El 28-M prueba el aguante de Sánchez frente a ...	Moncloa toma las riendas de la campaña ante la...	Ni Pedro Sánchez, ni Alberto Núñez Feijóo se p...	El 28-M prueba el aguante de Sánchez frente a ...
3	2023-05-12	ABC	La Ley de Partidos impide candidaturas con con...	El Gobierno y la Fiscalía son los únicos que p...	La Ley de Partidos, que salió adelante en 2002...	La Ley de Partidos impide candidaturas con con...
4	2023-05-12	El Español	Feijóo arranca la campaña ligando al PSOE con ...	El líder del PP protagoniza una agenda maraton...	Alberto Núñez Feijóo arrancó la campaña electo...	Feijóo arranca la campaña ligando al PSOE con ...
...	...	...	...	...	...	...
940	2023-11-16	El Mundo	Pedro Sánchez, investido presidente con la ame...	Revalida el Gobierno con 179 votos a favor, in...	179 votos a favor, 171 en contra y ninguna abs...	Pedro Sánchez, investido presidente con la ame...
941	2023-11-16	El País	Pedro Sánchez, presidente del Gobierno por ter...	La mayoría amplia que avala al líder del PSOE ...	Le costó mucho tiempo lograr que se lo tomaran...	Pedro Sánchez, presidente del Gobierno por ter...

Figura 12: Representación del *corpus* con la nueva columna "*All*"

Fuente: Elaboración propia

De cara a las visualizaciones previas, hemos creado una lista de columnas llamada "TEXT_COLUMNS" en la que hemos incorporado el título, subtítulo, cuerpo y la columna "*All*". Posteriormente y a través de un bucle, hemos representado el número de palabras de cada una de ellas:

Título: La mayoría de las noticias tienen entre 10 y 21 palabras. Hay una noticia con dos (un editorial de ABC llamado "NO DEBEMOS") y diez noticias que superan las 25 palabras, con una noticia de El País relativa a las municipales alcanzando las treinta palabras.

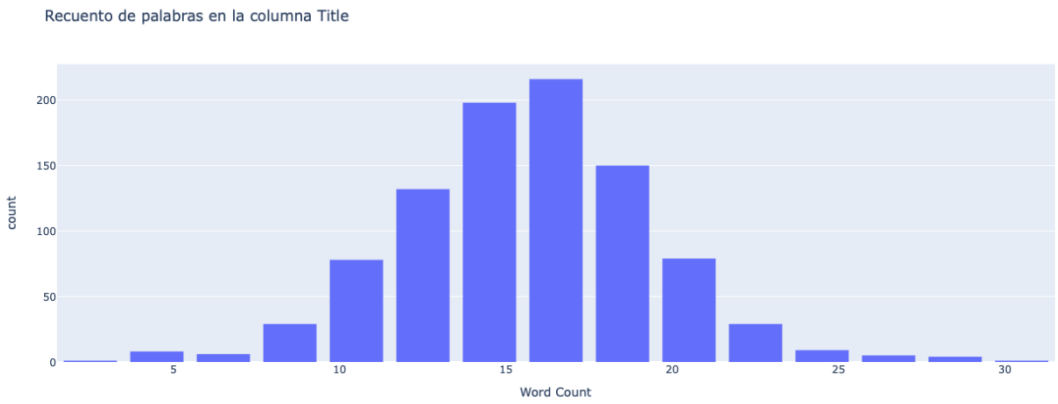


Figura 13: Histograma del número de palabras de los títulos

Fuente: Elaboración propia

Número de palabras	Cantidad de títulos
2 – 3	1
4 – 5	8
6 – 7	6
8 – 9	29
10 – 11	78
12 – 13	132
14 – 15	198
16 – 17	216
18 – 19	150
20 – 21	79
22 – 23	29
24 – 25	9
26 – 27	5
28 – 29	4
30 – 31	1

Tabla 5: Tabla del número de palabras de los títulos

Fuente: Elaboración propia

Subtítulo: El subtítulo cuenta con más palabras que el título. La mayoría oscilan entre las 15 y 40 palabras. Singularmente, hay tres que tienen sólo entre cinco y nueve palabras. De entre los diez subtítulos con más palabras, cinco son de El País, cuatro de ABC y el más largo, con 68 palabras y relativo a la rebelión del grupo Wagner, es de El Mundo.

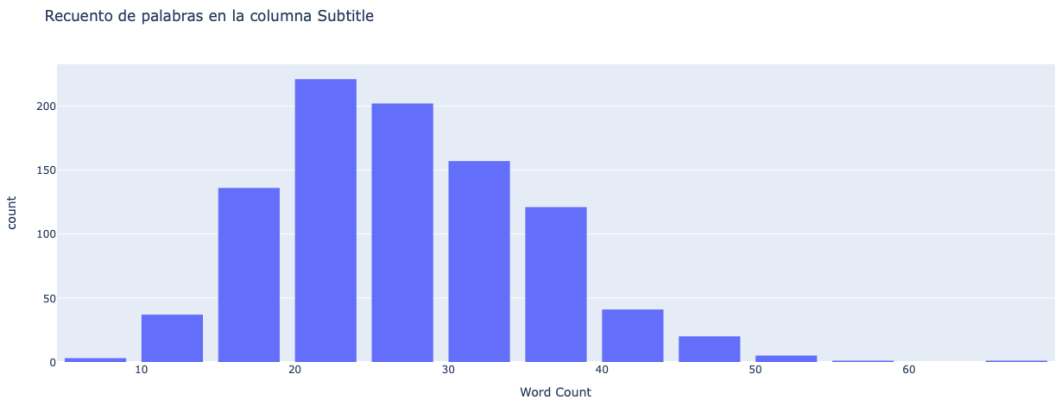


Figura 14: Histograma del número de palabras de los subtítulos

Fuente: Elaboración propia

Número de palabras	Cantidad de subtítulos
5 – 9	3
10 – 14	37
15 – 19	136
20 – 24	221
25 – 29	202
30 – 34	157
35 – 39	121
40 – 44	41
45 – 49	20
50 – 54	5
55 – 59	1
65 – 69	1

Tabla 6: Tabla del número de palabras de los subtítulos

Fuente: Elaboración propia

Cuerpo: El cuerpo de la noticia es la segunda columna más extensa. Aquí vemos, teniendo en cuenta la selección parcial que hemos hecho, que la mayor parte de las noticias tienen entre 120 y 259 palabras. De hecho, ninguna noticia tiene más de 400 palabras ni menos de 60. Las noticias más largas suelen ser las de El País o El Confidencial, mientras que ABC presenta las más cortas.

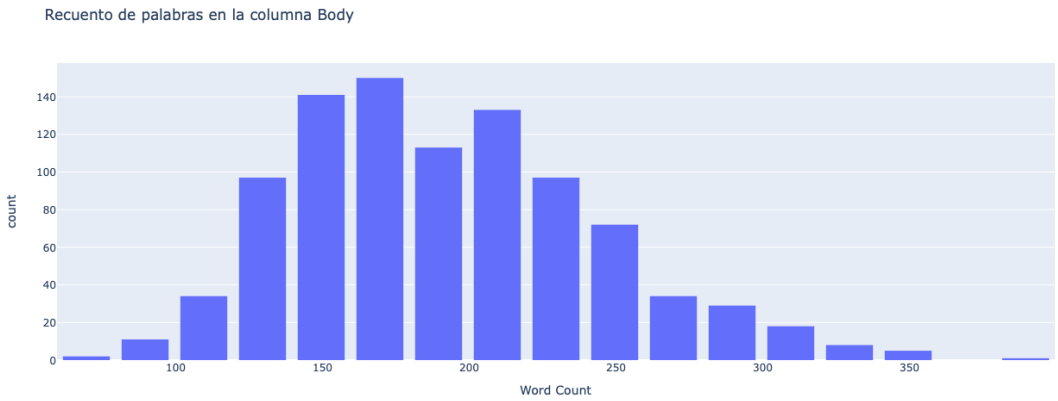


Figura 15: Histograma del número de palabras de los cuerpos

Fuente: Elaboración propia

Número de palabras	Cantidad de cuerpos
60 – 79	2
80 – 99	11
100 – 119	34
120 – 139	97
140 – 159	141
160 – 179	150
180 – 199	113
200 – 219	133
220 – 239	97
240 – 259	72
260 – 279	34
280 – 299	29
300 – 319	18
320 – 339	8
340 – 359	5
380 – 399	1

Tabla 7: Tabla del número de palabras de los cuerpos

Fuente: Elaboración propia

"All": Esta es la columna que va a ser sometida al *topic analysis* y en definitiva, la principal columna de trabajo, por contener toda la información. En este sentido, la columna resulta de añadir título, subtítulo y cuerpo en las mismas celdas, de ahí que su número de palabras será la suma de las tres anteriores. La mayoría de las noticias tienen entre 140 y 320 palabras. No hay registros inferiores a las 100 palabras ni superiores a las 439.

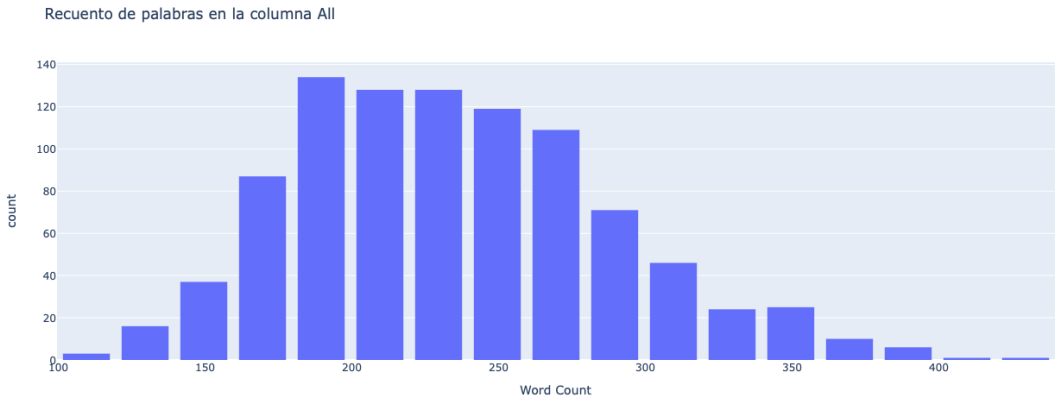


Figura 16: Histograma del número de palabras totales

Fuente: Elaboración propia

Número de palabras	Cantidad de "*All*"
100 – 119	3
120 – 139	16
140 – 159	37
160 – 179	87
180 – 199	134
200 – 219	128
220 – 239	128
240 – 259	119
260 – 279	109
280 – 299	71
300 – 319	46
320 – 339	24
340 – 359	25
360 – 379	10
380 – 399	6
400 – 419	1
420 – 439	1

Tabla 8: Tabla del número de palabras totales

Fuente: Elaboración propia

Puede verse que la mayoría de los histogramas tienen una distribución aproximadamente normal, con pocos registros muy cortos o largos. No nos parece conveniente eliminar estos registros, pues no suponen ni un 0.5% del total de las noticias y la información perdida puede resultar siendo mayor que la optimización de resultados derivada de la eliminación de los *outliers*.

Posteriormente, hemos examinado las palabras más frecuentes de la columna "*All*", lo que nos da una pista de si tendremos que eliminar o no las conocidas como *stopwords*. Los resultados fueron los siguientes:

Palabra	"de"	"la"	"el"	"que"	"en"	"a"	"y"	"del"	"los"	"con"
Frecuencia absoluta	14260	9450	7226	6543	6429	5290	5103	3925	3588	2517

Tabla 9: Palabras más comunes

Fuente: Elaboración propia

La conclusión es clara: hemos de eliminar las *stopwords* durante el preprocesamiento del texto, pues las palabras más repetidas aportan poco y pueden viciar el *topic analysis*.

Por último, vamos a representar la media de palabras de la columna "*All*" para cada uno de los cinco medios objeto de estudio, lo que nos permitirá detectar diferencias iniciales en la extensión de las noticias seleccionadas:

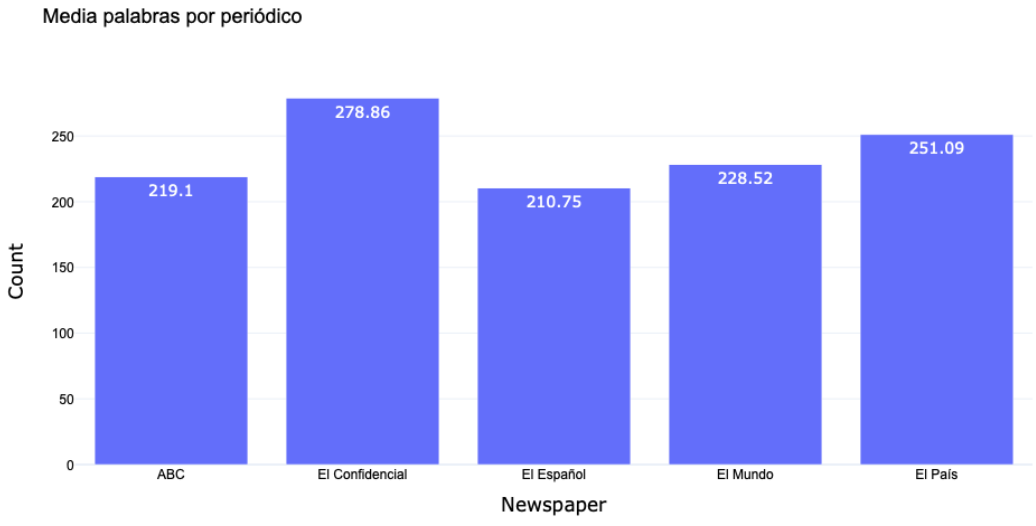


Figura 17: Media de palabras para cada periódico

Fuente: Elaboración propia

Si bien es cierto que todos oscilan en torno a los mismos valores, El País y El Confidencial son los medios en los que las noticias seleccionadas son más largas en media, a pesar de que en El País se ha elegido en la mayoría de los casos un solo párrafo. Por otro lado y aun habiendo escogido tres párrafos de cada noticia, El Español se mantiene como el medio con noticias de menor extensión.

4. PREPROCESAMIENTO DEL TEXTO

Una vez concluido el análisis preliminar y habiendo construido ya nuestro *corpus*, conviene realizar una importante tarea de preprocesamiento del texto para que podamos pasar al *topic analysis*. Aquí, realizamos varias tareas entre las que cabe destacar las siguientes:

En primer lugar, hemos habilitado con carácter previo el empleo de bigramas. El diccionario Collins (s.f.) define los bigramas como: "*a unit of two words, letters, or symbols that occur together in a text*". Es decir: un conjunto de dos palabras, letras o símbolos que suceden de forma conjunta en un texto. Por ejemplo, campaña suele ir seguida de electoral. Habida cuenta de la relevancia que tienen los pares de términos en el lenguaje político[12] hemos decidido activar esta función.

En segundo término, hemos *tokenizado* el texto. Un *token* generalmente tiende a identificarse con una palabra. No obstante y sin perjuicio de que en nuestro estudio casi siempre sea una palabra, la *tokenizacion* es (IBM, s.f.): "el proceso de desglosar un texto de formato largo en oraciones y palabras llamadas "*tokens*", las cuales se utilizan en los modelos como bolsa de palabras para tareas de agrupación de textos y cotejo de documentos". Así pues, hemos *tokenizado* la columna "*All*", que es la que vamos a usar en el análisis, así como las columnas "*Title*", "*Subtitle*" y "*Body*", por si en algún momento se quisiera emplear la muestra en otros estudios.

En tercera instancia, hemos lematizado la muestra. Por lematización entendemos la reducción de una palabra a su lema, que no es más que (IBM, s.f.): "el proceso de separar los prefijos y sufijos de la palabra para extraer el lema y el significado. Esta técnica mejora la recuperación de información, ya que reduce el tamaño de los archivos de indexación". El resultado de la lematización es que todas las formas flexibles de una palabra acaban reconducidas al mismo lema. Por ejemplo, comeré y comíamos vienen del lema "comer". No obstante, hemos realizado una intensiva tarea de filtrado para que los lemas extraídos sean "limpios". Por tanto, hemos aplicado las siguientes transformaciones y sometido al proceso de lematización sólo a los *tokens* que cumplían con las siguientes condiciones:

[12] En el ámbito político el empleo de pares de palabras es muy común. Se emplean, entre otras cosas: a) para designar la convocatoria electoral (elecciones municipales, autonómicas, generales…); b) para definir pactos o coaliciones (sumar y podemos, pp y vox, psoe y sumar…); c) al acompañar a los candidatos (candidato Feijóo, Vicepresidenta Díaz, Presidente Sánchez…).

Primero: Transformación de todos los caracteres en minúsculas. A pesar de que *spaCy* es un modelo definido como *case insensitive* (es decir, no distingue entre mayúsculas y minúsculas), la realidad es que al comparar nuestros *tokens* con el listado de *stopwords*, no era capaz de eliminar aquellas palabras que, siendo *stopwords*, contenían alguna letra mayúscula. Por tanto, hemos pasado el texto a minúsculas. Esto también nos asegura una mejor comparabilidad entre los *tokens*, previniendo posibles fallos adicionales relativos a la *case sensitivity*.

Segundo: Filtrado y eliminación de *stopwords.* Las *stopwords* son palabras cuyo significado no es relevante. Son palabas vacías de contenido que tienden a repetirse muchas veces, viciando el análisis de los términos con mayor relevancia. Una vez eliminadas, se puede proceder al análisis de los términos cuya aportación es más significativa. Las hemos eliminado a través de la subida de un archivo denominado "*spanish.txt*" que contiene una serie de palabras en castellano típicamente identificadas como *stopwords*. Algunos ejemplos son las palabras: "a", "actualmente", "adelante" o "fin", "fue", entre otros. Todo *token* cuyo contenido coincidiese con algún registro del citado listado ha quedado desechado.

Tercero: Los *tokens* deben ser alfabéticos. Hemos exigido que los *tokens* estén compuestos por caracteres alfabéticos, descartando así los numéricos, cuya información suele ser poco relevante al emplearse o bien para designar un número de representantes, cuestión estudiada al explicar el ciclo electoral o bien como acrónimo de las convocatorias electorales (ej. 28M, 23J…), aportando en este caso información de escasa importancia.

Cuarto: Ha de encontrarse en los *allowed part-of-speech tags*. Las *part-of-speech* son categorías gramaticales como los sustantivos, pronombres, verbos, adjetivos o adverbios. Esta es una forma de filtrar texto habitualmente empleada en los algoritmos de *text mining* mediante la que se puede optimizar el procesamiento al guardar sólo aquellas categorías gramaticales que aportan información realmente relevante. En el caso concreto que nos ocupa, las *part-of-speech* seleccionadas han sido las siguientes: adjetivos, adverbios, sustantivos, nombres propios y verbos. Por lo tanto, el modelo sólo lematizará y almacenará aquellas palabras que se encuentren dentro de esas categorías gramaticales.

Quinto: Condiciones adicionales. Estos últimos filtros son filtros "menores", en el sentido de que no precisan de un tratamiento individualizado por tratarse de un filtrado del estilo "cajón de sastre". En este apartado vamos a citar los filtros adicionales que, en alguno de los casos, pueden incluso coincidir con los ya tratados en los cuatro párrafos

inmediatamente anteriores. Pues bien, se ha obligado a que los *tokens* no puedan ser: signos de puntuación, signos dinerarios, dígitos, palabras fuera del vocabulario (ej. MLT-19 u otras expresiones que normalmente no se ven en un procesamiento de lenguaje natural) y espacios.

Una vez realizados estos filtrados, se ha creado la columna adicional "*All_clean*", así como las correspondientes al título, subtítulo y cuerpo en las que, por no ser objeto de análisis, no conviene detenernos. Esta columna tiene la siguiente estructura: es una línea en la que para cada noticia o registro hay una lista de los *tokens* que cumplen con todas las condiciones expuestas.

```
0      [sánchez, feijóo, modelos, país, electoral, pr...
1      [campaña_electoral, cis, clave, local, peso, m...
2      [m, prueba, sánchez, frente, feijóo, necesita,...
3      [ley, partidos, gobierno, únicos, instar, ley,...
4      [feijóo, campaña, psoe, bildu, sánchez, gestió...
                           ...
940    [pedro_sánchez, presidente, puigdemont, gobier...
941    [pedro_sánchez, presidente_gobierno, votos, an...
942    [alivio, moncloa, elecciones, psoe, gobierno, ...
943    [pedro_sánchez, presidente_gobierno, puigdemon...
944    [pedro_sánchez, presidente, votos, moncloa, gr...
Name: All_clean, Length: 945, dtype: object
```

Figura 18: Ejemplo de la columna "*All_clean*"

Fuente: Elaboración propia

Ahora y como parte final del preprocesamiento del texto conviene introducir dos cuestiones: a) algunas visualizaciones preliminares; y, b) la matriz TF-IDF.

Las visualizaciones preliminares se realizan con los términos más frecuentes a nivel absoluto. De entre todas las posibles, hemos optado por una nube de palabras, pues expresa de forma visual y clara las palabras más frecuentes en términos absolutos:

Figura 19: Nube de palabras preliminar con frecuencias absolutas

Fuente: Elaboración propia

Las palabras más repetidas son, como era previsible, las correspondientes a los nombres de los partidos políticos (pp, psoe…), o los nombres o apellidos de los candidatos, bien en modo bigrama (pedro_sánchez, yolanda_díaz…) o en solitario (feijóo, sánchez…). Igualmente, surgen de manera reiterada términos relativos a la comunicación política o al desarrollo electoral (campaña, presidente, gobierno, ejecutivo, izquierda…). Sorprende, aunque parece coherente con el marco teórico, la ausencia de términos económicos, aunque algunos como millones o datos sí aparecen, pero con una frecuencia tan reducida que los hace poco perceptibles a nivel visual. Algunas palabras que encontramos con frecuencia mayor a la esperada serían las expresiones fuentes o acuerdo.

Por último, calculamos la matriz TF-IDF. Su origen radica en la *Term Frequency-Inverse Document Frequency matrix*. Está compuesta por dos elementos (Provost y Fawcett, 2013) y constituye otra representación del conjunto de *tokens* o *bag-of-words*, además de la frecuencia absoluta anteriormente citada. Es una forma óptima para identificar las palabras no tanto más comunes, sino las más importantes en un conjunto de documentos:

Su primer componente es la TF o *term frequency*. Representa la relevancia de un término *t* en un documento *d*. Es decir, el número de veces que el término aparece en el

documento. Es la misma frecuencia representada en la Figura 19 y de ella sólo cabe precisar que cuantas más veces aparezca un término, mayor será su TF.

Como su nombre indica, el segundo componente es la IDF o *inverse document frequency*. Esta tiene en cuenta cómo de común es el término dentro del *corpus* en su conjunto, pues si aparece en todos los documentos quizás no sea lo suficientemente relevante, mientras que aquellos términos que aparecen en menor medida pueden dar información sobre un tema que, de estudiarse sólo en términos de frecuencia absoluta, pasaría desapercibido. Existen varias formas de calcularlo, entre las que puede destacarse la siguiente: logaritmo en base dos de la división del total de noticias entre las noticias que contienen el término. En caso de que no hubiéramos eliminado las *stopwords*, tendrían una IDF baja.

La nube de palabras resultante fue la siguiente:

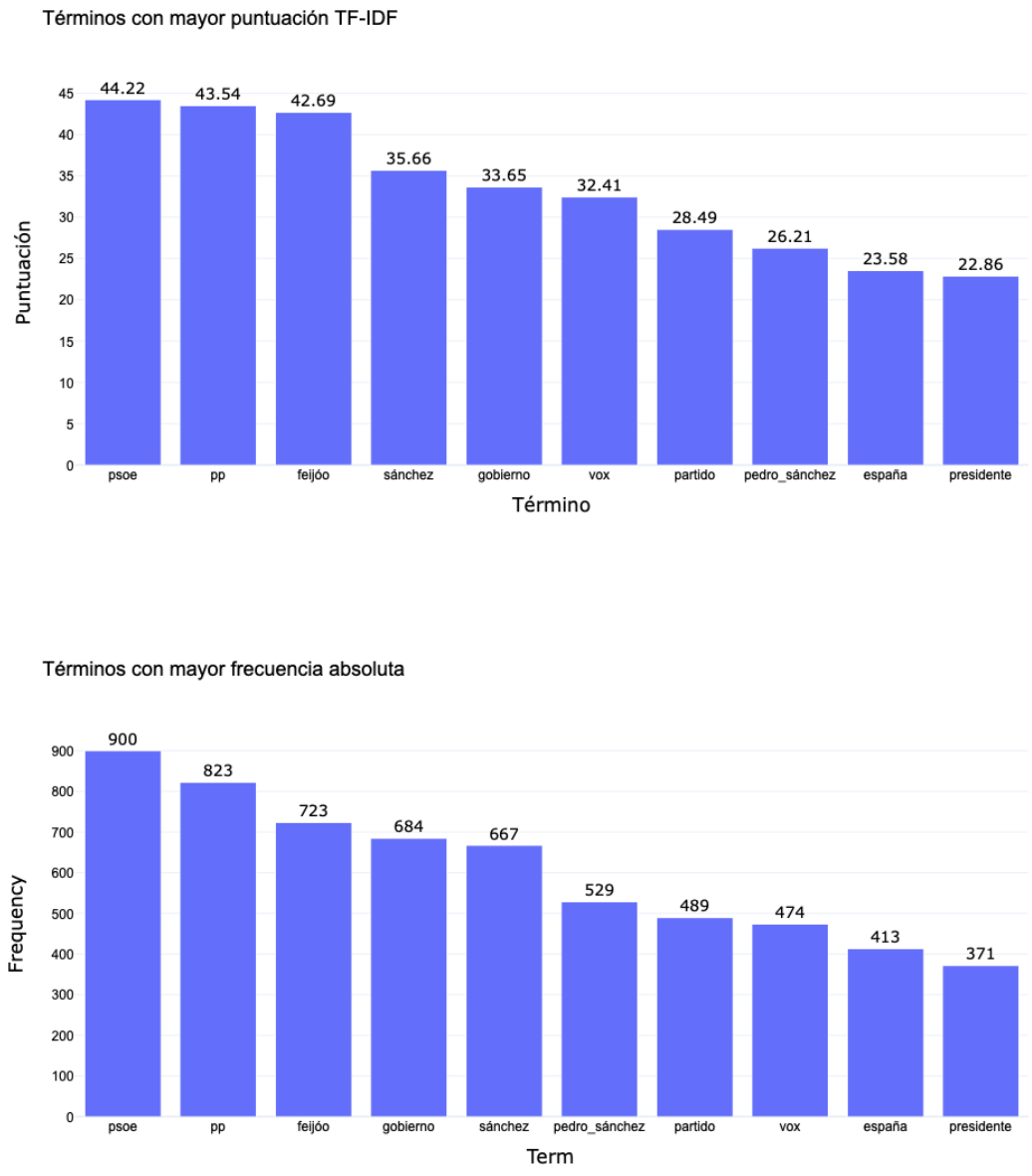
Figura 20: Nube de palabras de las puntuaciones totales de la matriz TF – IDF

Fuente: Elaboración propia

Las palabras son muy similares aquí y en la anterior nube de palabras, aunque hay algunas pequeñas diferencias. Por ejemplo, partido es más grande aquí que en la anterior nube, pero fundamentalmente mantienen la misma esencia. Vuelven a estar ausentes los temas económicos, en un número todavía más reducido. La presencia de palabras como "pp", "psoe" o "Sánchez" nos parece lógico en términos de TF-IDF, pues al final la

multiplicación por la frecuencia absoluta (TF-IDF = TF * IDF) lo propulsa a unos valores superiores.

Motivados por la excesiva similitud, hemos decidido representar dos últimas visualizaciones que comparen los diez términos más comunes tanto en frecuencia absoluta como en frecuencia TF-IDF.



Figuras 21 y 22: Comparación términos más frecuentes

Fuente: Elaboración propia

El resultado es esperado: están presentes los mismos diez términos, pero en orden diferente. Como ya explicamos anteriormente, esto puede ser por la propia fórmula de la TF – IDF y la multiplicación por la frecuencia, que en el caso del PSOE sobrepasa las 800 veces hasta alcanzar las 900.

5. *TOPIC ANALYSIS*

A continuación realizaremos el *topic analysis*. El objetivo de este análisis es meridianamente claro: saber qué temas son los más frecuentes, consiguiendo así responder a la pregunta que nos hacíamos al inicio del trabajo: ¿cuál fue la agenda de los medios españoles durante el ciclo electoral de 2023? Este análisis (Jacobi et al., 2015) lo que hace es identificar patrones latentes a través de la distribución de palabras en una colección de documentos. De su aplicación resultan una serie de *topics*, que no son otra cosa que *clusters* de palabras que aparecen de forma simultánea en una serie de documentos de acuerdo con los patrones objeto de análisis. No debe confundirse la palabra *topic* con tema pues, aunque en nuestro caso sí coinciden, hay otros casos de *topic analysis* (ej. emociones) en la que se representan formas de expresarse. Como ya dijimos, estos *clusters* contendrán una serie de términos que son homogéneos dentro del *cluster* y heterogéneos en relación con los demás *clusters*, aunque a veces puedan solaparse.

Para iniciar este análisis hemos de crear la *Document-Term Matrix* (en adelante, DTM). La DTM (Reintech, s.f.) es una matriz matemática que describe la frecuencia de los términos que aparecen en una colección de documentos. En la DTM, las filas se corresponden con documentos y las columnas con términos, de manera que cada valor *Aij* representará la frecuencia de un término en un documento. Se suelen emplear este tipo de matrices en *text mining* y sistemas de obtención de la información, al simplificar el proceso de comprensión y visualización de datos de texto. Para nuestro caso, hemos creado un diccionario a partir de la columna "*All_clean*", que como ya anunciamos antes, era el resultado de lematizar la columna "*All*". Una vez configurado el diccionario, se ha realizado la DTM y comprobado el número de términos y de bigramas[13]: 1605 términos y 122 bigramas.

Una vez creada la DTM, procedemos a aplicar el modelo *Latent Dirichlet Allocation* (en adelante, LDA). Este modelo (Jacobi et al., 2015; Tong, y Zhang, 2016; Zhu et al., 2016) es un modelo Bayesiano cuyo parámetro es variable. Es una técnica de aprendizaje no supervisado que crea *topics* en base a la aparición simultánea de palabras igualmente distribuidas, dejando al usuario la interpretación. Presupone que un documento puede ser

[13] De cara a configurar la DTM hemos señalado unos parámetros de modo previo. En cuanto a los bigramas, hemos indicado que debe haber un mínimo de diez asociaciones para que nos guarde el bigrama. Y en cuanto a los términos, para que sea seleccionado un término este debe estar como mínimo en cinco documentos diferentes y como máximo en el setenta por cien de los documentos, pues una repetición superior podría indicar irrelevancia. Por último, hemos limitado los términos a 3500.

explicado por una distribución polinómica de temas latentes y que esos temas pueden a su vez ser explicados por una distribución polinómica de palabras, de ahí que algunos autores afirmen que se trata de un modelo de tres capas. Para que sea útil, el usuario debe interpretar los términos que integran cada *topic* de acuerdo con el contexto del texto introducido, extrayendo así conclusiones lógicas.

No obstante y a la hora de realizar el LDA, existe una pluralidad de métricas que nos pueden asistir a los usuarios en la tarea de decidir cuál es el número óptimo de *topics*. Estas métricas permiten evitar la excesiva complejidad derivada de tener que inferir manualmente a través de la lectura de un documento de más de 500 páginas cuántos son los *topics* que deberían escogerse en un análisis. El número de *topics* es un parámetro que seleccionamos *a priori*, de ahí la necesidad de tener claro su númeo. En el caso que nos ocupa, se han empleado las métricas de coherencia y perplejidad (Pinto Gurdiel et al., 2021):

En primer lugar, atenderemos a la coherencia. La coherencia examina los *topics* en función de la comprensión humana. Un *topic* coherente estará compuesto por una lista de palabras que, de modo colectivo, tienden a representar un tema semántico. El método de puntuación suele realizarse en función de la coherencia interna del propio *corpus*, de manera que cuanto mayor es la puntuación más coherentes serán los *topics*.

En segundo término analizaremos la perplejidad. El modo más común de evaluar los *topics* es el de valorar el desempeño del modelo para documentos no analizados, es decir, su capacidad predictiva. La perplejidad está inversamente relacionada con la capacidad predictiva del modelo, por lo que a menor perplejidad, mejor será la capacidad generalizadora del modelo. Si se introducen demasiados *topics*, la probabilidad generalizadora disminuye, por lo que el análisis es menos útil.

Por todo ello, vamos a calcular la coherencia y la perplejidad para los siguientes números de *topics*: dos, tres, cuatro, cinco, seis, siete, ocho, nueve, diez, 11, 12, 13, 14, 15, 20, 25, 30, 40, 50 y 100. Los resultados han sido los siguientes:

Número de *topics*	Coherencia media	Perplejidad media
2	0.419472	-6.647108
3	0.421850	-6.615646
4	0.428249	-6.613370
5	0.425582	-6.598291
6	0.425227	-6.604485
7	0.451325	-6.632449
8	0.460747	-6.622285
9	0.445497	-6.646201
10	0.452358	-6.667201
11	0.487968	-6.663247
12	0.479417	-6.664934
13	0.429123	-6.676941
14	0.452083	-6.683399
15	0.472001	-6.690777
20	0.458571	-6.744383
25	0.442747	-6.790783
30	0.441991	-6.839228
40	0.430765	-6.923492
50	0.427287	-7.024405
100	0.398627	-7.353475

Tabla 10: Valores de coherencia y perplejidad

Fuente: Elaboración propia

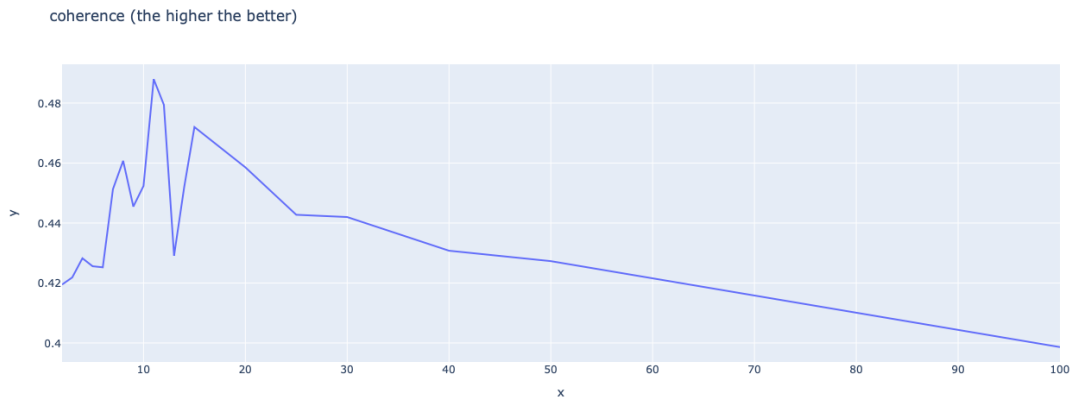


Figura 22: Gráfico de los valores de coherencia

Fuente: Elaboración propia

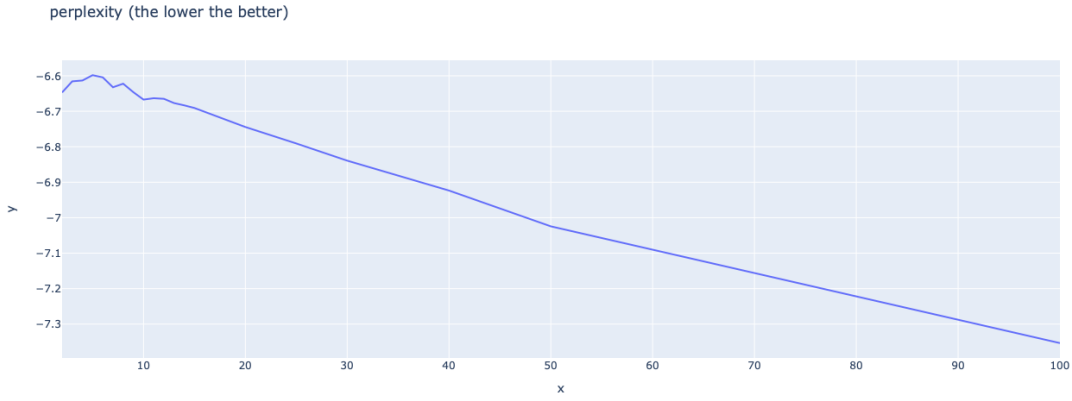


Figura 23: Gráfico de los valores de perplejidad

Fuente: Elaboración propia

Como puede verse, los movimientos y fluctuaciones son mayores al inicio de ambos gráficos (especialmente porque hacemos un análisis individualizado de los 15 *topics*). En cuanto a la selección, la duda ha girado en torno a los cuatro, siete u ocho *topics*, por tener ambos valores aceptables en términos de las dos métricas y no suponer una cantidad excesiva. En caso de fijarnos en los valores máximos de complejidad habrían de seleccionarse 14 *topics*, hipótesis descartada *a priori* por tratarse de demasiados y *a posteriori* tras verificarla mediante prueba y error. Si nos fijásemos en la perplejidad y buscásemos el mínimo valor, habríamos de seleccionar hasta 100 *topics*, lo que resulta poco operativo habida cuenta de que tenemos menos de 1200 términos.

Siguiendo la hipótesis expresada al principio del capítulo, relativa a la necesidad de interpretar el LDA de manera conforme y coherente con el conjunto del texto, hemos realizado pruebas con los tres valores que nos suscitaban dudas. Esto es, hemos probado a ver los resultados en caso de seleccionar cuatro, siete u ocho *topics*.

La conclusión fue tomar siete *topics* como valor definitivo. Por añadidura, hemos seleccionado 35 *LDA PASSES*. Es decir, que el modelo va a recorrer 35 veces la totalidad del *corpus* para seleccionar cada *topic*. El *random_state* lo dejamos en 42, para garantizar la reproductividad y que siempre se realice la misma *randomization* a la hora de presentar los *topics*, pues en caso de no introducirlo la semilla puede variar en cada partición entre conjunto de entrenamiento y validación. Estos resultados se presentan en dos dimensiones que representan los dos componentes principales PC1 y PC2, de manera que los temas estarán encuadrados en dos ejes. Ahora procederemos a un análisis pormenorizado de los resultados. Para comenzar, mostraremos los cinco términos más frecuentes en cada *topic* y sus pesos:

Como va a poder apreciarse en las figuras 25 a 31 (insertadas a continuación de este párrafo), las palabras más frecuentes aportan poca información o relevancia. Por lo tanto, vamos a modificar el parámetro lambda y a hacer una valoración individual de cada *topic*. Lambda es un parámetro que puede tomar valores que oscilan entre el cero y el uno. Si se acerca más al uno, representará términos con un ratio de frecuencia para la totalidad del *corpus* mayor que si toma valores más cercanos al cero, donde representa términos que son más específicos para el *topic*, pero menos frecuentes en el resto del *corpus*. Nosotros hemos optado por dar a lambda un valor de 0.5, de manera que así podamos estar en equilibrio entre la especificidad y la generalidad, habida cuenta de la enorme similitud de los términos más frecuentes de cada *topic*.

Los resultados, que posteriormente serán analizados, son los siguientes:

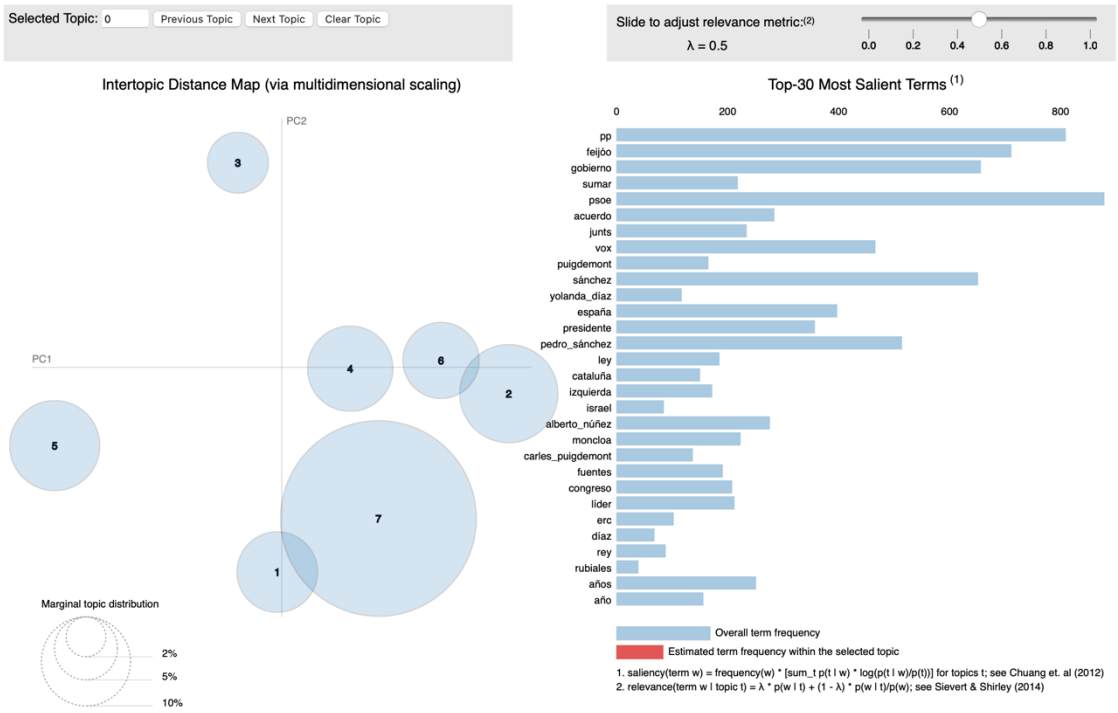


Figura 24: Resultados de la aplicación del *topic analysis*

Fuente: Elaboración propia

Los términos más frecuentes para cada uno de los *topics* son los siguientes:

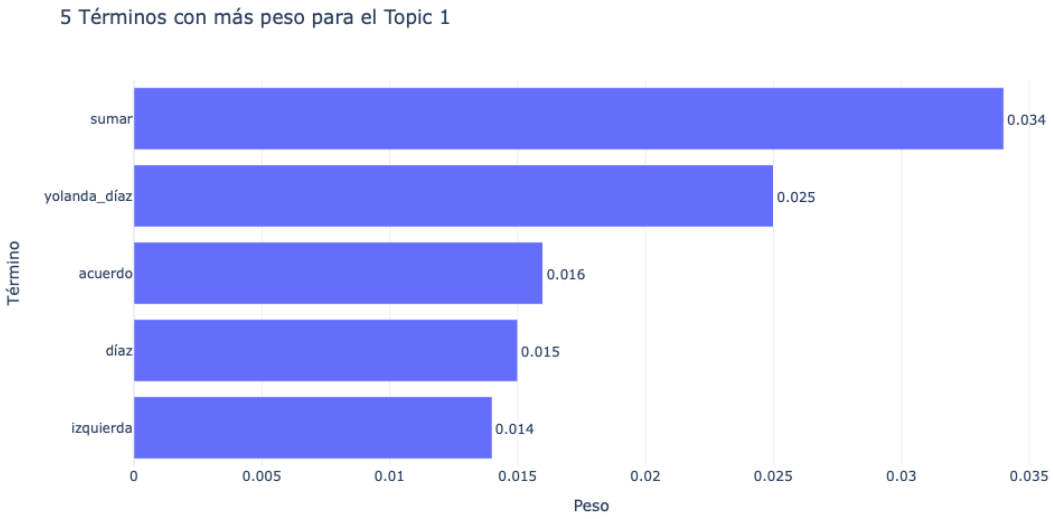


Figura 25: Términos más frecuentes para el *topic* 1

Fuente: Elaboración propia

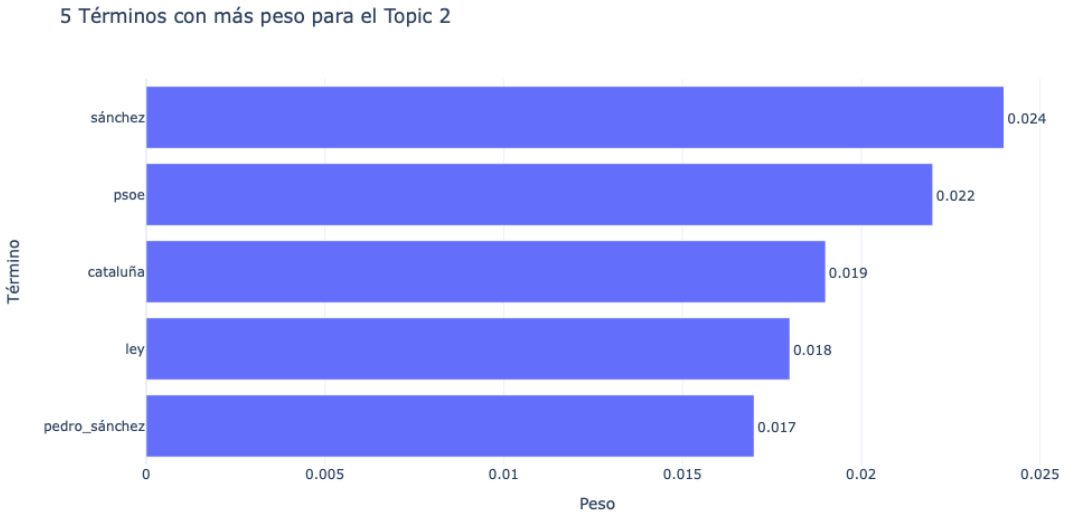


Figura 26: Términos más frecuentes para el *topic* 2

Fuente: Elaboración propia

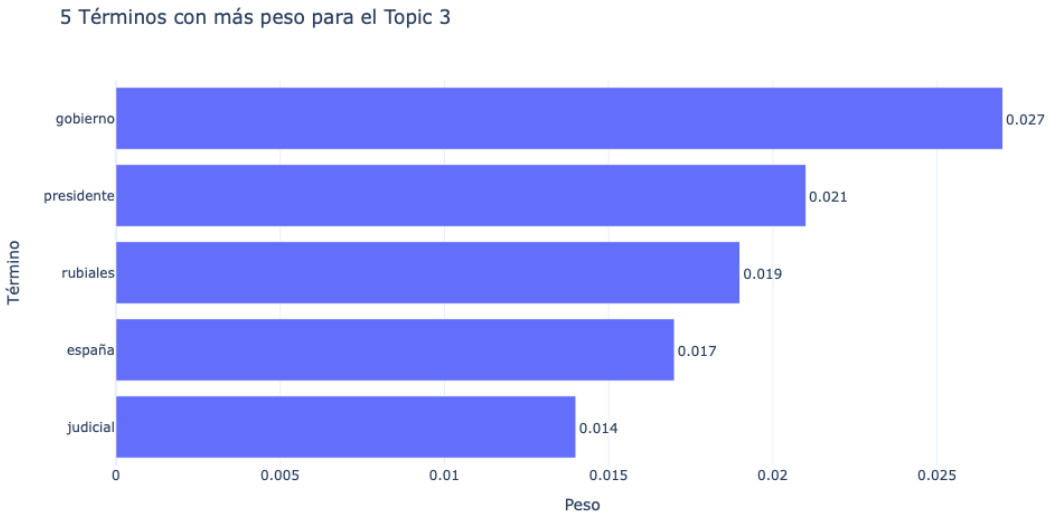


Figura 27: Términos más frecuentes para el *topic* 3

Fuente: Elaboración propia

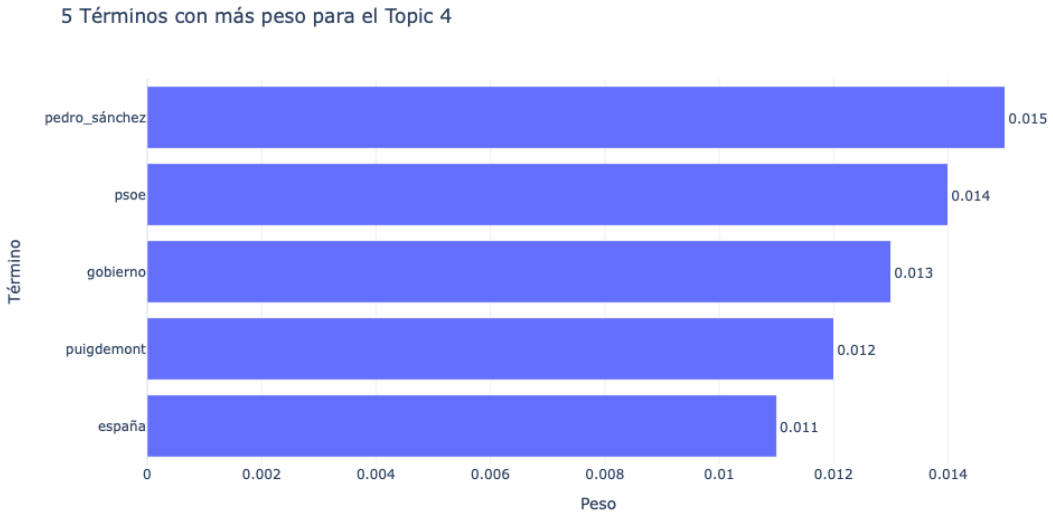


Figura 28: Términos más frecuentes para el *topic* 4

Fuente: Elaboración propia

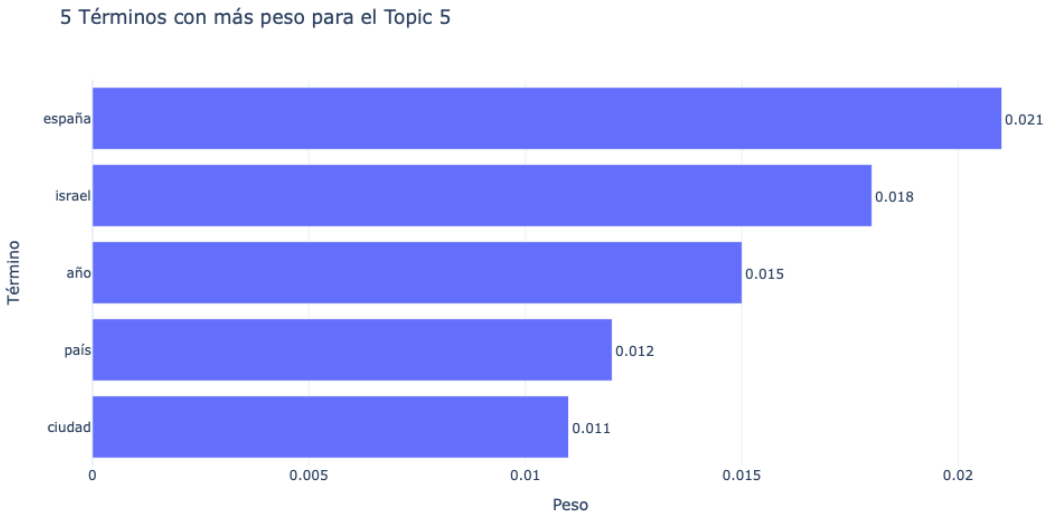


Figura 29: Términos más frecuentes para el *topic* 5

Fuente: Elaboración propia

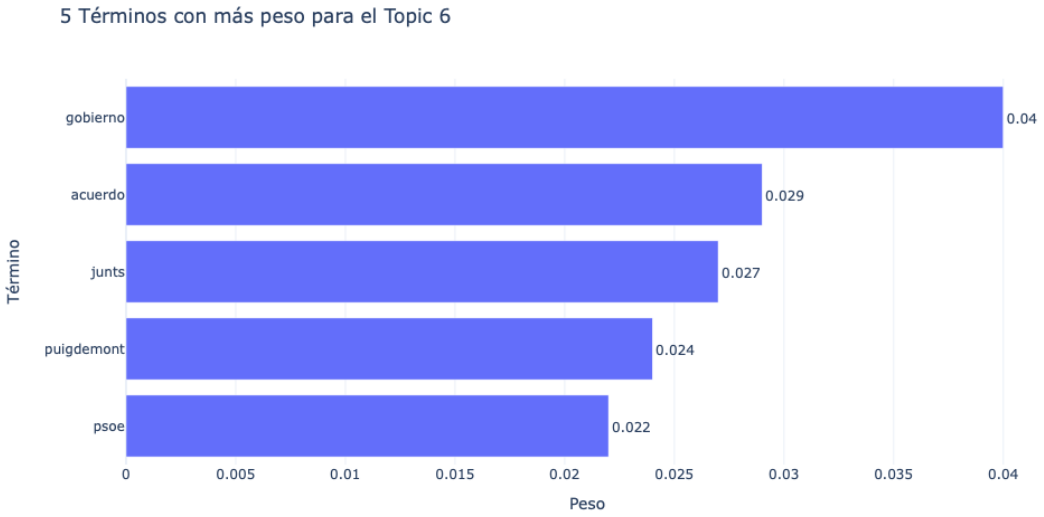


Figura 30: Términos más frecuentes para el *topic* 6

Fuente: Elaboración propia

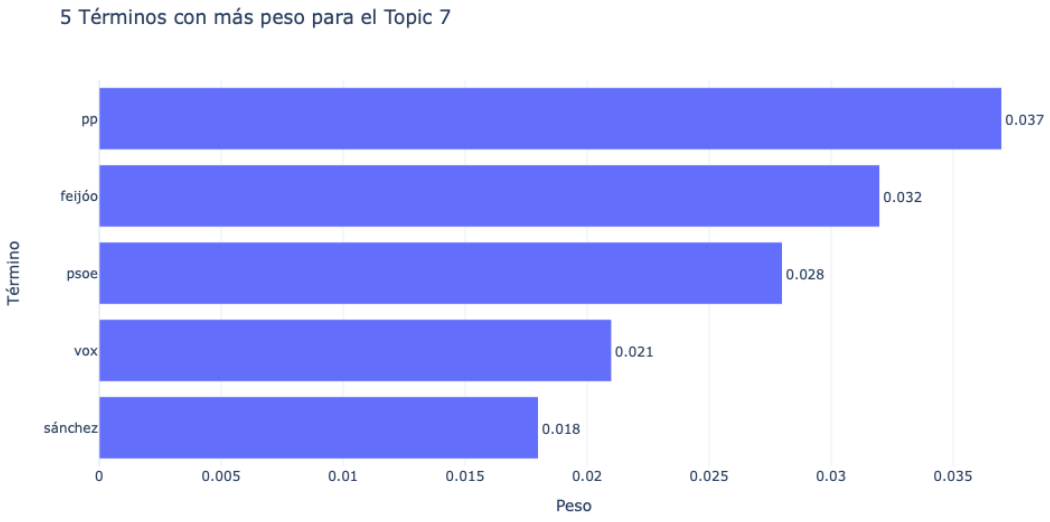


Figura 31: Términos más frecuentes para el *topic* 7

Fuente: Elaboración propia

Antes de proceder al análisis pormenorizado, conviene precisar a grandes rasgos la evolución temporal de todos los *topics*, para lo que hemos realizado una media de los cinco valores -uno por noticia- que toma cada *topic* en cada día. Casi todos tienen picos

separados y diferenciados que se explican en los siguientes apartados, donde dotaremos a estos *topics* de contexto y análisis:

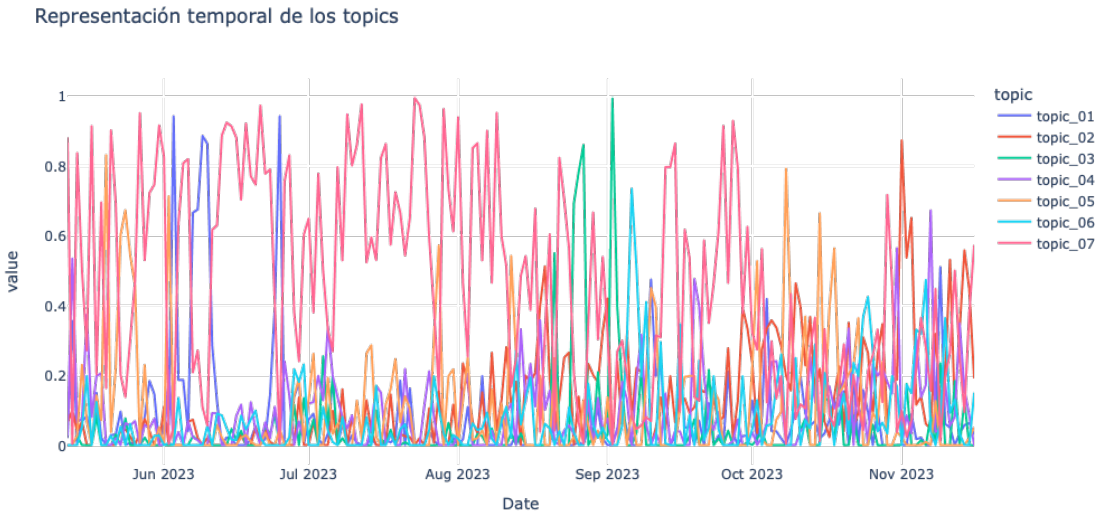


Figura 32: Evolución temporal de los *topics*

Fuente: Elaboración propia

5.1. *Topic* 01: Frente amplio de izquierdas

Como ya se avanzó durante el marco teórico, una de las principales novedades de las elecciones generales de 2023 fue la creación del frente amplio de izquierdas liderado por Yolanda Díaz. Este frente tiene su reflejo en este primer *topic*, compuesto por un 8.2% de los *tokens*. El tamaño tiene sentido, pues se gestó durante apenas dos semanas que si bien fueron convulsas, no dejan de ser dos semanas de las aproximadamente 26 que duró el estudio de campo. De hecho, resulta sorprendente y da pie a una reflexión acerca de la novedad que supuso esta coalición, pues con sólo dos semanas de duración ha logrado hacerse un hueco entre los siete *topics*. Los términos más mencionados con lambda ajustado a 0.5 fueron: "sumar", "yolanda_diaz", "Díaz", "prigozhin" e "Irene_montero"

Las palabras más mencionadas en este *topic* fueron las siguientes:

Figura 33: Nube de palabras para el frente amplio de izquierdas

Fuente: Elaboración propia

Puede verse que los términos típicamente relacionados con la izquierda aparecen representados en esta nube de palabras, como "Yolanda_diaz", "unidad", "plataforma" o "programa". Entre algunos términos cuya aparición no se esperaba o resulta sorprendente están los relacionados con el grupo Wagner, tales como "prigozhin" o "putin". Sin embargo, es posible que aparezcan porque los acontecimientos relacionados con el grupo Wagner sucedieron en el mismo lapso temporal que la coalición. Asimismo, avanzamos

que la coalición suscitó problemas por la no inclusión en las listas de Irene Montero, lo que puede verse claramente en las palabras “Irene_montero”, “ione”, “veto” o “ira”, entre otras. En cuanto a la distribución temporal de este *topic*, es la siguiente:

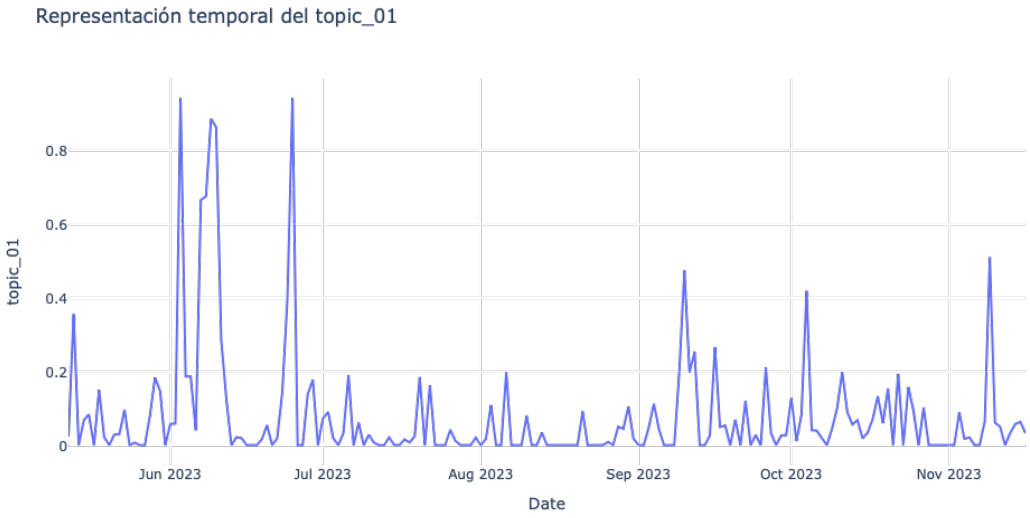


Figura 34: Distribución temporal del frente amplio de izquierdas

Fuente: Elaboración propia

La distribución temporal es en su mayoría lógica. En general, el *topic* alcanza su máxima relevancia en junio, sin perjuicio de otros picos alcanzados en distintos meses. Esto es porque fue en ese mes en el que se rubricó el acuerdo del frente amplio, lo que hace lógico que durante ese período de tiempo se hablase más del acuerdo que en otros momentos. En cuanto al verano (julio y agosto) no hay casi mención alguna. El resto de los picos guardan relación con otros acontecimientos simbólicos para este frente amplio, como sucede con el penúltimo, que coincide con el día en el que se rubricó el acuerdo PSOE-Sumar para la nueva legislatura. También hay algunos picos que no aportan información relevante, como el de septiembre, relativo al terremoto de Marruecos, sin que encontremos relación con el resto del tema.

5.2. *Topic* 02: Negociaciones sorprendentes

Aquí llegamos a uno de los tres *topics* relacionados con las negociaciones que acabaron dando lugar a la investidura de Pedro Sánchez como Presidente del Gobierno. En este caso, nos referimos a las que se llevaron a cabo con los independentistas catalanes, representados por los partidos *Esquerra Republicana de Catalunya* (en adelante, ERC) y *Junts*. Si bien es cierto que en el marco teórico introdujimos la polémica de la amnistía, es en este tema donde la problemática alcanza su máxima representación, sin perjuicio de que la palabra no aparezca expresamente. La razón es que posiblemente figura en más del 70% de las noticias, lo que hace que no sea un término significativo. En este sentido, este *topic* contiene el 12.1% de los *tokens*. En nuestra opinión, parece lógico, puesto que no fue hasta después del 23 de julio cuando el PSOE consideró la amnistía como una tesis válida, en lo que fue percibido como un cambio sorpresivo y sincronizado de opinión. Si bien es cierto que el proceso fue paulatino, al final del estudio de campo estaba plenamente interiorizado, habida cuenta de que la proposición de ley se había presentado en el Congreso de los Diputados.

Las palabras más comunes con lambda igual a 0.5 son: “Cataluña”, “ley”, “erc”, “constitución” y “Felipe_vi”. Su nube de palabras es la siguiente:

Figura 35: Nube de palabras para las negociaciones sorprendentes

Fuente: Elaboración propia

Podemos ver que en las palabras más frecuentes están aquellas relacionadas con estas negociaciones. Por un lado, las que hacen referencia a la ley como "texto", "propuesta" o "iniciativa". Por otro lado, otras que hacen referencia a los partidos o personas involucradas, "erc_junts", "independencia", "Sánchez" o "carles_puigdemont". Otras aluden al Jefe del Estado, puesto que durante las movilizaciones contra la amnistía, entre otras cosas, se suplicó la intercesión del Rey. De ahí que aparezcan palabras como "Felipe_vi" o "rey". Por último, el *topic* incluye también cuestiones relativas a la constitucionalidad de la amnistía, un asunto muy discutido. Prueba de ello son las referencias a "tribunal_constitucional", "derecho" o "legal".

Alguna mención que no guarda mucha relación con el tema, pero que aparece incluida en la nube, es la relacionada con la jura de la Constitución ante las Cortes Generales por Princesa Leonor con motivo de su mayoría de edad, acontecimiento que sucedió el 31 de octubre. Así sucede con palabras como "Leonor" o "princesa", probablemente por la similitud con las otras palabras de la Familia Real que sí guardan relación con el *topic*. La distribución temporal es la siguiente:

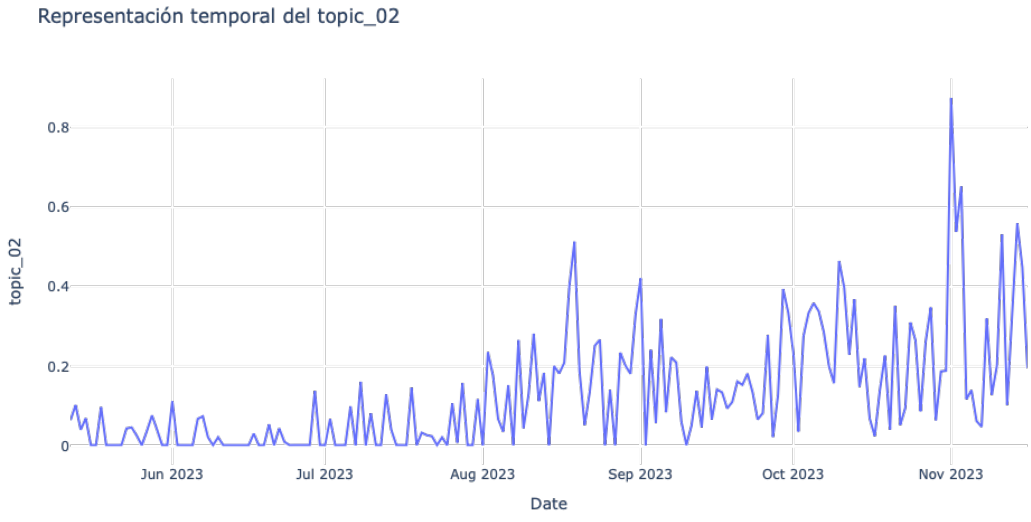


Figura 36: Distribución temporal de las negociaciones sorprendentes

Fuente: Elaboración propia

La conclusión aquí es rotunda: la distribución cuadra perfectamente con el contexto. Se observa como durante mayo, junio y julio las menciones a las negociaciones o a la amnistía son nulas. En cambio, a partir de las negociaciones de la Mesa del Congreso (día

17 de agosto), las referencias no dejan de crecer hasta el final del periodo. Si bien es cierto que hay fluctuaciones en cuanto a la importancia media, la tendencia es nítidamente ascendente. El pico máximo se alcanzó el uno de noviembre, donde coincidieron acontecimientos relativos a la monarquía -la citada jura- y la rúbrica de uno de los pactos, el del PSOE con ERC.

5.3. *Topic* 03: El último gol de Rubiales

Este es un *topic* poco esperado en el estudio de un ciclo electoral, pues está relacionado con el deporte. Sin embargo y como vinimos adelantando, durante el periodo previo a la investidura fallida de Núñez Feijóo, es decir, entre agosto y septiembre, la selección española ganó el mundial de fútbol femenino. Esta victoria fue agridulce en tanto que, al mismo tiempo, se produjo el conocido escándalo provocado por Luis Rubiales, presidente de la Real Federación Española de Fútbol. Este es el *topic* más pequeño, con un 4.7% de los *tokens*. Las palabras más comunes con lambda igual a 0.5 son: "rubiales", "Luis_rubiales", "cultura", "rfef" y "csd" y su nube de palabras es la siguiente:

Figura 37: Nube de palabras para el último gol de Rubiales

Fuente: Elaboración propia

Aquí el *wordcloud* está perfectamente relacionado con el tema, con pocos o ningún término incorrectamente asociado. Vemos: a) acrónimos de entidades deportivas, "csd", "rfef" o "tad"; b) referencias a los implicados y al suceso, "rubiales", "Jenni_hermoso", "beso"; y, c) palabras relacionadas con la respuesta del Gobierno, que fue rotundamente contraria a la actuación, como "cultura", "ejecutivo", "grave", "justicia" o "ley". Todo esto, sin olvidar aquellas palabras que se relacionan con el enorme éxito que supuso ganar

el mundial. Es el caso de “final”, “mundo” o “selección”. La distribución temporal del *topic* es la siguiente:

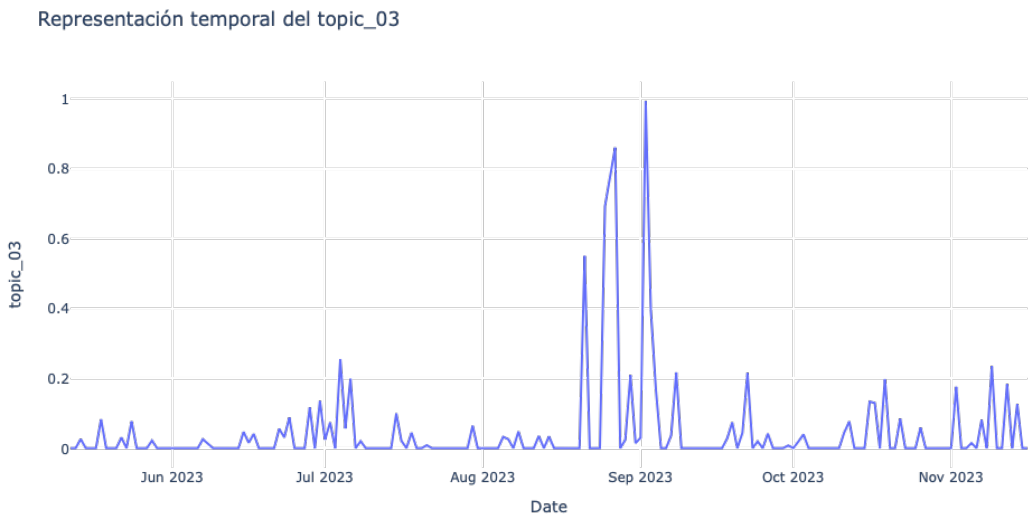


Figura 38: Distribución temporal del último gol de Rubiales

Fuente: Elaboración propia

Esta Figura nos muestra como la evolución del *topic* ha sido enormemente discreta, ya que los picos se limitan al tiempo que transcurrió entre la final y la resolución del escándalo derivado del controvertido beso. Es decir, finales de agosto y principios de septiembre. La mayoría de los registros dan un número muy próximo al cero y los únicos picos que no coinciden con el acontecimiento son como consecuencia de la palabra “partido”, que también está incluida en proposiciones como “partido electoral” o “partido político”, o por el efecto del suceso, que hizo que se recordase en otras noticias posteriores.

5.4. *Topic* 04: La Mesa del Congreso y sus expresiones lingüísticas

Como ya mencionamos al introducir el segundo *topic*, las negociaciones para la investidura de Sánchez, como las obras de teatro, estuvieron divididas en tres actos o *topics*. Pues bien, este es el segundo acto, relativo a la cuestión de las lenguas cooficiales y su introducción en la Mesa del Congreso. La Mesa del Congreso se constituyó el 17 de agosto bajo un pacto claro: permitir hablar en todas las lenguas reconocidas en cualquier parte del territorio español, singularmente catalán, gallego y euskera. De hecho, se trataba del resultado de un pacto con los nacionalistas de EH Bildu, PNV, BNG, ERC y *Junts*. El pacto igualmente incluía el intento de dar oficialidad a estas lenguas en el marco de la Unión Europea, proceso que se inició a partir del mismo día 17 de agosto. Adicionalmente se abordaron los problemas judiciales de los implicados en las negociaciones que, a la postre, se solventarían mediante la amnistía. Este *topic* incluye el 9.2% de los *tokens* y los cinco términos más mencionados con lambda igual a 0.5 son “Armengol”, “eta”, “lenguas”, “ue” y “santos”.

Las palabras más frecuentes son:

Figura 39: Nube de palabras para la Mesa del Congreso y sus expresiones lingüísticas

Fuente: Elaboración propia

Puede apreciarse que la frecuencia absoluta no coincide en su totalidad con las palabras más importantes con lambda igual a 0.5, que están con un tamaño más reducido. Si bien es cierto que se mencionan claramente palabras como "pedro_sánchez", "gobierno" o "socialista", aun siendo frecuentes, no constituyen la verdadera información latente del *topic*, fácilmente verificable al modificar el valor de lambda. Por tratarse de un *topic* que afecta a las tres lenguas, es extensivo a los tres nacionalismos, de ahí que haya términos como "eh_bildu" o "eta", cuestiones que suscitaron polémica al inicio del estudio de campo y que probablemente la mención generalizada de los nacionalistas los haya arrastrado a este bloque. Hay referencias claras al Congreso, como "Armengol" o "congreso_diputados" y a las negociaciones, con palabras como "Jordi" -negociador de *Junts*- y "Santos" -negociador del PSOE-, tanto españolas como europeas, lo que se ve en palabras como "UE" u "oficial". En cuanto a los procesos judiciales, vemos que hay palabras como "juez", "justicia" o "tsunami", referidas al caso del Tsunami Democrático y por el que se acusa a los líderes independentistas de presunto terrorismo. También hay palabras no relacionadas como "seguridad_social". La distribución temporal es la siguiente:

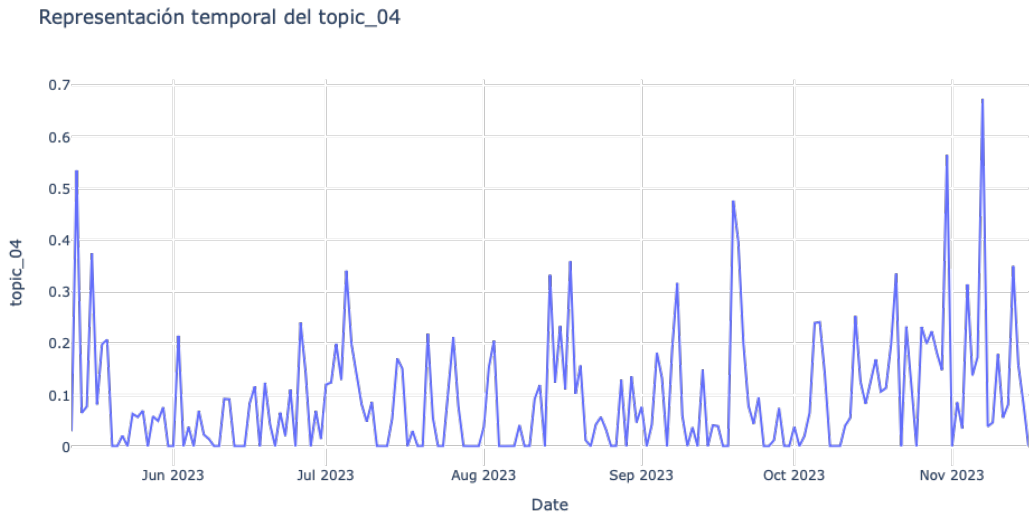


Figura 40: Distribución temporal de la Mesa del Congreso y sus expresiones lingüísticas

Fuente: Elaboración propia

La distribución temporal es extraña pero explicable. Los picos iniciales vienen dados por el carácter extensivo de los nacionalismos -EH Bildu fue muy mencionado-.

Posteriormente, los picos de agosto y septiembre son por dos cuestiones: a) el nombramiento de Armengol como Presidenta; y, b) la lucha en Europa por la oficialidad de las lenguas y la aprobación de su uso en el Congreso de los Diputados. Por último, los demás picos y los no explicados se entienden por las referencias a la justicia y los jueces.

5.5. *Topic* 05: El conflicto árabe-israelí

Este es el segundo y último *topic* que se aleja del ciclo electoral para dar paso a otros temas. El siete de octubre, como introdujimos en el marco teórico, el grupo terrorista Hamás lanzo un ataque sobre suelo israelí, que fue respondido por Israel con un ataque sobre la franja de Gaza. Este *topic* también incluye otras referencias al mundo árabe, singularmente las dudas sobre el proceso electoral en Melilla durante la campaña de las elecciones autonómicas, donde hubo detenidos por fraude electoral en las dos primeras semanas del estudio de campo. Incluye el 10.2% de los *tokens*, probablemente por las enormes repercusiones del conflicto. Las cinco palabras más repetidas con lambda igual a 0.5 son "Israel", "gaza", "melilla", "voto_correo" o "ciudad". Su nube de palabras es:

Figura 41: Nube de palabras para el conflicto árabe-israelí

Fuente: Elaboración propia

Aquí hay dos grupos fundamentales de noticias, como ya se ha venido advirtiendo en el párrafo anterior: a) las relacionadas con el conflicto árabe-israelí, como "Netanyahu", "Israel", "franja" o "gaza"; y, b) las que tienen que ver con la presunta compra de votos, como "compra_votos", "interior" o "coalición_melilla". Algunas palabras que no tienen mucho que ver, como "persona", "ine" o "crecimiento" responden a indicadores económicos. Probablemente el modelo, a partir de los datos del paro que se refieren a personas concretas, ha hecho extensivos estas noticias a este *topic*, por incluir otras

noticias relacionadas con personas (ej. desplazamientos de personas en la franja de Gaza). La distribución temporal es la siguiente:

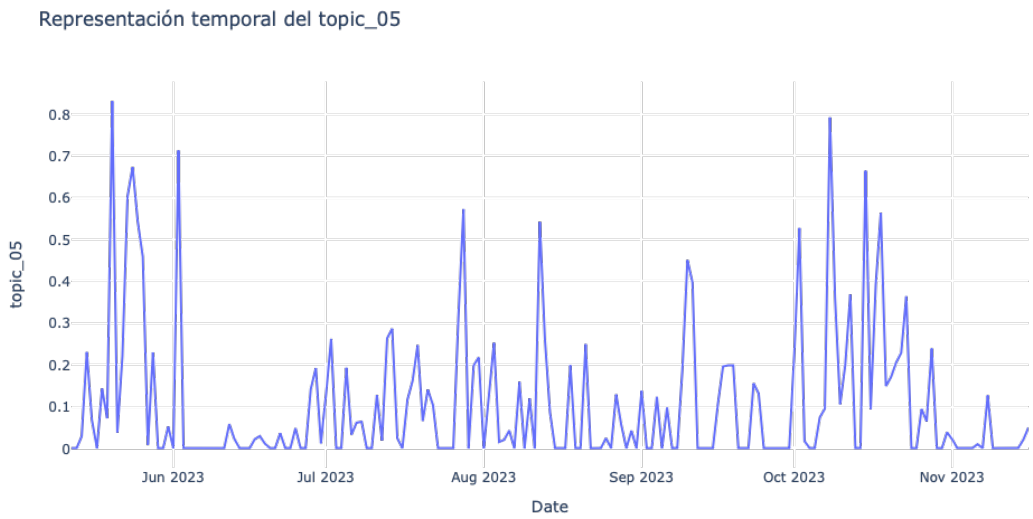


Figura 42: Distribución temporal del conflicto árabe-israelí

Fuente: Elaboración propia

Los mayores picos se dan en mayo, por los votos y en octubre, por el conflicto. En este sentido, nos parece lógica la distribución temporal. Otros picos tienen que ver con los datos económicos mencionados, singularmente el pico del 28 de julio o del 12 de agosto. La inclusión de estos datos económicos es lo que ha provocado que, aunque los valores sean palmariamente reducidos sin perjuicio de algunos picos, exista una distribución que pueda parecer confusa en algunos registros.

5.6. *Topic* 06: El último acto de las negociaciones

Este es el último *topic* relacionado como las conversaciones que produjeron la investidura de Pedro Sánchez. Si bien es cierto que los dos anteriores se centraban en cuestiones más generales -nacionalistas e independentistas catalanes- este *topic* se centra en las negociaciones con un partido en concreto: *Junts*. En consonancia con lo ya citado en el marco teórico, *Junts* era un partido que tradicionalmente no había venido apoyando a Pedro Sánchez, de ahí que las negociaciones con ellos resultasen sorprendentes. Los rumores de la negociación con Carles Puigdemont, líder de la formación, coincidieron temporalmente con la compra de Telefónica que también avanzamos, por lo que en el *topic* se verá reflejado este hecho. Con todo, resulta llamativo que las negociaciones con *Junts* hayan dado pie a un *topic* totalmente independiente. Este *topic* contiene el 7.3% de los *tokens* y sus palabras más repetidas con lambda igual a 0.5 son "Puigdemont", "junts", "acuerdo", "telefónica" y "gobierno".

La nube de palabras de este *topic* es la siguiente:

Figura 43: Nube de palabras para el último acto de las negociaciones

Fuente: Elaboración propia

Las palabras más destacadas están relacionadas de forma clara con las negociaciones con *Junts*: "acuerdo", "Cataluña", "carles_puigdemont" o "pedro_sánchez". Es curioso como el hecho de que las negociaciones fueran secretas haya provocado que la palabra

"fuente" y "fuentes" aparezcan en esta nube. Como avanzamos al introducir el *topic*, también hay palabras relacionadas con Telefónica tales como "telefónica", "inversión" o "stc". No logramos explicar la inclusión de estos términos, con la única razón siendo la coincidencia temporal o un fallo en la conformación de los *topics* por parte del modelo. En general, puede decirse que es el único o de los únicos *topics* en el que aparecen cuestiones económicas, al mencionar términos como "déficit" o "cuentas". Sin embargo, aparecen en pocas ocasiones, tienen tamaño reducido y seguramente están incluidos por extensión de lo sucedido en Telefónica. La distribución temporal es la siguiente:

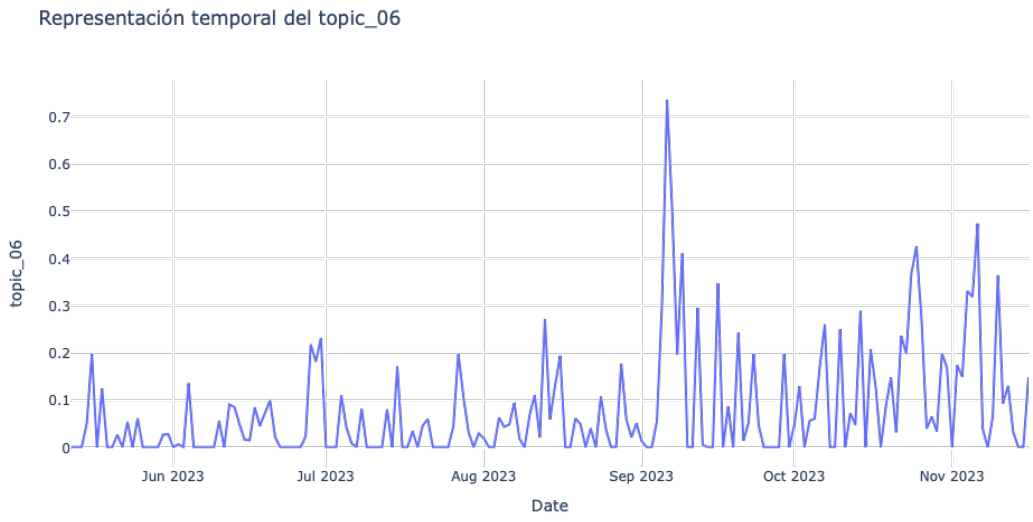


Figura 44: Distribución temporal del último acto de las negociaciones

Fuente: Elaboración propia

La distribución tiene sentido. Sin perjuicio de que se mencionase un poco durante los primeros meses, la auténtica relevancia se produce a partir de septiembre, con el inicio de las negociaciones con *Junts*. Alcanza el pico en el día de la entrada de STC en Telefónica, alrededor de la primera semana de septiembre y en el mes de noviembre, donde aparecen picos en aquellos días en los que hubo hitos en estas negociaciones, como el seis y 10 de noviembre, cuando se rubricó el pacto en Bruselas y cuando se anunció el acuerdo, respectivamente. El 25 de octubre hubo otro pico puesto que se consensuó el pacto con Sumar, que se esperaba provocase un efecto propulsor en las negociaciones, acelerándolas de forma exponencial. En general, se trata de un *topic* correctamente ajustado partiendo de la base de que hay palabras que seguramente no deberían estar incluidas.

5.7. *Topic* 07: Crónica de una elección perdida

Este es el mayor *topic* de todos los mencionados y probablemente uno de los más polémicos. Abarca el 48.2% de los *tokens*, es decir, prácticamente la misma cantidad que el resto de *topics* sumados. El resumen de este *topic* es que, si por un lado hemos hablado de las negociaciones para la investidura del Presidente Sánchez, en este toca relatar lo que sucedió con la derecha. Aquí se tratan tanto los pactos entre el PP y VOX, como el relato demoscópico de la victoria de Feijóo y su intento de investidura. Esto, con la oposición constante de la izquierda tanto a los pactos como a la victoria, sumado a su estado efervescente de felicidad a partir de las elecciones generales. Es, en definitiva, la consagración de un relato que se demostró falso en cuanto finalizó el recuento y que ha sido percibido como agonizante durante el resto del periodo, habida cuenta del fracaso del intento de investidura de Feijóo, fracaso ya conocido de antemano. Es un *topic* cuya intensidad es elevada durante todo el estudio de campo, con una distribución que de no ponerse en contexto podría interpretarse como el más claro ejemplo de ruido blanco, es decir, información que no aporta realmente mucho. Los cinco términos más frecuentes con lambda igual a 0.5 son “pp”, “feijóo”, “vox”, “psoe” y “partido”.

La nube de palabras para este *topic* es la siguiente:

Figura 45: Nube de palabras para la crónica de una elección perdida

Fuente: Elaboración propia

Como puede verse claramente existe una cantidad sustancial de términos relacionados con la derecha. Estos son “Alberto_núñez”, “vox”, “feijóo” o “populares”. También los hay relacionados con la izquierda como “Sánchez”, “psoe” o “Bildu”. Puede verse de forma nítida que existen palabras relacionadas con todos los momentos del ciclo electoral: a) autonómicas, “Madrid”, “Extremadura”, “comunidad_valenciana” o “guardiola” tres comunidades en las que VOX o apoyaba al Gobierno ya existente del PP o rubricó un pacto con el PP, así como el nombre de la ya mencionada Presidenta del Consejo de Gobierno de Extremadura María Guardiola; b) generales, “debate”, “cara_cara”, “noche_electoral” o “mayoría_absoluta”, este último término especialmente relacionado con los pronósticos ya enunciados en el marco teórico; c) posgenerales, “pnv”, “apoyo” o “imposible”, relacionadas bien con la imposibilidad de formar Gobierno o la necesidad del apoyo del PNV para conformar una coalición que desbancase a Sánchez, cosa que no sucedió; d) final del estudio de campo, “ferraz” o “violencia”.

La izquierda tiene representación con términos como “psoe_sumar”, “socialista” o “estrategia”, lo que hace también referencia al rearme del PSOE ante un escenario complejo. Hay palabras no relacionadas tales como “salvador” o “Ignacio” que aunque pudieran coincidir con nombres, no nos terminan de parecer relevantes o identificables. La distribución temporal es la siguiente:

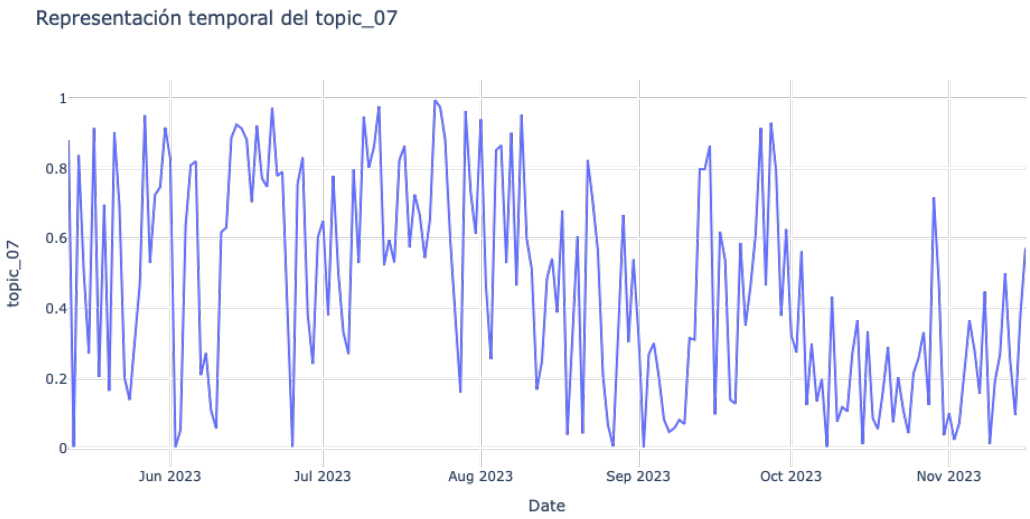


Figura 46: Distribución temporal de la crónica de una elección perdida

Fuente: Elaboración propia

Aquí se ve claramente lo que ya se ha avanzado: parece como si la distribución fuera ruido blanco. No obstante y dentro de este aparente caos, pueden sacarse conclusiones

muy claras. En primer lugar, la tendencia es decreciente, lo que parece lógico si se tiene en cuenta como se fueron desinflando las expectativas de la derecha a medida que avanzaba el estudio de campo. Puede verse que desde finales de verano, lo que encaja con la constitución de la Mesa del Congreso, acontecimiento que puso fin a las expectativas de Gobierno de la derecha, la tendencia desciende de forma brusca y pronunciada. Los picos posteriores, especialmente los acontecidos a finales de septiembre y principios de octubre tienen que ver con la investidura fallida del PP.

La brusca caída producida en julio es del día 25, donde los diarios empezaron a vislumbrar que el único con opciones de una investidura real era el Presidente Sánchez. No obstante, la expectativa se estuvo manteniendo durante el verano. Por último, el último pico es del día 28 de octubre y es relativo a varias cuestiones como el discurso que iba a pronunciar la Presidenta del Congreso en el día de la jura de la Princesa Leonor, prevista para el día 31 de octubre y en la que se hicieron varios comentarios que los medios interpretaron como un ataque encubierto a la derecha.

5.8. Análisis de cada medio de comunicación

Habiendo estudiado ya cada *topic* de manera exhaustiva, es procedente analizar cuánta importancia le ha dado a cada *topic* cada uno de los diferentes medios. De esta manera, podremos ver cuál ha sido su contribución al panorama global, es decir, cuál fue su agenda durante el ciclo electoral del año 2023.

5.8.1. *Diario El Mundo*

La distribución temporal de los siete *topics* en este periódico es la siguiente:

Figura 47: Distribución temporal de los *topics* para el Diario El Mundo

Fuente: Elaboración propia

Atendiendo a un análisis pormenorizado:

Frente amplio de izquierdas: El primero se trata durante el mes de junio con la relevancia propia que tuvo el frente de izquierdas. Esto es algo común en todos los periódicos. La pregunta en realidad reside más en el tratamiento que tiene el *topic* una vez constituido Sumar. El Mundo, de forma similar a El Confidencial, trata el tema poco a partir de septiembre, con un único pico, en realidad un *outlier,* el día 12 de septiembre, donde se informó de la manifestación de la *Diada*.

Negociaciones sorprendentes: Durante los primeros meses es el diario que menos trata estos asuntos. Otros periódicos, como veremos, sí que refieren con cierta intensidad las negociaciones de Sánchez y la amnistía. A diferencia de lo que pueda esperarse de un periódico con percepción ideológica cercana al siete, la amnistía aquí no se menciona hasta más adelante. El primer pico, del ocho de julio, está relacionado con el empleo. En cambio, el valor máximo se alcanza el uno de septiembre, cuando Íñigo Urkullu, entonces *lehendakari* del Gobierno vasco, escribió una carta dirigida a Sánchez para exigirle más concesiones a las "nacionalidades históricas" como consecuencia del resultado electoral.

El último gol de Rubiales: Se aborda el tema de Rubiales con relevancia, quizá menor a la de otros diarios. Los picos que no coinciden con el tiempo de los sucesos están relacionados con la justicia, debido a la enorme discusión que se generó sobre la existencia o no de responsabilidad penal en el caso del beso no consentido.

La Mesa del Congreso y sus expresiones lingüísticas: Aquí el tratamiento del *topic* es muy llamativo. La cuestión de la Mesa del Congreso pasa casi imperceptiblemente para El Mundo, con un peso muy bajo. Tienen más importancia aspectos accesorios, como los picos del siete y 13 de noviembre, en los que se hace referencia a la Presidenta del Congreso ejerciendo funciones propias de su cargo, como fijar plenos. Los demás picos están relacionados con la UE y la posible oficialidad de las lenguas.

El conflicto árabe-israelí: El Mundo, junto con el Diario ABC y El Confidencial, no dan tanta importancia al conflicto árabe-israelí, como si hace El País, mencionándolo escasamente. Como ya explicamos, la inclusión de noticias económicas es común en este *topic*, lo que explica los picos previos al conflicto.

El último acto de las negociaciones: Es probablemente el *topic* más trabajado por este diario. Hay picos muy juntos, debido a la concentración de las negociaciones antes y después de acontecimientos singulares, como la Mesa del Congreso o la investidura. El mayor pico previo en el tiempo tiene que ver con la modificación de la oficialidad del

valenciano y catalán en la Comunidad Valenciana y Baleares, enmarcado en pactos del PP y VOX. Por las similitudes lingüísticas, se le ha dado más peso en este *topic*.

Crónica de una elección perdida: Como sucede en todos los periódicos y ya veníamos infiriendo del propio *topic*, este tema se trata con mucha frecuencia por todos, habida cuenta no sólo de la cantidad porcentual de *tokens* que contiene, sino también de la relevancia durante todo el ciclo electoral.

5.8.2. *Diario El País*

La distribución de los diferentes *topics* para el Diario El País es la siguiente:

Figura 48: Distribución temporal de los *topics* para el Diario El País

Fuente: Elaboración propia

Procedemos a realizar una serie de apreciaciones en relación con cada *topic*:

Frente amplio de izquierdas: Este diario es, nítida y palmariamente, el que más menciona el tema. No sólo durante el tiempo en que se rubrica el pacto de la izquierda, sino también de forma constante a partir de septiembre, para ilustrar como avanzaban las negociaciones para investir a Pedro Sánchez. Por ello, resulta paradójico que no sea el que mayores picos presenta con posterioridad al pacto, donde destaca el diario ABC. No obstante, es el que de forma sostenida alude más al pacto.

Negociaciones sorprendentes: Igualmente El País, junto con el ABC y El Español, es el diario que más persistentemente alude a las negociaciones para la investidura de Pedro Sánchez con carácter previo a las elecciones generales. Pese a ello, los picos son reducidos y tienen grandes llanuras, ya que no se circunscribe a las negociaciones de investidura, sino a otras negociaciones como las relacionadas con el carácter social de las medidas propuestas por el PSOE y la izquierda. Así, el pico del 6 de julio corresponde a las promesas sociales del PSOE ante las elecciones generales. La carta del *lehendakari* Urkullu es la que explica el mayor pico, especialmente porque la publicó en este diario, siendo su peso mayor que en otros al tratarse de una fuente directa.

El último gol de Rubiales: El País es el diario que trató este *topic* con una intensidad mayor, tanto en el momento que sucedió como anteriormente. De hecho, fue el único que hizo un seguimiento un poco más profundo de la trayectoria en el mundial de la selección femenina de fútbol, sin limitarse al incidente. El pico del 19 de octubre, que es posterior, aborda la discriminación por razón de nacionalidad. Por lo tanto, podría entenderse más como un *outlier* que como una noticia estrictamente relacionada con el tema de Rubiales.

La Mesa del Congreso y sus expresiones lingüísticas: Aquí El País vuelve a diferir, en parte, del resto de los diarios. En primera instancia, el tratamiento es ciertamente desigual. Existen picos previos a la negociación de la Mesa, referidas especialmente a las acciones de la justicia con Puigdemont, singularmente la retirada de la inmunidad parlamentaria, expresada en el pico del cinco de julio. En segundo término, la llanura que existe entre finales de septiembre y principios de octubre, en la que otros diarios seguían resaltando la aprobación de la modificación del Reglamento. Finalmente, es escasamente referida la cuestión lingüística.

El conflicto árabe-israelí: El efecto arrastre de la palabra "personas" se produce fundamentalmente en este diario, que también presta una mayor atención al conflicto árabe-israelí. También da mucha importancia a la cuestión de Melilla y es sin duda el que

más aborda temas económicos como los datos del paro, afiliación a la seguridad social y cuestiones relacionadas en general con el empleo o el producto interior bruto.

El último acto de las negociaciones: Este *topic* destaca por la enorme llanura que existe hasta prácticamente el mes de agosto o septiembre. Las negociaciones con *Junts* se mencionan sucintamente, centrándose en otras cuestiones como la constitucionalidad de la amnistía en el segundo *topic*, ignorando todas las negociaciones. El pico es del día seis de septiembre y corresponde a la compra de Telefónica por el fondo STC. Este diario lo que suele mencionar en este *topic* son exclusivas, como la del seis de octubre, en la que informaron de que el PSOE acudía a Bruselas a reunirse con Puigdemont.

Crónica de una elección perdida: De nuevo aquí el *topic* es irregular pero frecuente, con altos y bajos que no terminan de dibujar un patrón claro salvo que se menciona mucho y muy intensamente, al igual que en otros medios. Ninguno destaca de manera significativa.

5.8.3. *El Confidencial*

La distribución de los *topics* para El Confidencial es la siguiente:

Figura 49: Distribución temporal de los *topics* para El Confidencial

Fuente: Elaboración propia

´Las cuestiones a destacar son las siguientes:

Frente amplio de izquierdas: El Confidencial no confiere especial importancia a este tema pues, aunque lo trata, no vuelve a mencionar cuestiones sobre la "izquierda a la izquierda del PSOE" a partir de la publicación del pacto y comunicación de la coalición a la Junta Electoral. Hay algunos picos posteriores, como el del cuatro de octubre, en el que se valoran las potenciales carteras ministeriales reservadas para Sumar de conformarse un un Gobierno en coalición con el PSOE. Algunas menciones en julio, singularmente el día siete, prestan atención a la campaña de las diferentes formaciones que se adhirieron a Sumar.

Negociaciones sorprendentes: El Confidencial trata este *topic* con una intensidad estándar, pero con distinto enfoque al resto, sin excesiva ausencia ni presencia. El mayor pico es del 19 de agosto, dos días después de la constitución de la Mesa del Congreso, con un sonoro mensaje: los Letrados del Congreso consideran que la amnistía es inconstitucional. En consecuencia, su enfoque es más jurídico que político. También publica una pluralidad de artículos sobre la Casa Real, muchos más que el ABC, diario tradicionalmente monárquico. Ejemplos de ello son los picos del 26 de septiembre, 10 de octubre, 27 de octubre o uno de noviembre. El último pico es del 11 de noviembre, donde también habla de la amnistía.

El último gol de Rubiales: En el caso de Rubiales, El Confidencial mantiene, de nuevo, un tratamiento estándar. Es verdad que, durante el tiempo de duración de la noticia, dedicó tres noticias monográficas e hizo referencia a avances del caso en todas las portadas. Sin embargo, estas noticias en muchos casos no han pasado el filtro porque no ocupaban la portada principal de la edición matutina, lo que nos permite reflexionar acerca de futuras líneas de investigación que posteriormente trataremos. Los picos previos hacen referencia a cuestiones culturales con sesgo político. Por ejemplo, la petición de VOX al PP de ostentar el Ministerio de Cultura en un hipotético Gobierno de coalición. Como el fútbol suele asociarse con el mundo de la cultura, parece lógica la asociación.

La Mesa del Congreso y sus expresiones lingüísticas: Aquí se destaca un tratamiento sensiblemente inferior al resto de medios. Hay tres picos: 18 de agosto, ocho y 20 de septiembre. El primero se relaciona con la problemática de las lenguas y los demás con la justicia. Es llamativo como El Confidencial tiende a dar un enfoque jurídico o económico a los temas en detrimento de lo que sería el contenido puramente político que se observa en otros medios.

El conflicto árabe-israelí: Este medio vuelve a destacar por el efecto arrastre de la palabra personas. El Confidencial, junto con El País, es el diario que más publica sobre economía. Todos los picos previos al conflicto árabe-israelí abordan aspectos relativos a la economía, al paro y al crecimiento económico. No obstante, su enfoque es diferente, puesto que mientras El País celebra los datos económicos, El Confidencial tiende a poner su veracidad en barbecho. De nuevo surge una potencial nueva línea de investigación.

El último acto de las negociaciones: El *topic* en El Confidencial combina cuestiones económicas, como la de Telefónica, con referencias a la amnistía. Los picos previos hablan de las cesiones previas y la amnistía a *Junts* comienza a aparecer en los picos del seis, nueve y 12 de septiembre. Aun así, el grueso de la información surge a partir del anuncio formal de la amnistía, cercano al final del estudio de campo.

Crónica de una elección perdida: Como hemos venido anunciando de forma reiterada, la información de este *topic* es reiterativa durante todo el estudio.

5.8.4. *Diario ABC*

La distribución temporal de los *topics* es la siguiente:

Figura 50: Distribución temporal de los *topics* para Diario ABC

Fuente: Elaboración propia

Aquí podemos resumir la siguiente información:

Frente amplio de izquierdas: El Diario ABC es, junto a El País, el que más referencia este *topic*. Si bien es cierto que no hay un peso muy elevado salvo fechas puntuales, se mantiene una constancia reseñable. No sólo se aborda el pacto, sino también los conflictos que surgieron dentro del mismo, como es el caso del pico del seis de agosto. En septiembre vemos picos también, pero relacionados especialmente con el desfile del 12 de octubre, que generó cierta polémica entre los integrantes de Sumar.

Negociaciones sorprendentes: Junto con El Español, ABC se corona como el medio que más menciona este *topic*. No sólo por la intensidad, que se mantiene prácticamente desde la noche electoral, sino por la constancia. Se enfoca tanto la amnistía como las posiciones de todos los actores y se publican editoriales relacionados con el tema. Siempre que se tratan las negociaciones, como en el día cuatro de octubre, ABC las trata como capitulaciones o cesiones, diferenciándose del resto. En términos generales es más prolijo en el uso de adjetivos, como sucede en el pico del 14 de septiembre. También hay muchas referencias a la monarquía, siempre acompañadas de alusiones a la unidad de España, como en el pico del ocho de octubre.

El último gol de Rubiales: ABC y El Español son los medios que menos alusiones hacen a este *topic*. Las noticias de ABC versan en torno a la actuación gubernamental, que en algunos titulares como los de los días dos o tres de septiembre, califica de poco eficaz. En términos generales, ABC no profundiza en exceso en este tema, sin perjuicio de la preceptiva mención al incidente y a la victoria en el mundial.

La Mesa del Congreso y sus expresiones lingüísticas: Las lenguas son un tema recurrente en ABC, siendo el medio que más habla de este *topic*, en términos absolutos y en términos de consistencia. Se aborda el tema catalán desde una perspectiva tanto crítica con los independentistas como constante: es el único medio que sigue hablando de los acontecimientos catalanes de 2017 desde una perspectiva dura, destacándose el 19 de junio, donde dedicó un reportaje entero a *Tsunami Democratic,* mientras que el resto del medios hablaban del Ayuntamiento de Barcelona.

El conflicto árabe-israelí: Sin perjuicio de que pudiese haber informado sobre el conflicto árabe-israelí en otras noticias, no le dedicó la portada en ningún día. Mientras que todos los medios experimentan enormes picos en el momento del conflicto, ABC dedica sus portadas a las negociaciones de Sánchez. La única similitud que vemos, además del efecto arrastre de algunos datos económicos, es la del dos de octubre, donde

se informa sobre las discotecas incendiadas de Murcia, que debido a los términos empleados ha terminado por incluirse aquí. En general, llama mucho la atención el tratamiento de este tema.

El último acto de las negociaciones: La ausencia de menciones de ABC sobre este *topic* es inusual. La lógica nos indica que la mayoría de las referencias al independentismo son realizadas directamente en los *topics* dos y cuatro. Los mayores valores son: a) el 16 de mayo, relativo a los costes de despido y que probablemente se trate de un *outlier*; b) el 13 de agosto, donde se habla del boicot de los independentistas a la Vuelta ciclista; y, c) el 25 de octubre, donde se habla de las nefastas consecuencias del nuevo Gobierno que estaba en formación. No hay referencias a Telefónica.

Crónica de una elección perdida: En este punto es interesante que sea el único medio que tiene pesos elevados en los días del intento de investidura de Feijóo: días 26 y 27 de septiembre. El resto de los medios o lo tienen en uno o no los tienen tan elevados. El Diario ABC, entre toda la dispersión del *topic*, alcanza en estos días sus mayores pesos.

5.8.5. *El Español*

Como último medio de este trabajo, conviene analizar la distribución temporal de El Español:

Figura 51: Distribución temporal de los *topics* para El Español

Fuente: Elaboración propia

Respecto de las cuestiones más representativas de cada *topic*:

Frente amplio de izquierdas: En primer lugar, El Español menciona de forma estándar el tema, ni en exceso ni quedándose corto. No obstante, es llamativo el peso del día cuatro de septiembre, donde incluye una noticia cuyo sitio más apropiado sería el *topic* 03, pues habla de Rubiales. El modelo probablemente lo haya incluido aquí porqué trata sobre las reacciones de Irene Montero, entonces Ministra de Igualdad y parte de Sumar, al asunto de Rubiales. Aun así, es paradójico que entre los dos temas, el modelo haya optado por este.

Negociaciones sorprendentes: El Español ha hecho alusión a las negociaciones de Pedro Sánchez en una pluralidad de ocasiones a lo largo del estudio, al igual que ABC. Pero presenta un matiz que pasa inadvertido a primera vista. La escala de este gráfico presenta una singularidad: el máximo valor es 0.8, no 1, siendo el único de los treinta y cinco -uno por cada *topic* y medio- cuyo valor máximo no pasa del 0.8. La conclusión es clara: si bien es cierto que hace muchas referencias a los pactos de Pedro Sánchez, tiene una menor intensidad que el resto, dándole a las noticias otro contenido. El máximo valor figura al día siguiente de la constitución de la Mesa del Congreso, 18 de agosto. Se atiende a la negociación como trámite para lograr la investidura, no como pacto.

El último gol de Rubiales: El Español es el que menos trata este *topic*. Hay abundantes llanuras, singularmente la del día cuatro de septiembre, adjudicado al frente amplio de izquierdas. El único pico reseñable es el del seis de julio, cuyo objeto tiene más que ver con la justicia, asociada a este por el posible reproche penal de la actuación de Rubiales más que a la polémica de Rubiales en sí.

La Mesa del Congreso y sus expresiones lingüísticas: Aquí hay presentes también varias llanuras, habida cuenta de la perseverancia que ya presenta el segundo *topic,* con picos relacionados con otros nacionalismos, como los iniciales y previos a la constitución de la Mesa. Uno de los máximos valores se alcanza el 19 de septiembre, expresando de forma contundente que el Congreso se había convertido en el primer Parlamento que, a pesar de contar con una lengua común, empleaba pinganillos.

El conflicto árabe-israelí: El máximo valor aparece el día del conflicto, abandonándolo al día siguiente, donde se enfoca en otros temas. Hay otros picos, pero todos arrastrados por los desplazamientos de personas en la franja de Gaza. De hecho, el empleo reiterativo de la palabra personas provoca que muchos temas económicos se subsuman aquí de forma errónea. En líneas generales, se observan muchas llanuras.

El último acto de las negociaciones: Este es el *topic* al que El Español más ha contribuido. Desde el 27 de julio, día de máximo valor donde la noticia eran las exigencias concretas de Puigdemont, El Español no ha parado de publicar titulares relacionados con estas negociaciones. Destacan los de los días seis, siete y ocho de septiembre, donde el primero se refiere a potenciales negociaciones de Sánchez con Telefónica para frenar la entrada de los saudíes. Como ya detallamos en su momento, la cuestión de Telefónica, probablemente por coincidencia temporal, se ve arrastrada a este *topic*.

Crónica de una elección perdida: Para concluir con el análisis y dar paso a las conclusiones, podemos ver que en este caso hay un progresivo descenso a medida que las posibilidades de la derecha se frustran. Sin embargo, se observa una revitalización posterior con las movilizaciones en contra de la amnistía.

CAPÍTULO IV: CONCLUSIONES, PROBLEMAS Y FUTURAS LÍNEAS DE INVESTIGACIÓN

1. CONCLUSIONES

Primera. Los medios de comunicación han tenido, al menos durante la totalidad del estudio de campo, un comportamiento eminentemente político. Resulta paradójico que en el estudio, a pesar de tratarse de un ciclo electoral, no se traten en ningún caso los problemas de los electores. La economía es un tema completamente olvidado, las preocupaciones ciudadanas se encuentran totalmente desnaturalizadas y la totalidad del estudio de campo habla sobre la política en mayúsculas, pero no sobre medidas o proposiciones. Si se aplicase analógicamente a la analítica de datos, el estudio se limitaría a describir los metadatos, no entrando salvo contadas excepciones a explicar los datos o el porqué de los datos, lo que sería equivalente a no explicar por qué se solicita el voto.

Segunda. Los medios de comunicación tienen, sin género de dudas, una marcada agenda ideológica. Esta agenda no sólo se entiende por la intensidad con la que abordan los temas, sino también por cómo los abordan. Cuando se estudian las contribuciones de cada medio a la imagen global del estudio y se pone en contexto con la información real, las consecuencias en términos de objetividad son demoledoras para la práctica totalidad de medios y noticias estudiadas. Si bien es cierto que algunos lo disimulan mejor, hay ejemplos palmarios en los que el componente político tiene una influencia quizás excesiva, hasta el punto de desvirtuar la información o no referenciarla en absoluto, es decir, ignorándola. Habida cuenta de que no podemos confundir objetividad con neutralidad, tampoco se debería confundir subjetividad con activismo.

Tercera. Hay paralelismos sorprendentes entre los medios. Tras analizar el estudio, puede concluirse que son los medios que más a favor o más en contra están de un tema los que se ocupan de esto en mayor medida. En muchos casos, para entender dónde está la objetividad, hay que estudiar, además de la información, su reiteración temporal, en detrimento de otros parámetros más lógicos como la intensidad. Así, es reseñable el comportamiento de los diarios El País y ABC. Si bien es cierto que el ciudadano medio no se equivoca cuando enuncia que sus enfoques son ideológicamente opuestos, la

realidad es que -con algunas salvedades- se refieren a algunos temas con una insistencia idéntica o enormemente parecida. Por ejemplo, son los medios que más se refirieron al frente amplio de izquierdas.

Cuarta. Vivimos en una sociedad enormemente globalizada, en muchos casos profundamente insensibilizada y, en cierta medida, esto puede ser atribuible a los medios de comunicación. Durante el ciclo electoral ha habido una rebelión de un grupo armado, un ataque terrorista y la correspondiente invasión de un territorio en respuesta. Sin embargo, los medios le dan un tratamiento, a nuestro parecer, demasiado superficial. Lo que hoy es noticia deja de serlo tan pronto como nuestro estado mental cambia, generalmente en poco más de una semana o incluso un día. Las noticias pueden incluso llegar a confundirse con *outliers* aditivos, en tanto que muchas de ellas son simplemente la desviación de una serie de temas predeterminados o cuya presencia en las portadas se da por hecha. No existen apenas situaciones de transitoriedad o de reducción paulatina de la relevancia de las noticias, sino que llegan un día y son desplazadas al siguiente.

Quinta. Que no es lo mismo la opinión pública que la opinión publicada. Al hacer la extracción y análisis de las noticias se infieren cuestiones muchas veces sorprendentes, pues en muchos casos los titulares no coinciden con lo que termina sucediendo, especialmente en el marco de las elecciones generales, donde la realidad superó -con creces- el paisaje mediático. Es probable que estemos entrando en una época en la que los medios, especialmente a raíz de divergencias tan grandes como la de julio, deban buscar fórmulas alternativas para conectar con una audiencia que, cuando se pronuncia, al menos electoralmente, lo hace de forma totalmente contraria a ellos.

Sexta. Que la percepción de los ciudadanos acerca de los medios no tiene por qué ser cierta en todos los extremos. Si se observa detenidamente el tratamiento dado a algunos temas, existen singulares excepciones de medios que, a pesar de ser catalogados como de una ideología determinada, luego no hacen honor a la misma. Por ejemplo, El Confidencial es un medio que, para los ciudadanos, es ideológicamente “de derechas”. Sin embargo del estudio se infiere que no realiza una crítica política, en tanto que se opone al Gobierno en el sentido más técnico. De aquí también pueden extraerse conclusiones más desoladoras, como si estamos o no entrando en una época de polarización en la que,

o los medios dicen lo que el usuario quiere leer, o automáticamente pertenecen a la ideología contraria.

2. PROBLEMAS Y FUTURAS LÍNEAS DE INVESTIGACIÓN

Por último, nos gustaría destacar ciertos problemas y posibles líneas de investigación futuras.

En primer lugar, expondremos los problemas evidenciados. El primero es claro: la ausencia de diversidad de temas en los medios. Esta escasez, motivada por el excesivo componente político de las noticias, ha hecho que este análisis parezca parco. Hemos analizado siete temas que parecían reconducirse, de una u otra manera, a las negociaciones, los pactos, las encuestas o cuestiones siempre eminentemente políticas. Es difícil valorar si el análisis ha sido infructuoso, o si los medios realmente no han hablado de otra cosa que no sea la política, dejando otros temas apartados o circunscritos a lo que tradicionalmente se conoce como "edición salmón", más centrada en la economía. El más claro ejemplo es que, de sumarse todos los temas políticos, ocuparían aproximadamente el 85 por cien de los *tokens*, con uno de ellos -el 07- conteniendo casi el 50 por cien de estos, lo que dificulta mucho el estudio individualizado de los acontecimientos.

El segundo problema es que, a pesar de haber valorado siete temas, todos presentaban *outliers* que en muchos casos hacían complejo el análisis. En ciertos puntos era frustrante ver como había puntos con pesos muy elevados pero de contenido poco relacionado con el tema. No obstante, de probarse otras combinaciones de *LDA PASSES* y *N_TOPICS*, los resultados eran igual de desesperanzadores. Realmente no existe en este momento una respuesta clara para este problema. Esto ha supuesto un problema para lograr el objetivo del trabajo, que no era otro que extraer conclusiones lógicas acerca de unas noticias concretas.

El tercer problema ha sido la falta de noticias. Aunque 945 noticias pueden parecer muchas, en la práctica no es así. No son suficientes para obtener una relación suficiente de términos -ligeramente superiores a 1.000- ni para conseguir un análisis exhaustivo. Es imperativo que, de abordarse un nuevo análisis, se haga con un número de noticias que ofrezcan una imagen más global. Esta imagen más global quizás logre paralelamente resolver el problema del exceso de contenido político en las portadas.

Como último problema, se evidencia la excesiva tendencia de los medios a criticar al Gobierno. Esta proposición no busca dar la impresión de opinión política o de denuncia acerca de la "derechización de los medios", sino que busca exponer su labor opositora. Esta labor está especialmente relacionada con el primer problema. La oposición al

Gobierno nubla el análisis y el diagnóstico, pues convierte al Gobierno, como órgano, en el centro de la crítica. Esto obliga a que los medios que tienden a defender al Gobierno, en lugar de hablar de sus posibles logros, se vean desplazados a un marco mental de suma cero en el que o estás con el Gobierno o contra él. En este sentido, creemos que la única perdedora es la información.

Por último y en cuanto a las futuras líneas, la primera es que, debido a la ausencia de espacio, no ha sido posible realizar un *sentiment analysis* que seguramente mejoraría el estudio. Creemos que lo mejoraría por las divergencias que podrían observarse en el tratamiento de la información. Hemos ido dando pistas a lo largo del trabajo acerca de la existencia de un tratamiento desigual de la información, pero esto no dista de ser empíricamente pobre, pues se basa en las observaciones del escritor. Sería analíticamente interesante -aunque a nivel de espacio imposible- observar el contenido en términos más emocionales. Esto también permitiría realizar un análisis mucho más profundo e interesante sobre los diferentes grados de oposición al Gobierno.

En segundo lugar, habría sido positivo un análisis de ciertos términos, estudiando su evolución temporal. De esta manera, podrían valorarse mejor las diferencias entre la opinión publicada y la opinión pública o realidad. Como ejemplos de interés serían: "mayoría_absoluta", "perdón", "amnistía" o, singularmente, "gaza" e "Israel". En nuestra opinión, ilustraría mucho mejor la idea de la sociedad globalizada mencionada en las conclusiones.

En tercer y último lugar, consideramos que realizar una revisión más profunda optimizaría el diagnóstico. No obstante, no habrá análisis más profundo si no se incluyen más noticias, más *inputs* al modelo, pues sin ello es más complicado profundizar en el paisaje de la realidad social española durante el ciclo electoral de 2023. De hecho, esto permitiría apreciar si, a pesar de no aparecer en portada, los medios mantienen transitoriedad en la información. Hasta entonces no cabe concluir otra cosa que, a nuestro juicio y respecto de ciertos temas, los medios tienden a tomar la información, publicarla y olvidarse de ella. En la práctica, puede entenderse como un ciclo triturador de sucesos de escaso recorrido.

BIBLIOGRAFÍA

Ardèvol-Abreu, A. (2015). Framing o teoría del encuadre en comunicación. Orígenes, desarrollo y panorama actual en España. *Revista Latina de Comunicación Social, 70,* 423-450.

Arroyo Cabello, M. (2006). La prensa en la democracia (1986-2005). En Arroyo Cabello, M. y Roel Vecino, M. (coords.), *Los medios de comunicación en la democracia (1986-2005): prensa, radio y televisión* (13-59). Fragua.

Asociación para la Investigación de los Medios de Comuncación (2024, 1ª Ola). *Estudio General de Medios.* Recuperado el 28 de mayo de 2024, de: EGM

Ayala Sörenssen, F. (2023, 3 de marzo). ABC cumple 120 años. *ABC.* Recuperado el 28 de mayo de 2024, de: aniversario

Bhanot, N., Singh, H., Sharma, D., Jain, H., & Jain, S. (2019). Python vs. R: A Text Mining Approach for analyzing the Research Trends in Scopus Database. *arXiv preprint arXiv:1911.08271*.

Carreño, A. (2023, 15 de junio). La 'geometría variable' de Feijóo, el plan de acuerdos plurales limitado por la dependencia de VOX. *El Independiente*. Recuperado el 26 de mayo de 2024, de: precampaña

CNN Español (2023, 12 de octubre). Cronología: así fueron, paso a paso, el ataque de Hamás y la respuesta de Israel. *CNN en Español.* Recuperado el 28 de mayo de 2024, de: hamás

Collins (s.f.). Bigram. En *Collins Dictionary*. Recuperado el 29 de mayo de 2024, de: bigrama

Del Arco Bravo, M. A., Yunquera Nieto, J. y Pérez Bahón, F. (2016). Los cien primeros días de *El Español*. Análisis de la estructura y los contenidos en los inicios del diario digital. *Revista Latina de Comunicación Social, 71,* 527-551.

Díaz, B. (2023, 5 de septiembre). Saudí Telecom compra el 9,9% de Telefónica por 2.100 millones y se convierte en el primer accionista. *CincoDías*. Recuperado el 28 de mayo de 2024, de: STC

Domènech, J. (2023, 11 de septiembre). Cronología del caso Rubiales: las fechas clave del escándalo. *El Periódico*. Recuperado el 28 de mayo de 2024, de: rubiales

Doménech, S. (2023, 1 de marzo). Joaquín Manso, director de El Mundo: "Una de las mejores recetas para crecer en influencia es el modelo premium". *Dircomfidencial*. Recuperado el 28 de mayo de 2024, de: elmundo

EFE (2023, 29 de mayo). Sánchez convoca elecciones anticipadas el 23 de julio tras la derrota del PSOE el 28M. *Agencia EFE.* Recuperado el 26 de mayo de 2024, de: elecciones

El Español (2016, 23 de octubre). Las 30 obsesiones de El Español. *El Español.* Recuperado el 28 de mayo de 2024, de: valores

El País (1977, marzo). *Estatuto de Redacción de El País.* Recuperado el 28 de mayo de 2024, de: estatutos

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communitation, 43(4)*, 51-58.

Esteve, M. (2023, 13 de septiembre). El catalán podrá hablarse a partir del próximo martes en el Congreso. *Diario ARA*. Recuperado el 28 de mayo de 2024, de: cooficiales

Herrero-Beaumont, E. (2024, 8 de febrero). Nacho Cardero: "Cada vez resulta más complicado ejercer de periodista, cada vez resulta más necesario". *Ethic*. Recuperado el 28 de mayo de 2024, de: editorial

Huesca, K. (2023, 28 de octubre). El PSOE abre a la militancia la decisión sobre la investidura tras fijar Sánchez la amnistía. *Agencia EFE.* Recuperado el 28 de mayo de 2024, de: amnistía

IBM (s.f.). ¿Qué es la minería de texto?. *IBM.* Recuperado el 28 de mayo de 2024, de: token

Ionos (2022, 17 de agosto). ¿Qué es una paywall? Todo sobre las barreras digitales. *Ionos*. Recuperado el 28 de mayo de 2024, de: paywall

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 89-106. Routledge.

Lardiez, A. (2023, 3 de octubre). El rey designa a Sánchez candidato a la investidura. *Demócrata.* Recuperado el 28 de mayo de 2024, de: investidura

León, S. (2023, 25 de mayo). La traca final del PSOE ante el 28-M: compra de votos, empadronamientos masivos, una pelea y un secuestro. *Libertad Digital.* Recuperado el 26 de mayo de 2024, de: escándalos

Lippman, W. (2003). *La opinión pública*. Madrid, Langre.

Marchante, D. (2023, 11 de septiembre). Cuándo tendrá lugar la investidura de Feijóo: estas son las fechas. *El Debate*. Recuperado el 28 de mayo de 2024, de: Feijóo

McCombs, M. (2006). *Estableciendo la agenda. El impacto de los medios en la opinión pública y en el conocimiento*. Barcelona, Paidós.

McCombs, M., y Shaw, D. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly, 36(2),* 176-187.

Menéndez, M. (2023a, 11 de mayo). Las claves de las elecciones del 28M: ¿Qué se vota, qué hay en juego y cómo llegan los partidos? *Noticias RTVE.* Recuperado el 26 de mayo de 2024, de: 28M

Menéndez, M. (2023b, 29 de mayo). El PP gana las elecciones del 28M y el PSOE pierde casi todo su poder territorial y sus principales bastiones. *Noticias RTVE.* Recuperado el 26 de mayo de 2024, de: resultados

Monrosi, J. E. (2023, 21 de julio). De la depresión a la remontada: la campaña ciclotímica del PSOE en busca de un vuelco el 23J. *elDiario.es.* Recuperado el 27 de mayo de 2024, de: remontada

Pinto Gurdiel, L., Morales Mediano, J. y Cifuentes Quintero, J. A. (2021). *A comparison study between coherence and perplexity for determining the number of topics in practitioners interviews analysis.* Generando y transfiriendo conocimiento en un entorno donde la única certeza es la incertidumbre. IV Congreso Iberoamericano de Jóvenes Investigadores en Ciencias Económicas y Dirección de Empresas, Madrid, Comunidad de Madrid, España.

Provost, F., y Fawcett, T. (2013). Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly.

Raintech (s.f.). Document-Term Matrix. En *Glossary*. Recuperado el 29 de mayo de 2024, de: DTM

Real Academia Española (s.f.). Oficialismo. En *Diccionario de la lengua española*. Recuperado el 25 de mayo de 2024, de: oficialismo

Real Academia Española (s.f.). Proceso electoral. En *Diccionario panhispánico del español jurídico*. Recuperado el 25 de mayo de 2024, de: Proceso electoral

Sabés Turmo, F. (2023). "El País", ¿el periódico global en español? Análisis de los últimos cambios en su formato. *Anagramas, 6(12),* 15-30.

Sánchez, J. A. (2020, 22 de junio). 20 años construyendo un sueño. *El Confidencial*. Recuperado el 28 de mayo de 2024, de: fundación

Santana, A. (2023, 10 de junio). Podemos se une a Sumar, la coalición de la izquierda española para las elecciones de julio. *El Independiente*. Recuperado el 26 de mayo de 2024, de: candidatura unitaria

Sartori, G. (1999). *Partidos y sistema de partidos*. Madrid, Alianza Editorial.

Sartori, G. (2005). *Elementos de teoría política*. Madrid, Alianza Editorial.

Shaw, D. y Martin, S. (1992). The Function of Mass Media Agenda Setting. *Journalism Quarterly, 69(4),* 902-920.

Soto, R. y García, Y. (2023, 24 de septiembre). Claves del acto del PP en protesta por una posible ley de amnistía. *Newtral*. Recuperado el 28 de mayo de 2024, de: Felipe II

Tong, Z. y Zhang, H. (2016). A document exploring system on LDA Topic model for Wikipedia articles. *The International Journal of Multimedia & Its Applications, 8(3/4),* 1-13.

Triviño, M. (2023, 13 de julio). Los sindicatos denuncian la situación límite en Correos para poder votar. *El Mundo*. Recuperado el 27 de mayo de 2024, de: correos

Valero, E. G. (2023, 17 de agosto). Constitución de las Cortes Generales, en directo: últimas noticias de la elección de la Mesa del Congreso. *La Razón*. Recuperado el 28 de mayo de 2024, de: Congreso

Van der Eijk, C. (1987). Testing theories of electoral cycles. The case of the Netherlands. *European Journal of Political Research, 15,* 253-270.

Vega, I. (2023, 3 de agosto). España batió el récord europeo de encuestas publicadas en vísperas de unas elecciones: 105 en las dos semanas previas al 23-J. *El Confidencial Digital.* Recuperado el 27 de mayo de 2024, de: encuestas

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Vijay Gaikwad, S., Chaugule, A. y Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications, 85(17),* 42-45.

Zhu, M., Zhang, X. y Wang, H. (2016). A LDA Based Model for Topic Evolution: Evidence from Information Science Journal. *Advances in Computer Science Research, 58*, 49-54.

CÓDIGO EMPLEADO

TFG Luis

Parte I: Importaciones previas

Importación de todos los paquetes requeridos para la realización del trabajo

```
!pip install spacy
!python -m spacy download en_core_web_lg
!pip install gensim
!pip install pyLDAvis
!pip install kaleido
```

```
import pandas as pd
from pathlib import Path
import plotly.express as px
import matplotlib.pyplot as plt
import numpy as np
from pathlib import Path
import spacy
import gensim
from gensim.models.phrases import Phraser
import pyLDAvis
import pyLDAvis.gensim_models
from wordcloud import WordCloud
import kaleido
import collections
from sklearn.feature_extraction.text import TfidfVectorizer
import seaborn as sns
import re
import plotly.graph_objects as go
from plotly.subplots import make_subplots
```

Parte II: Instalación del modelo y configuración de pasos previos

Configuración de los parámetros requeridos

```
SPACY_MODEL = "en_core_web_lg"
```

```
# ADJ: adjective
# ADP: adposition
# ADV: adverb
# AUX: auxiliary
# CONJ: conjunction
# CCONJ: coordinating conjunction
# DET: determiner
```

```
# INTJ: interjection
# NOUN: noun
# NUM: numeral
# PART: particle
# PRON: pronoun
# PROPN: proper noun
# PUNCT: punctuation
# SCONJ: subordinating conjunction
# SYM: symbol
# VERB: verb
# X: other
# SPACE
ALLOWED_POS_TAGS = ['ADJ', 'ADV', 'NOUN', 'PROPN', 'VERB']
```

Uso o no de bigramas. En nuestro caso, emplearemos un mínimo de diez bigramas para considerarlo relevante. Los vamos a tener en cuenta por la asociación propia de palabras que se hace en el ámbito de la comunicación.

```
USE_BIGRAMS = True # Usamos bigramas
BIGRAM_MIN_COUNT = 10 # Mínimo diez veces
BIGRAM_THRESHOLD = 5
```

Igual que por la naturaleza de los bigramas la exigencia es de diez, para tener términos relevantes duplicamos la exigencia a veinte y limitamos el número a tres mil quinientos, que es un número alto y que no debería eliminar muchos términos. Esto se deriva de la intencionalidad de tratar de preservar el máximo número de documentos.

```
NO_BELOW = 5 # No menos de cinco
NO_ABOVE = 0.7
KEEP_N = 3500 # Conservar máximo tres mil quinientas palabras
```

Aquí vamos a ir actualizando el número de topics en cada uno de los modelos analizados.

```
# Óptimo número de topics
DO_COHERENCE = False
COHERENCE_NUM_TOPICS = list(range(2, 15, 1)) + list(range(15, 30,
5)) + list(range(30, 60, 10)) + [100]
COHERENCE_RUNS = 6

# Número de topics para el análisis LDA en cada modelo
N_TOPICS = 7
LDA_PASSES = 35
```

Importamos tanto el documento generado de stopwords (que es una mezcla de stopwords clásicas extraídas de internet) como el excel trabajado con la estructura propia de los modelos analizados (fecha, periódico, título, subtítulo y cuerpo).

```
CORPUS_IN = "/content/CORPUS_DEFINITIVO.xlsx"
STOPWORDS_PATH = "/content/spanish.txt"
```

Aquí generamos las distintas rutas de salida.

```
OUTPUT_PATH = Path(f'/OUTPUT')
OUTPUT_PATH.mkdir(exist_ok=True, parents=True)

RUN_PATH = OUTPUT_PATH / f'{N_TOPICS:02d}/'
RUN_PATH.mkdir(exist_ok=True, parents=True)

CORPUS_OUT = RUN_PATH / 'corpus_results.xlsx'
WORDCOUNT_HISTOGRAM_PATH = OUTPUT_PATH /
'all_word_count_histogram.svg'
COHERENCE_PATH = OUTPUT_PATH / 'coherence.svg'
PERPLEXITY_PATH = OUTPUT_PATH / 'perplexity.svg'
NUM_TOPICS_PATH = OUTPUT_PATH / 'num_topics.xlsx'
```

Parte III: Preparación del modelo NLP

```
# Cargamos el modelo spacy
NLP = spacy.load(SPACY_MODEL)
print('Cargado: spaCy model')
```

```
# Leemos los stopwords
with open(STOPWORDS_PATH, 'r') as file:
    CUSTOM_STOPWORDS = [line.strip() for line in file.readlines()]

# Añadimos las stopwords a las stopwords del modelo
for word in CUSTOM_STOPWORDS:
    NLP.vocab[word].is_stop = True

print('Added: custom stop words to spacy\n', CUSTOM_STOPWORDS)
```

Parte IV: Lectura del *corpus*

Lectura del corpus. Hemos creado una columna con TODO el texto para poder analizar los topics de modo general.

```
corpus = pd.read_excel(CORPUS_IN)
corpus['All'] = corpus['Title'] + ' ' + corpus['Subtitle'] + ' ' +
corpus['Body'] # Creación de la columna ALL
corpus['Count'] = corpus['All'] # Creación de columna para gráficos
corpus
```

Vemos cuales son las palabras más repetidas.

```
palabras = []
```

```
for doc in corpus['All']:
  palabras.extend(doc.split())

palabras_contadas = collections.Counter(palabras)
print(palabras_contadas.most_common(50)) # Cálculo de las palabras
más frecuentes
```

```
corpus['Count'] = corpus['Count'].apply(lambda x: len(x.split()))
media_palabras =
corpus.groupby('Newspaper')['Count'].mean().reset_index()
media_palabras # Determinación media palabras por periódico a
través de una tabla
```

```
media_palabras['Count']=media_palabras['Count'].round(2) #
Redondeamos
fig = px.bar(media_palabras, x='Newspaper', y='Count', title='Media
palabras por periódico',
             labels={'Newspaper': 'Newspaper', 'Count': 'Count'},
             text='Count') # Mostramos la media de palabras

fig.update_layout(
    bargap=0.2,
    template='plotly_white',
    font=dict(
        size=14,
        color="black"
    ),
    title_font=dict(
        size=18,
        color='black',
        family="Arial"
    ),
    xaxis=dict(
        tickfont=dict(
            size=12,
            color='black',
            family="Arial"
        ),
    ),
    yaxis=dict(
        tickfont=dict(
            size=12,
            color='black',
            family="Arial"
        ),
    )
)
```

```
fig.show()
```

Parte V: Histogramas

Aquí realizamos el recuento de las palabras a través de histogramas. Lo hacemos para aquellas columnas en las que existen palabras que pueden recontarse, descartando por ello los periódicos (pues se limitaría a contar por un lado los periódicos con dos palabras y por otro el ABC) y las fechas (pues trataría de contar palabras donde sólo hay fechas).

```
TEXT_COLUMNS = ["Title", "Subtitle", "Body", "All"] # Lista para
histogramas
```

```
# Contamos las palabras
def _count_alphabetic_words(text, nlp):
    doc = nlp(text)
    # Contamos tokens
    return sum(token.is_alpha for token in doc)

# Realizamos un bucle en cada columna de TEXT_COLUMNS
for column in TEXT_COLUMNS:
    # Aplicamos la función a cada uno de los registros de cada
columna
    corpus[f'{column}_word_count'] =
corpus[column].apply(_count_alphabetic_words, args=(NLP,))

    # Graficamos el histograma
    fig = px.histogram(corpus, x=f'{column}_word_count',
title=f'Recuento de palabras en la columna {column}',
                       labels={f'{column}_word_count': 'Word
Count'}, nbins=20)  # Optamos por 20 barras

    fig.update_layout(bargap=0.2)

    fig.write_html(f'{column}_word_count_histogram.html')

    fig.show()  # Mostramos los histogramas
```

Parte VI: Preprocesamiento del texto

Eliminamos las stopwords (palabras que se repiten mucho y no son relevantes para nuestro análisis). Luego, también nos aseguramos de que todas los tokens se correspondan con alguna de las Part of Speech anteriormente indicadas.

También lematizamos, es decir, reducimos a un lema común. Estos resultados finalmente son guardados en una nueva columna con la estructura: columna_clean

```
def _pipeline_spacy(text, nlp, allowed_pos_ags = ALLOWED_POS_TAGS,
stopwords = CUSTOM_STOPWORDS):
```

```
    spacy_doc = nlp(text)

    lemmas = [token.lemma_.lower() for token in spacy_doc
              if str(token).lower() not in stopwords
              and token.is_alpha
              and token.pos_ in allowed_pos_ags
              and str(token.lemma_).lower() not in stopwords
              and not token.is_punct
              and not token.is_currency
              and not token.is_digit
              and not token.is_oov
              and not token.is_space
              and not token.is_stop
             ]

    return lemmas

for column in TEXT_COLUMNS:

    tokenized_texts = corpus[column].apply(_pipeline_spacy,
args=(NLP,))

    if USE_BIGRAMS:
        print(f'Using bigram model for {column}')

        bigram_model = gensim.models.Phrases(tokenized_texts,
min_count=BIGRAM_MIN_COUNT, threshold=BIGRAM_THRESHOLD)
        bigram_phraser = Phraser(bigram_model)


        texts_with_bigrams = [bigram_phraser[doc] for doc in
tokenized_texts]
        corpus[f'{column}_clean'] = texts_with_bigrams
    else:
        print(f'Using tokens for {column}')
        corpus[f'{column}_clean'] = tokenized_texts
```

```
corpus['All_clean'] # Mostramos como queda la columna a analizar
```

Calculamos la frecuencia absoluta y la matriz TF-IDF.

```
# Frecuencia absoluta:

datos = {'All_clean': corpus['All_clean']}

datos_frecuencia = pd.DataFrame(datos)
```

```
datos_frecuencia['document_str'] =
datos_frecuencia['All_clean'].apply(lambda x: ' '.join(x))

documentos = datos_frecuencia['document_str'].tolist()

palabras = [word for sublist in datos_frecuencia['All_clean'] for
word in sublist]

term_frequency = collections.Counter(palabras)

tf = pd.DataFrame(term_frequency.items(), columns=['Term',
'Frequency']).sort_values(by='Frequency', ascending=False)

print(tf)

# Vectorizamos para la matriz TF-IDF:

vectorizer = TfidfVectorizer()

matriz_tfidf = vectorizer.fit_transform(documentos)

palabras_tfidf = vectorizer.get_feature_names_out()

tfidf = pd.DataFrame(matriz_tfidf.toarray(),
columns=palabras_tfidf)

print(tfidf)

term_sum = np.sum(tfidf)
```

Seleccionamos las diez palabras más repetidas o con mayor peso para representar gráficamente.

```
top_10_terms = term_sum.nlargest(10)
top_10_terms # Términos más frecuentes (tfidf)
top_10_terms_dict = top_10_terms.to_dict()
top_10_terms_abs = tf.head(10) # Términos más frrecuentes
(absoluto)
```

```
# Representación gráfica de ambos

fig_freq = px.bar(top_10_terms_abs, x='Term', y='Frequency',
title='Términos con mayor frecuencia absoluta',
                  labels={'Término': 'Término', 'Frecuencia':
'Frecuencia'},
                  text='Frequency')
```

```
fig_freq.update_layout(
    bargap=0.2,
    template='plotly_white',
    font=dict(size=14, color="black"),
    title_font=dict(size=18, color='black', family="Arial"),
    xaxis=dict(tickfont=dict(size=12, color='black',
family="Arial")),
    yaxis=dict(tickfont=dict(size=12, color='black',
family="Arial"))
)

fig_freq.update_traces(texttemplate='%{text:.0f}',
textposition='outside')

fig_freq.show()
```

```
fig = go.Figure()

fig.add_trace(go.Bar(
    x=list(top_10_terms_dict.keys()),
    y=list(top_10_terms_dict.values()),
    text=[f'{value:.2f}' for value in top_10_terms_dict.values()],
    textposition='outside',
    name='TF-IDF Score'))

fig.update_layout(
    title='Términos con mayor puntuación TF-IDF',
    xaxis_title='Término',
    yaxis_title='Puntuación',
    bargap=0.2,
    template='plotly_white',
    font=dict(size=14, color="black"),
    title_font=dict(size=18, color='black', family="Arial"),
    xaxis=dict(tickfont=dict(size=12, color='black',
family="Arial")),
    yaxis=dict(tickfont=dict(size=12, color='black',
family="Arial"))
)

fig.show()
```

```
# Wordcloud en términos absolutos
diccionario_frecuencia = dict(zip(tf['Term'], tf['Frequency']))

wordcloud_frecabs = WordCloud(width=800, height=400,
background_color='white').generate_from_frequencies(diccionario_fre
cuencia)
```

```
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud_frecabs, interpolation='bilinear')
plt.axis('off')
plt.title('Wordcloud palabras')
plt.show()
```

```
# Wordcloud resultante de las puntuaciones de la matriz TF-IDF
diccionario_tfidf = term_sum.to_dict()

wordcloud_tfidf = WordCloud(width=800, height=400,
background_color='white').generate_from_frequencies(diccionario_tfi
df)

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud_tfidf, interpolation='bilinear')
plt.axis('off')
plt.title('Wordcloud de puntuaciones TF-IDF')
plt.show()
```

Parte VII: Creación de la Document-Term Matrix (DTM)

Creamos la Matriz de Términos para la columna en la que se encuentra todo el texto de forma agrupada, para realizar un análisis del corpus. Para ello, descartamos palabras que salgan en menos de cinco documentos o en más del setenta por cien.

```
# Creación de diccionario para la columna que contiene el texto
dictionary = gensim.corpora.Dictionary(corpus['All_clean'])

# Limitaciones
dictionary.filter_extremes(no_below=NO_BELOW, no_above=NO_ABOVE,
keep_n=KEEP_N)

# DTM
doc_term_matrix = [dictionary.doc2bow(doc) for doc in
corpus['All_clean']]
```

```
# Recuento de documentos (verificación de que suma 945)
print(f'Number of documents: {dictionary.num_docs}')
```

```
# Recuento de términos
terms = [dictionary[id] for id in dictionary]
terms = sorted(terms)
print(f'Número de términos: {len(terms)} --- Mostrando los 100
primeros\n', terms[:100])
```

```
# Recuento de bigramas
bigrams = [term for term in terms if '_' in term]
```

```
print(f'Número de bigramas: {len(bigrams)} --- Mostrando los 100
primeros\n', bigrams[:100])
```

Parte VIII: Determinación del número óptimo de *topics*

Aquí consieramos el overfitting.

```
def _compute_coherence_values(dictionary, corpus, texts,
num_topics_list, num_runs):

    average_coherence_values = []
    average_perplexity_values = []

    for num_topics in num_topics_list:
        # Para cada número de topics

        coherence_values = []
        perplexity_values = []

        for _ in range(num_runs):
            # Lo lanzamos varias veces
            lda_model = gensim.models.LdaModel(
                corpus=corpus,
                num_topics=num_topics,
                alpha='auto',
                eta='auto',
                id2word=dictionary,
                passes=LDA_PASSES)
            coherence_model = gensim.models.CoherenceModel(
                model=lda_model,
                texts=texts,
                dictionary=dictionary,
                coherence='c_v')


coherence_values.append(coherence_model.get_coherence())

perplexity_values.append(lda_model.log_perplexity(corpus))

        average_coherence = sum(coherence_values) / num_runs
        average_perplexity = sum(perplexity_values) / num_runs
        print(f'{num_topics=}
{average_coherence=:.6f}\t{min(coherence_values):.6f} ..
{max(coherence_values):.6f}')
        print(f'{num_topics=}
{average_perplexity=:.6f}\t{min(perplexity_values):.6f} ..
{max(perplexity_values):.6f}')

        average_coherence_values.append(average_coherence)
```

```
        average_perplexity_values.append(average_perplexity)

    return average_coherence_values, average_perplexity_values

if True:
    coherence_values, perplexity_values =
_compute_coherence_values(
        dictionary=dictionary,
        corpus=doc_term_matrix,
        texts=corpus['All_clean'],
        num_topics_list=COHERENCE_NUM_TOPICS,
        num_runs=COHERENCE_RUNS)

    fig = px.line(x=COHERENCE_NUM_TOPICS, y=coherence_values,
title='coherence (the higher the better)')
    fig.write_image(COHERENCE_PATH)
    fig.show()

    fig = px.line(x=COHERENCE_NUM_TOPICS, y=perplexity_values,
title='perplexity (the lower the better)')
    fig.write_image(PERPLEXITY_PATH)
    fig.show()
```

```
# Lo guardamos en excel
num_topics_df = pd.DataFrame({
    'n_topics': COHERENCE_NUM_TOPICS,
    'coherence': coherence_values,
    'perplexity':perplexity_values})

num_topics_df.to_excel(NUM_TOPICS_PATH)
```

Parte IX: *Topic* analysis

```
# Creación del modelo LDA
lda_model = gensim.models.LdaModel(
    corpus = doc_term_matrix,
    num_topics = N_TOPICS, # Topics deseados
    id2word = dictionary,
    passes = LDA_PASSES, # Ojear
    random_state = 42,  # Reproductividad
    alpha = 'auto',
    eta = 'auto')

# Generamos la visualización HTML
vis = pyLDAvis.gensim_models.prepare(lda_model, doc_term_matrix,
dictionary, sort_topics=False)
pyLDAvis.save_html(vis, str(RUN_PATH /
'interactive_LDA_plot.html'))
```

```
# Mostramos los topics y palabras asociadas
topics = lda_model.print_topics(num_words=5)
print(topics)
for topic_id, description in topics:
    print(f'topic_{1+topic_id:02d} "{description}"')

pyLDAvis.enable_notebook()
pyLDAvis.display(vis)
```

Representamos los cinco términos más frecuentes de cada *topic*.

```
topics = lda_model.print_topics(num_words=5)
print(topics)
top_terms = {}
for topic_id, description in topics:
    terms = re.findall(r'([\d\.]+)\*"?(\w+)"?', description)
    term_weights = [(float(weight), term) for weight, term in
terms]
    top_terms[topic_id] = sorted(term_weights, reverse=False)

plot_data = []
for topic_id, terms in top_terms.items():
    for weight, term in terms[:5]: # Limitamos a cinco términos
        plot_data.append((topic_id, term, weight))

datos_topic = pd.DataFrame(plot_data, columns=['Topic', 'Term',
'Weight']) # Creamos un dataframe para representar

# Visualización
for topic_id in datos_topic['Topic'].unique():
    visualizaciones = datos_topic[datos_topic['Topic'] == topic_id]
    fig = px.bar(visualizaciones, x='Weight', y='Term',
orientation='h',
                 title=f'5 Términos con más peso para el Topic
{topic_id+1}',
                 labels={'Weight': 'Peso', 'Term': 'Término'})
    fig.update_layout(bargap=0.2, template='plotly_white')
    fig.update_traces(text=datos_topic['Weight'].round(3),
textposition='outside')
    fig.show()
```

Parte X: Datos para cada documento

```
# Obtenemos la distribución de los topics en cada documento
doc_topics = [lda_model.get_document_topics(doc,
minimum_probability=0.) for doc in doc_term_matrix]
num_topics = lda_model.num_topics
```

```
doc_topic_dist = pd.DataFrame()

for doc_id, topics_per_doc in enumerate(doc_topics):
    for topic_id, topic_prob in topics_per_doc:
        doc_topic_dist.loc[doc_id, f'topic_{1+topic_id:02d}'] =
topic_prob

# Tratamos los valores faltantes
doc_topic_dist = doc_topic_dist.fillna(0)

# Copiamos el índice apropiado perdido en el procedimiento
doc_topic_dist.index = corpus.index

# Ordenamos las columnas
doc_topic_dist = doc_topic_dist.sort_index(axis=1)

# Las unimos con el corpus
corpus = corpus.join(doc_topic_dist)
```

```
# Creamos un corpus agrupado para representar los topics por
separado y agrupamos por fecha para ver la evolución temporal

columnas_media = ['topic_01', 'topic_02', 'topic_03', 'topic_04',
'topic_05', 'topic_06', 'topic_07']


corpus_agrupado =
corpus.groupby('Date')[columnas_media].mean().reset_index()
```

```
corpus_agrupado # Vemos el resultado
```

```
# Lo representamos de forma individualizada a través de un bucle y
de forma conjunta para mayor impacto visual
corpus_representado = pd.melt(corpus_agrupado, id_vars='Date',
var_name='topic', value_name='value')

fig = px.line(corpus_representado, x='Date', y='value',
color='topic', title='Representación temporal de los topics')
fig.update_layout(plot_bgcolor='rgba(0,0,0,0)',
paper_bgcolor='rgba(0,0,0,0)')
fig.show()
```

```
for topic in columnas_media:
    fig = px.line(corpus_agrupado, x='Date', y=topic,
title=f'Representación temporal del {topic}')
```

```
    fig.update_layout(plot_bgcolor='rgba(0,0,0,0)',
paper_bgcolor='rgba(0,0,0,0)')
    fig.show()
```

Parte XI: *Wordclouds* por *topic*

```
num_topics = lda_model.num_topics


for i in range(num_topics):
    # Creamos un wordcloud para cada topic

    # Términos y sus pesos para cada topic
    terms = lda_model.show_topic(i, topn=500)

    # Creamos un diccionario con las palabras y sus pesos para cda
wordcloud
    terms_dict = {term: weight for term, weight in terms}

    # Graficamos
    wc = WordCloud(width=1600,
                   height=900,
                   max_font_size=150,
                   max_words=500,

background_color='white').generate_from_frequencies(terms_dict)

    wc.to_file(RUN_PATH / f'wordcloud_{1+i:02d}.png')
```

Parte XII: Guardamos en excel la composición

```
NUM_WORDS = 20

with pd.ExcelWriter(RUN_PATH / 'lda_topics.xlsx') as writer:

    for i in range(num_topics):

        # Extramos las palabras y sus frecuencias para cada topic
        words_freqs = lda_model.show_topic(i, NUM_WORDS)
        df = pd.DataFrame(words_freqs, columns=['term',
'frequency'])

        df.to_excel(writer, sheet_name=f'Topic_{1+i}', index=False)
```

Parte XIII: Guardar estadísticas en excel

```
corpus.to_excel(CORPUS_OUT)
```

Parte XIV: Creamos un nuevo *corpus* para poder desglosar el análisis por medio

```
columnas_remanentes = ['Date', 'Newspaper'] + [f'topic_{i:02d}' for
i in range(1, 8)]

nuevo_corpus = corpus[columnas_remanentes]
nuevo_corpus # Nuevo corpus con las columnas de la fecha, medio y
topics con pesos
```

```
# Renombramos las columnas de los topics para mayor claridad
nuevo_corpus = nuevo_corpus.rename(columns={
    "topic_01": "Frente amplio de izquierdas",
    "topic_02": "Negociaciones sorprendentes",
    "topic_03": "El último gol de Rubiales",
    "topic_04": "La Mesa del Congreso y sus expresiones
lingüísticas",
    "topic_05": "El conflicto árabe-israelí",
    "topic_06": "El último acto de las negociaciones",
    "topic_07": "Crónica de una elección perdida"
})

nuevo_corpus
```

```
# Representación de la evolución de los topics por medio

corpus_agrupamiento = nuevo_corpus.groupby('Newspaper')

for newspaper, group in corpus_agrupamiento:

    fig = make_subplots(rows=7, cols=1, shared_xaxes=True,
                        vertical_spacing=0.02,
                        subplot_titles=topics)

    for i, topic in enumerate(topics, start=1):
        fig.add_trace(go.Scatter(
            x=group['Date'],
            y=group[topic],
            mode='lines',
            name=topic
        ), row=i, col=1)

    fig.update_layout(
        height=2000,
        title_text=f'Topic Trends for {newspaper}',
        showlegend=False
    )
```

```
    fig.update_xaxes(showgrid=False)
    fig.update_yaxes(showgrid=False)
    fig.update_layout(plot_bgcolor='rgba(0,0,0,0)',
paper_bgcolor='rgba(0,0,0,0)')

    fig.show()
```

DECLARACIÓN DE USO IA GENERATIVA

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, **LUIS VILLANUEVA RIBES**, estudiante de **DERECHO Y ANÁLISIS DE NEGOCIOS / *BUSINESS ANALYTICS*** de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "***TEXT MINING* Y EL CICLO ELECTORAL ESPAÑOL DE 2023: UN ANÁLISIS DE LOS PRINCIPALES DIARIOS**", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Interpretador de código:** Para realizar análisis de datos preliminares.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 18/06/2024

Firma: Luis Villanueva Ribes