Vol. 4, Num. 2. **Article n.º 171**
DOI: https://doi.org/10.35951/v4i2.171

C&M Communication & Methods

Métodos de Procesado del Lenguaje Natural aplicados al estudio de las coberturas mediáticas

Natural Language Processing Methods Applied to the Study of Media Coverage

Mar Castillo-Campos. Universidad Loyola Andalucía (España)

Mar Castillo-Campos has a degree in Communication, and a master's degree in Research Methods. She is currently working at Universidad Loyola Andalucía as a research assistant and PhD student, integrating quantitative methodologies and data science in the field of journalism.
ORCID: orcid.org/0000-0003-1931-1493

David Becerra-Alonso. Universidad Loyola Andalucía (España)

David Becerra-Alonso obtained his PhD in the School of Computing at the University of the West of Scotland, where he worked on dynamical chaotic systems. David currently holds a position as a lecturer at Universidad Loyola Andalucía. His research interests include dynamical systems, emergent collective behavior, and machine learning techniques and heuristics.
ORCID: orcid.org/0000-0001-5174-7743

David Varona-Aramburu. Universidad Complutense Madrid (España)

David Varona-Aramburu has a PhD in Journalism and currently works at the Journalism and New Media Department at Universidad Complutense de Madrid. David has worked for more than 20 years in the media, and does research in Communication and Media.
ORCID: orcid.org/0000-0001-8972-0490

Artículo recibido: 08/10/2022 – Aceptado: 16/11/2022

Resumen:
El Procesamiento del Lenguaje Natural comprende distintas técnicas cuantitativas para el análisis de textos que presentan puntos de partida distintos a los habitualmente empleados en periodismo. Con un carácter eminentemente exploratorio y en base a la teoría fundamentada, la conjunción de técnicas aquí empleadas, TF, TF*IDF, word2vec y proyección de términos con UMAP permite detectar la vinculación entre términos en distintas fuentes documentales, así como su frecuencia de uso y la exposición a determinados conceptos, ideas y personajes. Esta metodología pretende ayudar a

vislumbrar nuevas líneas de exploración y estudio, y puede combinarse con otras técnicas de análisis del discurso más profundas. La flexibilidad del método permite además experimentar con distintos grupos de palabras para cualquier otra fuente documental.

Palabras clave:
Procesado del Lenguaje Natural; PLN; TFIDF; cobertura mediática

Abstract:
*Natural Language Processing comprises different quantitative techniques for the analysis of texts that present different starting points to those usually used in journalism. With an eminently exploratory character and based on grounded theory, the combination of techniques used here, TF, TF*IDF, word2vec and projection of terms with UMAP allow us to detect the link between terms in different documentary sources, as well as their frequency of use and exposure to certain concepts, ideas and characters. This methodology is intended to help to envision new lines of study, and can be combined with other more in-depth discourse analysis techniques. The flexibility of the method also allows experimentation with different word groups for any other documentary source.*

Keywords:
Natural Language Processing; NLP; TFIDF; media coverage

1. Introducción y estado de la cuestión

El estudio de la comunicación política está comúnmente vinculado a un análisis del discurso de los candidatos. En períodos de campaña electoral, además, se centra la atención en la posible persuasión que los políticos pudieran ejercer sobre la población a través de los medios de comunicación (McNair, 2017), pero esto deja otros muchos aspectos importantes fuera de análisis (tipo de información que se comunica, *agenda setting*, criterios de evaluación para cada candidato...) (Iyengar & Simon, 2000). Tradicionalmente, el estudio de la comunicación política se hacía mediante encuestas y paneles por el lado de la percepción, y con análisis del discurso para el contenido (Iyengar & Simon, 2000), aunque cada vez más se emplean técnicas cuantitativas y de inteligencia artificial que permiten automatizar parte del análisis y trabajar con muestras mayores.

En el caso de esta investigación, se emplean técnicas pertenecientes a lo que se conoce como Procesado del Lenguaje Natural (PLN). Por "lenguaje natural" se entiende aquel empleado en el día a día por las personas para comunicarse entre ellas, en cualquier idioma (Bird *et al.*, 2019, p. 9). Se diferencia de los lenguajes "artificiales" —como el matemático o el de programación, por ejemplo— en que no sigue unas reglas tan estrictas y evoluciona con el tiempo, el tipo de uso o la geografía, entre otros factores. Esto lo hace más complejo de analizar y estudiar. El procesamiento computacional que trata este lenguaje puede ser dificultoso, precisamente por la amplitud, diversidad y complejidad de lo que se estudia (Bird *et al.*, 2019). El PLN comprende desde técnicas como los conteos automáticos de palabras, por ejemplo, que ayudan a detectar temáticas

o estilos de escritura, hasta sistemas más sofisticados que, apoyados en inteligencia artificial, logran procesar esos textos en una suerte de "entendimiento" para dar respuesta a ese lenguaje (Sun *et al.*, 2017). Parte de la complejidad de este procesamiento reside en la riqueza del lenguaje (sinónimos, polisemias, anáforas...), que se acentúa en textos literarios, publicitarios, poéticos o periodísticos, entre otros. Los artículos científicos o expositivos tienden a emplear un lenguaje más sencillo, acotado y canónico, lo que hace que el PLN se aplique más a menudo y con más éxito a textos de este tipo.

Estas técnicas, aún en desarrollo, están permitiendo vislumbrar *insights* o pronósticos que pasan desapercibidos en el uso de otros métodos de investigación, puesto que no siempre se había puesto el foco intencional en estudiar determinados factores que sí han puesto de relieve métodos exploratorios, no confirmatorios.

En todo caso, se trata de un método cuyo uso no está extendido en el campo del periodismo. Los estudios que anteceden a este habitualmente comparan manualmente una muestra acotada de artículos, leyéndolos, analizándolos y extrayendo conclusiones sobre qué temas tratan esas piezas informativas, cómo presentan a cada actor o qué rol desempeñan (Casero-Ripollés *et al.*, 2016; Mancera-Rueda y Villar-Hernández, 2020; Sánchez Gutiérrez, 2016). También buscan cuáles reciben más o menos cobertura, qué sentimiento subyace de la escritura (Gao *et al.*, 2019; Li *et al.*, 2022; Miguel-Sáez-de-Urabain, *et al.*, 2017; Sánchez Gutiérrez y Nogales Bocio, 2018; Shapiro *et al.*, 2020)... Asimismo, son realmente numerosas las investigaciones que estudian cuantitativamente el lenguaje en la conversación en redes sociales (Doan *et al.*, 2019; Emadi y Rahgozar, 2020; Goularas y Kamis, 2019; Müller, 2020; Paniagua-Rojano *et al.*, 2020; Singh *et al.*, 2018...), si bien en esta ocasión se ha querido optar por textos más largos y complejos como los periodísticos. Otros investigadores que han empleado la vectorización y cuantitivización de los documentos han utilizado un método similar para clasificar las noticias: Orden-Cruz *et al.* (2019) aplican, por ejemplo, PLN sobre titulares y resúmenes de noticias para extraer el sentimiento implícito. Riedel *et al.* (2017) utilizan la misma metodología aquí expuesta (Term Frequency, Term Frequency and Inverse Document Frequency, y cálculo de similitud coseno) para clasificar noticias. Aquí se expondrá un método complementario, con distintos fines y potencial en el campo de la comunicación.

2. Análisis de la metodología

La mayoría de los estudios previos apuestan por una metodología cualitativa, lo que por un lado permite muestras más acotadas y tiende a realizar un análisis en profundidad (Marshall, 1996, p. 523). Las metodologías cuantitativas, por otro lado, parten mayoritariamente de una intención clasificatoria (Berven *et al.*, 2020; Edell, 2018; Jung y Lee, 2019; Zhou *et al.*, 2019). Esto supone que el investigador ha diseñado previas categorías, en las que luego se "encajan" los resultados, técnica muy útil para algunas investigaciones, pero que, en ocasiones, resulta limitante y sesgada, pues podría comprometer los datos a pertenecer a alguna de las categorías establecidas, dejando fuera otras alternativas no planteadas u otro tipo de agrupaciones. Su uso está especialmente extendido para la clasificación de documentos, de cuya totalidad de la muestra se ha seleccionado una parte que se ha etiquetado, lo cual requiere de la

intervención previa del investigador. El análisis exploratorio, en cambio, no crea categorías, sino que agrupa los datos y pone de manifiesto los puntos de unión y disidencia entre ellos, siendo labor del investigador analizar, a posteriori, a qué corresponden y si es pertinente establecer distintas categorías. Con base en la teoría fundamentada, estas técnicas son especialmente útiles para descubrir datos, compararlos y plantear preguntas a partir de ellos, que con otros métodos podrían no salir a la luz.

Partiendo de esa convicción, esta investigación se desarrolla en tres fases: una primera, de conteo de términos; una segunda, de estudio de relación entre binomios de conceptos mediante la aplicación de redes neuronales; y una última, de agrupación de términos y proyección de resultados.

A continuación, se exponen los diferentes métodos empleados:

2.1. Frecuencia de términos

Term Frequency (TF) o "frecuencia de un término" es un conteo automático de palabras en el texto, tanto absoluto como relativo respecto al número total de palabras en el mismo documento. Se denomina también *Local Term Weight*, pues es una medida de un texto individual, no comparado, de correlación positiva con su frecuencia (Tian y Tong, 2010). Si bien se emplea como punto de partida para muchas investigaciones (detección de términos significativos, etiquetado de noticias según relevancia de una palabra, etc.), también resulta útil para el análisis de las coberturas mediáticas para:

> extraer palabras clave de cada fuente documental o medio de comunicación;
>
> como punto de partida para detectar temáticas de interés, términos o conceptos frecuentes, para cada fuente o noticia;
>
> calcular el potencial número de veces que un lector habría estado expuesto a diferentes términos, ideas o candidatos, o qué importancia se le ha dado a determinado tema o personaje por cada fuente documental.

TF aporta un dato absoluto para cada palabra y por cada texto, pudiéndose establecer una comparativa, por ejemplo, entre piezas informativas sobre un mismo hecho en distintos medios de comunicación. Una elevada tasa de TF para una palabra en un texto podría aportar información sobre la temática del mismo, así como la comparativa entre tasas para distintas palabras permitiría establecer una jerarquía en función a la importancia que se les da en el texto.

Cuando se trata de términos relevantes, pero frecuentes en todos los textos a comparar, se emplean otras técnicas, como *Term Frequency times Inverse Document Frequency* (TF*IDF), que permite localizar la fuente que más importancia otorga a un término concreto. TF*IDF se formula como un factor entre la frecuencia relativa de una palabra o conjunto de palabras en un texto y la comparación de toda una selección de textos que contienen ese mismo término (Salton y Buckley, 1988). De este modo, puede resaltar los términos más utilizados en cada texto o la fuente que más veces está empleando determinados conceptos en comparación con un gran cuerpo de textos.

En TF*IDF, el valor cero para un término indica que no aparece en el texto a estudiar. Un hipotético valor 1 indica que dicho concepto aparece una cantidad relevante de veces en ese texto y no aparece en otros de la muestra. Los valores más próximos a la unidad indican que los términos estudiados son relevantes en ese texto y se mencionan poco en los demás con los que se compara. Algunos términos, aunque muy mencionados, son empleados por muchas de las distintas fuentes comparadas: en ese caso, el valor del concepto es cercano a cero.

Habitualmente, este método se emplea para detectar palabras clave en los textos (Qaiser y Ali, 2018; Kuncoro y Iswanto, 2015), para construir resúmenes de un cuerpo más largo (Christian *et al.*, 2016) o para la clasificación de textos basada en conceptos diferenciales (Xia y Chai, 2011; Wongso *et al.*, 2017), entre otros.

Ambas metodologías expuestas, TF y TF*IDF, pueden ser muy útiles como punto de partida, pues permiten un posterior cálculo más sofisticado entre términos, y revelan datos interesantes sobre las prioridades informativas de cada fuente.

2.2. Aplicación de redes neuronales

Para medir la vinculación de los conceptos obtenidos entre sí, se puede recurrir a técnicas de aprendizaje automático. En concreto, el modelo *word2vec* es una de las redes neuronales de convolución que permite el procesado del lenguaje natural. Hasta entonces, la conversión numérica de los términos en bolsas de palabras (BOW) tenía carencias importantes, pues se perdía el sentido del orden de las palabras en el texto y, por tanto, la relación que se establecía entre ellas. También se ignoraba la semántica de las mismas, no diferenciando entre sinónimos, verbos o conectores (Le y Mikolov, 2014). Así, word2vec ("de palabra a vector") realiza la conversión de cada término a un vector numérico de múltiples dimensiones, proceso por el que mantiene la relación entre las palabras y es capaz de trazar vínculos sintácticos y conceptuales entre ellas (Mikolov *et al.*, 2013). Se puede aplicar este modelo de red neuronal para detectar la probabilidad de que un término dado se relacione con otro. Dicha probabilidad considera la distancia o similitud entre los términos en una muestra y contempla, para cada documento, cómo de vinculados están dos conceptos.

2.3. Proyección de resultados

La comparativa de vectores de múltiples dimensiones obtenidos con word2vec se realiza mediante la operación con matrices de distancia, pero otras técnicas permiten la visualización de los datos de forma más sencilla y aclaratoria. Los reductores dimensionales precisamente logran transformar las múltiples dimensiones de los vectores con la mínima pérdida de información posible. Los vectores numéricos creados en fases previas se simplifican y proyectan en un gráfico de dos dimensiones, manteniendo la distancia entre los términos estudiados. De este modo, es posible observar cómo diferentes términos se agrupan bien porque cumplan funciones sintácticas parecidas, porque estén habitualmente relacionados en las oraciones o porque registren una mayor probabilidad de aparecer juntos en el texto.

El reductor dimensional *Uniform Manifold Approximation and Projection* (UMAP) ha mostrado con anterioridad mejores resultados en la proyección de datos y visualización de estructuras que otros reductores (McInnes *et al.*, 2018; Vermeulen *et al.*, 2021). Su uso exploratorio permite analizar cómo unas palabras se colocan en el espacio. Sería

opción del investigador considerar si la distribución obtenida corresponde a agrupaciones de términos o no. Una de las principales ventajas que presenta es que podría revelar asociaciones de conceptos que no se hubieran planteado estudiar al principio de la investigación y que abran posibles líneas de interés e investigación.

Ha de recalcarse que las distancias obtenidas indican una relación entre conceptos, pero no exigen correlación, proporcionalidad o reciprocidad. Si bien tanto la matriz como el mapa de términos permiten vislumbrar *insights* posiblemente desapercibidos mediante un análisis cualitativo o clasificatorio, los resultados no deben quedar exentos de una interpretación por investigadores profesionales del ámbito que se estudia.

La combinación de técnicas es útil en tanto que los términos más frecuentes (obtenidos con TF) y los diferenciales (TF*IDF) son seleccionados para estudiar la distancia entre ellos (word2vec) y la distribución (UMAP).

3. Ejemplo de aplicación

La metodología propuesta en este estudio se ha empleado para analizar la cobertura en medios de comunicación de las elecciones a la Asamblea de Madrid, celebradas el 4 de mayo de 2021 (4M). Estos comicios fueron de especial interés por el contexto de pandemia en el que se celebraron y porque, pese a ser de alcance regional, alcanzaron cobertura nacional y fueron seguidos de cerca por los medios más importantes en España. La gran cantidad de noticias generada desde el anuncio de las elecciones, y particularmente durante la campaña electoral, desvió el foco informativo de la pandemia para centrarse en las candidaturas a la Asamblea.

Se propone la metodología anteriormente descrita para investigar la cobertura mediática en el contexto electoral: averiguar qué partidos políticos copan la agenda de los medios y qué actores se ven reforzados; determinar con qué propuestas electorales se relacionan y evaluar qué importancia se les da a estas respecto a los sucesos en los que se ven involucrados los candidatos. La muestra de noticias (n=450), que recoge todas las noticias vinculadas a las elecciones del 4M durante las dos semanas que dura la campaña electoral aparecidas en tres de los medios más relevantes en España, justifica una metodología que automatice y agilice el análisis de los textos, y que permita un análisis exploratorio —no clasificatorio— de las piezas informativas.

Con la aplicación de TF se pone de relieve la gran visibilidad de algunos actores en titulares, el abuso de adjetivos o la gran cantidad de palabras de connotación violenta, entre otros. Así, por ejemplo, dada una población de todas las noticias relativas a la campaña electoral 4M para el medio de comunicación eldiario.es, solo en una semana la candidata Isabel Díaz Ayuso apareció 26 veces en titulares, sucedida por el partido político Vox —22 veces— y el candidato Pablo Iglesias, en 11 ocasiones. Sin embargo, para el periódico El País, Pablo Iglesias es el personaje más nombrado en titulares esa misma semana, 14 veces (Figura 1). Seis de los partidos concurrentes y sus cabeza de lista tuvieron representación mediática, pero las agrupaciones restantes no fueron siquiera mencionadas.

Figura 1

Número de apariciones en titulares de los tres candidatos o agrupaciones más nombradas, en la primera semana de campaña

	Ayuso	Iglesias	Vox
ABC	8	7	4
eldiario.es	26	11	22
El País	12	14	14

En relación a las palabras con connotación ideológica o emocional, en el medio ABC se contabilizan en titulares las palabras "antiokupas", "holocausto", "nazismo" y "criminal" al menos una vez durante una semana de campaña electoral. En eldiario.es, "amenazas", "fascismo" y "ultraderecha" aparecen hasta tres veces en titulares en solo siete días (Figura 2).

Figura 2

Número de apariciones en titulares por término en la primera semana de campaña

	amenazas	antiokupas	criminal	fascismo	holocausto	odio	nazismo	ultradercha
ABC	x	x	x		x		x	
eldiario.es	x			x				x
El País	x					x		

En este estudio se ha utilizado TF*IDF para detectar qué medios han otorgado más importancia a ciertos términos, ya sea a los candidatos, a determinadas medidas políticas o a acontecimientos concretos que han sucedido durante la campaña. Así, ha sido visible cómo para algunos grupos de medios, medidas más concretas (económicas, sociales, educativas, fiscales...) cobran mucha más relevancia que para otros, y que ciertos candidatos de partidos minoritarios solo reciben cobertura en algunos medios. Se descubre, además, cómo términos de connotación peyorativa, negativa e incluso violenta son muy frecuentes en algunos diarios y aparecen de forma residual en otros (Figura 3).

Figura 3

*Tasa de TF*IDF para los tres medios de comunicación*

	ayusadas	ayusismo	ayusista
ABC	0	0	0
ELD	0,00249102	0	0
ELP	0	0,00329014	0,00329014

En el caso de este ejemplo, se usa de forma despectiva "ayusada", "ayusismo" o "ayusista", referido como algo propio, proclive o relacionado con la candidata Isabel

Díaz Ayuso. Para ABC, la tasa TF*IDF es cero: no se ha empleado ninguna de estas palabras en las noticias publicadas por este medio. El País y eldiario.es sí las han empleado en sus piezas informativas, aunque no con especial asiduidad –por ello la tasa TF*IDF es baja. Es un ejemplo de las palabras que resultan espontáneamente de la aplicación de TF*IDF: si bien no se habían planteado estos términos al principio de la investigación, una tasa cero para un medio y distinta de cero para otros apunta al investigador una diferencia que revisar.

En una fase posterior, el uso de word2vec ha permitido comprobar si un medio de comunicación relaciona un partido con ciertas propuestas políticas, problemáticas o conceptos. Tanto en El País como en eldiario.es, Edmundo Bal del partido Ciudadanos es el que más se relaciona con las propuestas de fiscalidad. Las políticas de empleo están para ambos medios más vinculadas a Isabel Díaz Ayuso, candidata por el Partido Popular (PP). Para el periódico eldiario.es, entre los términos que con mayor probabilidad aparecerán junto a la cabeza de lista del PP se hallan "victoria" o "reelección", y su contrario, Pablo Iglesias, estará habitualmente próximo a "muerte" o "abandonado" (Figura 4).

Figura 4

Binomio de términos con máxima probabilidad de aparecer juntos, para cada medio

	Ayuso	Iglesias	Vox
ABC	presidenta (0.983)	violencia (0.938)	afear (0.998)
eldiario.es	reelección (0.883)	abandonado (8.55)	radical (0.994)
El País	justificar (0.91)	antidisturbios (0.878)	negar (0.874)

Entre los términos más cercanos a Vox o a su cabeza de lista Rocío Monasterio se hallan "afear", "radical" o "duda", para ABC, eldiario.es y El País, respectivamente (Figura 5).

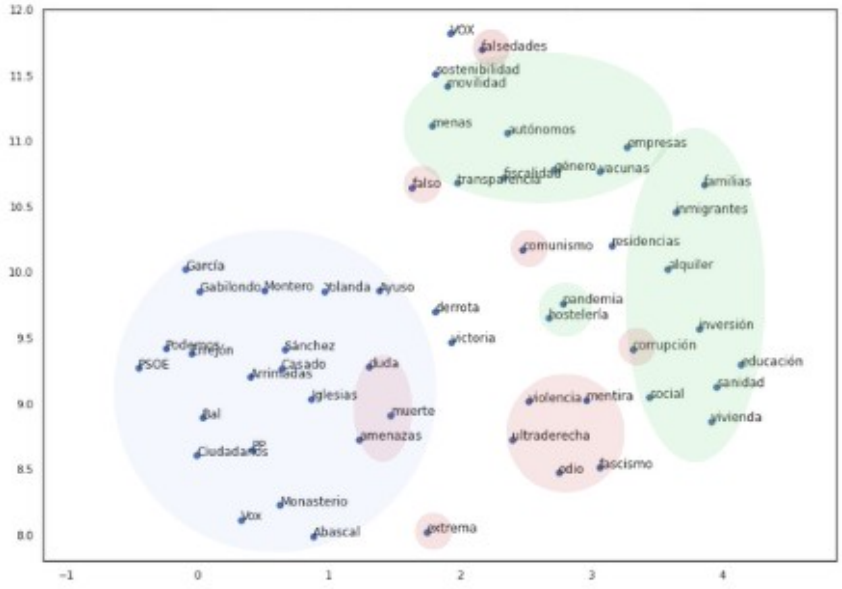
Figura 5

Distancia entre el partido político Vox y la candidata Rocío Monasterio respecto a otros términos, para cada medio

En una tercera fase se emplea el reductor dimensional UMAP, con el que se observa cómo los candidatos a unas elecciones se posicionan como términos semejantes, cumpliendo unas funciones sintácticas similares, empleados en los mismos contextos y con cierta vinculación unos con otros (Figura 6). Por otro lado, las medidas propuestas también se sitúan en cercanía ("autónomos", "empresas", "fiscalidad"; "sostenibilidad", "movilidad"; "pandemia", "hostelería", "residencias" ...). Otros términos, más cargados de sentimiento, se asocian habitualmente a los mismos actores o medidas políticas (Iglesias, cerca de "muerte" y "amenazas", y los relacionados con Vox, Monasterio y Abascal, con "extrema" y "falsedades").

Figura 6

Proyección de términos con UMAP del medio eldiario.es



Es interesante explorar esas vinculaciones, que con otras técnicas habría sido complejo extraer. La relación de cercanía o lejanía entre términos conviene ser estudiada y abre otras líneas de investigación: por qué para determinado medio de comunicación la corrupción parece más vinculada a la inversión, la educación o la sanidad más que a la transparencia o la fiscalidad, por ejemplo; o cómo determinados candidatos y agrupaciones políticas están estrechamente vinculados en el discurso sin estarlo necesariamente ideológicamente. A partir de la proyección obtenida por UMAP se han vuelto a someter a word2vec términos aparentemente cercanos para asegurar numéricamente la probabilidad de aparecer relacionados en el texto. De nuevo, se clarifica la cercanía de partidos o candidatos con calificativos —habitualmente negativos— y siempre más ligados a acontecimientos que a propuestas políticas. En cuanto a estas, por ejemplo, es visible cómo eldiario.es vincula el discurso relacionado con las propuestas de empleo y transparencia con el Partido Popular, mientras que para El País esta formación está más ligada a las medidas de turismo y sanidad durante la pandemia, y para ABC, con las propuestas de fiscalidad y movilidad.

4. Discusión

Este estudio aporta un punto de partida exploratorio y ofrece un acercamiento distinto al análisis de textos, en tanto que el estudio de los datos permite cuestionar al investigador qué está ocurriendo, por qué se distribuyen de esa forma las palabras, por qué algunos binomios de términos están relacionados o por qué son mucho más frecuentes unos que otros.

Se ha comparado este estudio con investigaciones que aplican diversos métodos similares e iguales a estos: TF*IDF (Qaiser y Ali, 2018), *support vector machines* (Cervantes *et al.*, 2020), *latent Dirichlet allocation* (Kim et al, 2018), o word2vec (Jang *et al.*, 2019) entre otras muchas técnicas (Kowsari *et al.*, 2019), si bien su uso es probadamente útil, pero empleado en su caso para la clasificación y no para la exploración. Otras nuevas propuestas (Campos *et al.*, 2020) serán muy útiles como alternativa a TF y TF*IDF para la extracción de *keywords* en fase previa a la aplicación de redes neuronales y proyección; o como investigación complementaria (Thavareesan y Mahesan, 2020). En todo caso, se ha querido también comparar los resultados de este análisis mediático con otras investigaciones previas de contextos —no metodologías— similares. Encontramos que Labio-Bernal (2018) y Sánchez Gutiérrez (2016) realizaron ambos análisis del discurso, de los que extraían que el partido político Podemos recibía un trato negativo de forma sistemática por parte de los medios, personalizado en su candidato Pablo Iglesias. La muestra, en estos casos, era menor que en este estudio, que, aunque no extrae conclusiones de ese peso, permite vislumbrar la efectiva relación entre dicha agrupación política y términos de connotación negativa, a veces peyorativa e incluso violenta.

Mancera-Rueda y Villar-Hernández (2020) realizaron un análisis del discurso de unos comicios anteriores con el que detectaron que al partido Vox se le relacionaba con los llamados *campaign-issues* y no con medidas electorales. Resultados similares se obtienen en la presente investigación, resaltando la probabilidad de que las siglas del partido o su cabeza de lista se vinculen a términos clave de acontecimientos de la campaña (mitin, debate, etc.) y otros relacionados con sucesos concretos (como inmigrantes o menores no acompañados, que generaron cierto conflicto durante la campaña), pero no con medidas electorales (fiscalidad, educación, sanidad, empleo, género...). Otros estudios cualitativos —en este caso de debates televisivos, realizados por García-Marín (2015) y García-Marín *et al.* (2018)— vislumbraron una tendencia hacia el infoentretenimiento al detectar la personalización de propuestas e ideas en personajes conocidos o carismáticos y palabras más propias del lenguaje del entretenimiento, "incluso pugilístico" (García-Martín, 2015), que de la información. Ideas similares se observan con la metodología aplicada en este estudio, que detecta numerosos adjetivos, términos cargados de connotación, peyorativos y personificados en el candidato político. Fenoll y Rodríguez-Ballesteros (2017) habían previamente detectado, con programas de análisis de texto que, entre las 50 palabras más empleadas por los medios para cubrir información electoral, muchas eran propias del lenguaje competitivo: líder, primero, segundo, último, ganar... Tal y como también se ha podido comprobar en esta investigación, Fenoll y Rodríguez-Ballesteros coincidían en una omisión por parte de los medios de los temas, en favor del contenido o de los sucesos de campaña, e incluso una "simplificación de la información al binomio ganador/perdedor", que coincide con los resultados aquí obtenidos en la polarización de

la cobertura, y que "priva a la audiencia de los elementos de juicio necesarios para una elección de voto fundamentada" (2017).

La combinación metodológica aquí empleada resulta sencilla de implementar y extrapolable a otros análisis de textos, especialmente cuando se trabaja con muestras numerosas, textos largos y redacciones complejas, distintas a las que comúnmente se encuentran en redes sociales, por ejemplo. La automatización de las técnicas propuestas es útil para la comparación, pues puede aplicarse en repetidas ocasiones sobre distintos casos para subrayar las diferencias y similitudes entre las fuentes estudiadas. Es, además, un punto de partida para otras investigaciones, tanto cualitativas como mixtas, que pueden explorar en profundidad de análisis los medios o temáticas que hayan destacado, desprendidas de esta aplicación metodológica. En base a la teoría fundamentada, permite un análisis a partir de los datos y un posterior análisis inductivo. La flexibilidad del método permite experimentar con distintos grupos de palabras, atendiendo a cualquier concepto que se trate desde los medios de comunicación o cualquier otra fuente documental. No solo es compatible, sino que resultaría muy útil combinar los resultados obtenidos con estos métodos con otros para el análisis de sentimiento, por ejemplo.

Con todo, esta metodología presenta ciertas limitaciones. La tecnología aplicada desvela algunos puntos de partida que deben ser tratados cuidadosamente. Una relación entre términos no implica que sea directa, ni permite extraer conclusiones *per se*. Es un método rápido, sencillo y escalable que no sustituye a los profundos análisis del discurso, por ejemplo, sino que se presenta como metodología complementaria, previa a estudios más hondos del contenido.

5. Referencias

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. *O'Reilly Media.*

Berven, A., Christensen, O., Moldeklev, S., Opdahl, A., & Villanger, K., (2020). A knowledge-graph platform for newsrooms. *Computers in Industry 123*. https://doi.org/10.1016/j.compind.2020.103321

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257-289. https://doi.org/10.1016/j.ins.2019.09.013

Casero-Ripollés, A., Feenstra, R., & Tormey, S. (2016). Old and New Media Logics in an Electoral Campaign The Case of Podemos and the Two-Way Street Mediatization of Politics. *The International Journal of Press/Politics, 21*(3), 378-397. https://doi.org/10.1177/1940161216645340

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing, 408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.1

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications, 7*(4), 285-294. https://doi.org/10.21512/comtech.v7i4.3746

Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., & Torii, M. (2019). Extracting health-related causality from twitter messages using natural language processing. *BMC medical informatics and decision making, 19*(3), 71-77. https://doi.org/10.1186/s12911-019-0785-0

Edell, A. (2018). I trained fake news detection AI with >95% accuracy, and almost went crazy. En *Towards Data Science.* https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c

Emadi, M., & Rahgozar, M. (2020). Twitter sentiment analysis using fuzzy integral classifier fusion. J*ournal of Information Science, 46*(2), 226-242. https://doi.org/10.1177/0165551519828

Fenoll, V., & Rodríguez-Ballesteros, P. (2017). Análisis automatizado de encuadres mediáticos. Cobertura en prensa del debate 7D 2015: el debate decisivo. *Profesional de la Información, 26*(4), 630-640. https://doi.org/10.3145/epi.2017.jul.07

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access 7*, 154290-154299. https://10.1109/ACCESS.2019.2946594

García-Marín, J., Calatrava García, A., & Luengo, Ó. G. (2018). Debates electorales y conflicto. Un análisis con máquinas de soporte virtual (SVM) de la cobertura mediática de los debates en España desde 2008. *Profesional de la información 27*(3). https://doi.org/10.3145/epi.2018.may.15

Goularas, D., & Kamis, S. (2019, August). Evaluation of deep learning techniques in sentiment analysis from twitter data. In *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications*, pp. 12-17. IEEE. https://doi.org/10.1109/Deep-ML.2019.00011

Iyengar, S., & Simon, A. F. (2000). New perspectives and evidence on political communication and campaign effects. *Annual review of psychology*, *51*(1), 149-169. https://doi.org/10.1146/annurev.psych.51.1.149

Jang, B., Kim, I., & Kim, J. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one, 14(*8). https://doi.org/10.1371/journal.pone.0220976

Jung, N., & Lee, G. (2019). Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. *Advanced Engineering Informatics, 41*, 100917. https://doi.org/10.1016/j.aei.2019.04.007

Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and

Doc2Vec. *Information Sciences, 477*, 15-29. https://doi.org/10.1016/j.ins.2018.10.006

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150. https://doi.org/10.3390/info10040150

Kuncoro, B.A., & Iswanto, B.H. (2015, November). TF-IDF method in ranking keywords of Instagram users' image captions. En *2015 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 1-5). IEEE. https://ieeexplore.ieee.org/document/7437705

Labio-Bernal, A. (2018). Anti-communism and the mainstream online press in Spain: Criticism of Podemos as a strategy of a two-party system in crisis. The Propaganda Model Today: Filtering Perceptions and Awareness. *University of Westminster Press*. https://doi.org/10.16997/book27

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, (pp. 1188-1196). PMLR. https://doi.org/10.48550/arXiv.1405.4053

Li, L., Johnson, J., Aarhus, W., & Shah, D. (2022). Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. *Computers & Education, 176*, 104354. https://doi.org/10.1016/j.compedu.2021.104354

Mancera-Rueda, A., & Villar-Hernández, P. (2020). Análisis de las estrategias de encuadre discursivo en la cobertura electoral sobre Vox en los titulares de la prensa española. *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, 315-340. https://doi.org/10.31921/doxacom.n31a16

Marshall, M. N. (1996). Sampling for qualitative research. *Family practice, 13*(6), 522-526. https://doi.org/10.1093/fampra/13.6.522

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McNair, B. (2017). An introduction to political communication. *Routledge.*

Miguel-Sáez-de-Urabain, A., Fernández-de-Arroyabe-Olaortua, A., & Lazkano-Arrillaga, I. (2017). La espectacularización de la información política. El caso de El País en las elecciones estadounidenses de 2016. *Revista Latina De Comunicación Social 72*, 1131-1147. https://doi.org/10.4185/RLCS-2017-1211

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503

Paniagua-Rojano, F., Seoane-Pérez, F., & Magallón-Rosa, R. (2020). Anatomía del bulo electoral: la desinformación política durante la campaña del 28-A en España.

Revista CIDOB d'Afers Internacionals 124, 123-146. https://doi.org/10.24241/rcai.2020.124.1.123

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29. https://doi.org/10.5120/ijca2018917395

Salton, G., Buckley, C. (1988). Term-Weighting approaches in Automatic Text Retrieval. *Information Processing and Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sánchez Gutiérrez, B. (2016). La representación mediática de los partidos políticos emergentes: el caso de Podemos y Ciudadanos en Atresmedia (Trabajo Final de Máster). *Universidad de Sevilla*.

Sánchez-Gutiérrez, B., & Nogales-Bocio, A. I. (2018). La cobertura mediática de Podemos en la prensa nativa digital neoliberal española: una aproximación al caso de OkDiario, El Español y El Independiente. En A.I. Nogales Bocio, C. Marta-Lazo, M.A. Solans García (Ed.), *Estándares e indicadores para la calidad informativa en los medios digitales*, (pp. 125-146).

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020). Measuring news sentiment. *Journal of Econometrics 228*(2), 221-243. https://doi.org/10.1016/j.jeconom.2020.07.053

Singh, K., Sen, I., & Kumaraguru, P. (2018, July). A Twitter corpus for Hindi-English code mixed POS tagging. En *Proceedings of the sixth international workshop on natural language processing for social media*, (pp. 12-17). https://doi.org/10.18653/v1/W18-3503

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion, 36*, 10-25. https://doi.org/10.1016/j.inffus.2016.10.004

Thavareesan, S., & Mahesan, S. (2020, July). Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. En 2020 *Moratuwa Engineering Research Conference*, (pp. 272-276). IEEE. https://doi.org/10.1109/MERCon50084.2020.9185369

Tian, X., & Tong, W. (2010). An improvement to TF: Term distribution based term weight algorithm. En *2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing 1*, (pp. 252-255). IEEE. https://doi.org/10.1109/NSWCTC.2010.66

Xia, T., & Chai, Y. (2011). An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm. *Journal of Software, 6*(3), 413-420. http://www.jsoftware.us/vol6/jsw0603-9.pdf

Vermeulen, M., Smith, K., Eremin, K., Rayner, G., & Walton, M. (2021). Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 252*, 119547. https://doi.org/10.1016/j.saa.2021.119547

Wongso, R., Luwinda, F. A., Trisnajaya, B. C., & Rusli, O. (2017). News article text classification in Indonesian language. *Procedia Computer Science*, *116*, 137-143. https://doi.org/10.1016/j.procs.2017.10.039

Zhou, P., Shi, W., Zhao, J., Huang, K-H., Chen, M., & Chang, K-W. (2019). Analyzing and Mitigating Gender Bias in Languages with Grammatical Gender and Bilingual Word Embeddings. *ACL*

Conflicto de intereses: los autores declaran que no existen.

Traducción al inglés: aportada por los autores.

HOW TO CITE (APA 7ª)

Castillo-Campos, M., Becerra-Alonso, D., & Varona-Aramburu, D. (2022). Métodos de Procesado del Lenguaje Natural aplicados al estudio de las coberturas mediáticas. *Communication & Methods - Comunicación y Métodos*, *4*(2), 85-99. https://doi.org/10.35951/v4i2.171