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

Natural Language Processing Methods Applied to the Study of Media Coverage

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

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

Article received: 08/10/2022 – Accepted: 16/11/2022

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

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

1. Introduction and state of art

The study of political communication is commonly linked to an analysis of candidates' discourse. In election campaign periods, moreover, attention is focused on the possible persuasion that politicians could exert on the population through the media (McNair, 2017), but this leaves many other important aspects out of analysis (type of information communicated, agenda setting, evaluation criteria for each candidate...) (Iyengar & Simon, 2000). Traditionally, the study of political communication was done through surveys and panels on the perception side, and with discourse analysis for content (Iyengar & Simon, 2000), although quantitative and artificial intelligence (AI) techniques are increasingly being used to automate part of the analysis and to work with larger samples.

In this research, techniques belonging to what is known as Natural Language Processing (NLP) are used. By "natural language" we mean the language used by people to communicate with each other daily, in any language (Bird et al., 2019, 9). It differs from "artificial languages" - such as mathematical or programming languages, for example - in that it does not follow such strict rules and evolves with time, type of use or geography, among other factors. This makes it more complex to analyze and study. The computational processing that this language deals with can be difficult, precisely because of the breadth, diversity, and complexity of what is being studied (Bird et al., 2019). NLP ranges from techniques such as automatic word counts, for example, which help to detect topics or writing styles, to more sophisticated systems that, supported by AI, manage to process these texts in a sort of "understanding" to respond to that language (Sun et al., 2017). Part of the complexity of this processing lies in the richness of the language (synonyms, polysemies, anaphora...), which is accentuated in literature, advertising, poetry, or journalistic texts, among others. Scientific literature tends to use simpler and more limited and canonical language, which means that NLP is applied more often and more successfully to texts of this type. These techniques, which are still under development, are revealing insights that go unnoticed in the use of other research methods, since the intentional focus has not always been placed on studying certain factors that have been now highlighted by exploratory, non-confirmatory methods.

In any case, it is a method that is not widely used in the field of journalism. The studies that precede this one usually manually compare a limited sample of articles, reading them, analyzing them and drawing conclusions about what topics these news pieces deal with, how they present each actor or what role they play (Casero-Ripollés et al., 2016; Mancera-Rueda and Villar-Hernández, 2020; Sánchez Gutiérrez, 2016). They also look for which ones receive more or less coverage, what feeling underlies the writing (Gao et al., 2019; Li et al., 2022; Miguel-Sáez-de-Urabain, et al., 2017; Sánchez Gutiérrez and Nogales Bocio, 2018; Shapiro et al., 2020)... Likewise, there are really numerous researches that quantitatively study language in social media conversation (Doan et al., 2019; Emadi and Rahgozar, 2020; Goularas and Kamis, 2019; Müller, 2020; Paniagua-Rojano et al., 2020; Singh et al., 2018...), although on this occasion we have opted for longer and more complex texts such as journalistic texts. Other researchers who have employed vectorization and quantization of documents have used a similar method to classify news: Orden-Cruz et al. (2019) apply, for example, NLP on headlines and news summaries to extract implicit sentiment. Riedel et al. (2017) use the same methodology (Term Frequency, Term Frequency and Inverse Document Frequency, and cosine similarity calculation) to classify news stories. A complementary method, with different purposes and potential in the field of communication, will be presented here.

2. Methodology

Most previous studies use qualitative methodology, which on the one hand allows for more limited samples and tends to carry out an in-depth analysis (Marshall, 1996, p. 523). Quantitative methodologies, on the other hand, mostly start from a classificatory intention (Berven et al., 2020; Edell, 2018; Jung and Lee, 2019; Zhou et al., 2019). This implies that the researcher has designed previous categories, into which the results are then "fitted". It is a very useful technique for some research, but on occasions, is limiting and biased, as it could commit the data to belonging to one of the established categories, leaving out other alternatives that were not proposed or other types of groupings. Its use is particularly widespread for the classification of documents, from whose entire sample a part has been selected and labelled, which requires the prior intervention of the researcher. Exploratory analysis, on the other hand, does not create categories, but rather groups the data and reveals the points of union and dissidence between them, with the researcher's task being to analyze, ex post, what they correspond to and whether it is appropriate to establish different categories. Based on grounded theory, these techniques are particularly useful for uncovering data, comparing it, and raising questions from it that might not come to light with other methods.

Based on this conviction, this research is carried out in three phases: first, counting terms; second, studying the relationship between binomials of concepts through the application of neural networks; and finally, grouping terms and projecting results.
The different methods used are described below:

2.1. Term Frequency

Term Frequency (TF) is an automatic count of words in the text, both absolute and relative to the total number of words in the same document. It is also called Local Term Weight, as it is a measure of an individual, non-comparative text's positive correlation with its frequency (Tian and Tong, 2010). While it is used as a starting point for many investigations (detection of significant terms, news labelling according to the relevance of a word, etc), it is also useful for the analysis of media coverage in order to:

extract keywords from each documentary source or media outlet,

as a starting point to detect topics, frequent terms or concepts, for each source or news item,

calculate the potential number of times a reader would have been exposed to different terms, ideas or candidates, or what importance has been given to a certain topic or character for each documentary source.

TF provides absolute data for each word and for each text, making it possible to establish a comparison, for example, between pieces of information about the same event in different media. A high TF rate for a word in a text could provide information on the subject matter of the text, and a comparison between rates for different words would allow a hierarchy to be established according to the importance given to them in the text.

To study the words that are relevant but frequent in all the texts to be compared, other techniques are used, such as Term Frequency times Inverse Document Frequency (TFIDF), which makes it possible to locate the source that gives most importance to a specific term. TFIDF is formulated as a factor between the relative frequency of a word or set of words in a text and the comparison of a whole selection of texts containing the same term (Salton and Buckley, 1988). It can highlight the most frequently used terms in each text or the source that is using certain concepts most often in comparison to a large body of texts.

Using TFIDF, a value of zero for a term indicates that it does not appear in the text. A hypothetical value of 1 indicates that the concept appears a relevant number of times in that text and does not appear in other texts in the sample. Values closer to 1 indicate that the terms studied are relevant in that text and are rarely mentioned in the others with which they are compared. Some terms, although mentioned a lot, are used by many of the different sources compared: in this case, the value of the concept is close to zero.

Typically, this method is used to detect keywords in texts (Qaiser and Ali, 2018; Kuncoro and Iswanto, 2015), to construct summaries of a longer corpus (Christian et al.,

2016) or for text classification based on differential concepts (Xia and Chai, 2011; Wongso et al., 2017), among others.

Both methodologies, TF and TFIDF, can be very useful as a starting point, as they could reveal insights about the informational priorities of each source.

2.2. Application of neural networks

Machine learning techniques can be used to measure the linking of the obtained concepts to each other. In particular, the word2vec model is one of the convolutional neural networks that allow the processing of natural language. Until then, the numerical conversion of terms into bags of words (BOW) had significant shortcomings, as the sense of the order of the words in the text and, therefore, the relationship established between them was lost. The semantics of the words were also ignored, not differentiating between synonyms, verbs, or connectors (Le and Mikolov, 2014). Thus, word2vec ("from word to vector") converts each term to a multi-dimensional numerical vector, a process by which it maintains the relationship between words and is able to trace syntactic and conceptual links between them (Mikolov et al., 2013). This neural network model can be applied to detect the probability that a given term is related to another term. This probability considers the distance or similarity between terms in a sample and considers, for each document, how closely two concepts are linked.

2.3. Projecting results

The comparison of multi-dimensional vectors obtained with word2vec is done by means of the operation with distance matrices, but other techniques allow the visualization of the data in a simpler and clearer way. The dimensional reduction algorithms precisely manage to transform the multiple dimensions of the vectors with as little loss of information as possible. The vectors created in previous phases are simplified and projected onto a two-dimensional graph, maintaining the distance between the terms studied. In this way, it is possible to observe how different terms are grouped together either because they fulfil similar syntactic functions, because they are usually related or because they are more likely to appear together in the text.

The dimensional reduction tool Uniform Manifold Approximation and Projection (UMAP) has previously shown better results in data projection and structure visualization than other reducers (McInnes et al., 2018; Vermeulen et al., 2021). Its exploratory use allows analysis of how words are placed in space. It is up to the researcher to consider whether the distribution obtained corresponds to clusters of terms or not. One of its main advantages is that it could reveal associations of concepts that would not have been considered at the beginning of the research and which open up possible lines of interest and research.

It should be stressed that the distances obtained indicate a relationship between concepts, but do not require correlation, proportionality, or reciprocity. While both the matrix and the term map allow for glimpses of insights possibly unnoticed through qualitative or classificatory analysis, the results should not be exempt from interpretation by professional researchers in the field under study.

The combination of techniques is useful in that the most frequent terms (obtained with TF) and the differential terms (TFIDF) are selected to study the distance between them (word2vec) and the distribution (UMAP).

3. Example of application

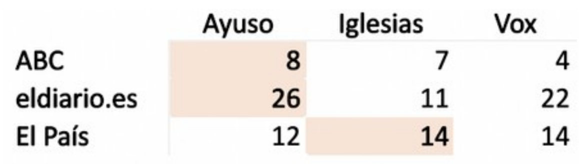
The methodology proposed in this study has been used to analyze the media coverage of the Madrid Assembly elections held on May 4th, 2021 (4M). These elections were of special interest due to the pandemic context in which they were held and because, despite being regional in scope, they achieved national coverage and were closely covered by the most important media in Spain. The large amount of news generated since the announcement of the elections, and particularly during the election campaign, shifted the news focus away from the pandemic to the Assembly candidates.

The methodology described above is proposed to investigate media coverage in the electoral context: to find out which political parties dominate the media agenda and which actors are reinforced; to determine which electoral proposals are related to them, and to evaluate the importance given to these with respect to the events in which the candidates are involved. The news sample (n=450), which includes all the news items related to the 4M elections during the two weeks of the electoral campaign that appeared in three of the most relevant media in Spain, justifies a methodology that automates and speeds up the analysis of the texts, and which allows for an exploratory - not classificatory - analysis of the news items.

TF highlights the high visibility of some actors in headlines, the abuse of adjectives or the large number of words with violent connotations, among others. For example, given a population of all the news related to the 4M election campaign for the media outlet eldiario.es, in just one week the candidate Isabel Díaz Ayuso appeared 26 times in headlines, followed by the political party Vox - 22 times - and the candidate Pablo Iglesias, on 11 occasions. However, for the newspaper El País, Pablo Iglesias is the most named person in headlines that same week, 14 times (Figure 1). Six of the competing parties and their list leaders were represented in the media, but the remaining groupings were not even mentioned.

Figure 1

Number of headline appearances of the three most named candidates or parties in the first week of campaigning

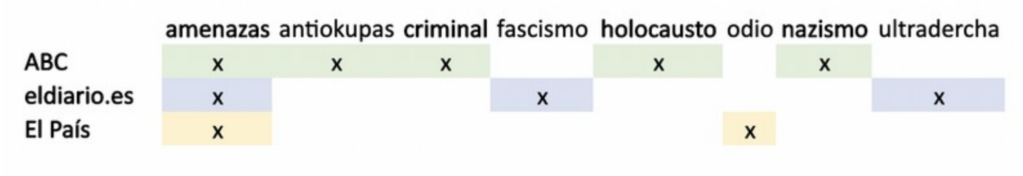


	Ayuso	Iglesias	Vox
ABC	8	7	4
eldiario.es	26	11	22
El País	12	14	14

In relation to words with an ideological or emotional connotation, ABC counted the words ‘anti-squatting’, ‘holocaust’, ‘Nazism’ and ‘criminal’ in its headlines at least once during one election campaign week. In eldiario.es, ‘threats’, ‘fascism’ and ‘far-right’ appear up to three times in headlines in just seven days (Figure 2).

Figure 2

Number of appearances in headlines by term in the first week of the campaign

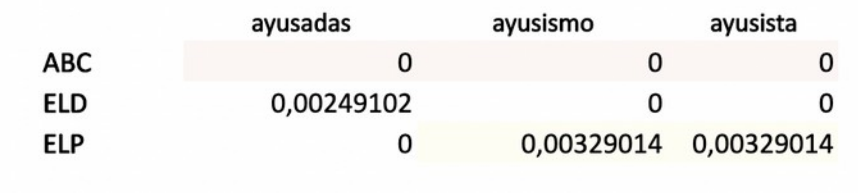


	amenazas	antiokupas	criminal	fascismo	holocausto	odio	nazismo	ultradercha
ABC	x	x	x		x		x	
eldiario.es	x			x				x
El País	x					x		

In this study, TFIDF has been used to detect which media have given more importance to certain terms, either to the candidates, to some political measures or to specific events that have taken place during the campaign. Thus, it has become apparent that for some media groups, more specific measures (economic, social, educational, fiscal, etc) are much more relevant than for others, and that certain minority party candidates receive coverage in some media only. It was also discovered that terms with pejorative, negative and even violent connotations are very frequent in some newspapers and appear a little in others (Figure 3).

Figure 3

TFIDF ratio for the three media outlets



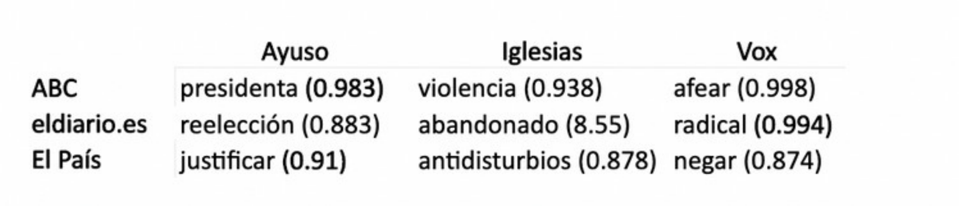
	ayusadas	ayusismo	ayusista
ABC	0	0	0
ELD	0,00249102	0	0
ELP	0	0,00329014	0,00329014

In the case of this example, ‘ayusada’, ‘ayusismo’ or ‘ayusista’ are used in a derogatory way, referring to something proper, prone to or related to the candidate Isabel Díaz Ayuso. For ABC, the TFIDF rate is zero: none of these words have been used in the news published by this media. El País and eldiario.es did use them in their news pieces, although not with special assiduity - that is why the TFIDF rate is low. This is an example of the words that spontaneously result from the application of TFIDF: although these terms had not been considered at the beginning of the research, a zero rate for one medium and a non-zero rate for others points the researcher to a difference to review.

At a later stage, the use of word2vec has allowed us to check whether a media outlet relates a party to certain political proposals, issues or concepts. In both El País and eldiario.es, Edmundo Bal of the Ciudadanos party is the one who is most related to taxation proposals. For both media, employment policies are more closely linked to Isabel Díaz Ayuso, candidate for the Partido Popular (PP). For the newspaper eldiario.es, the terms most likely to appear next to the head of the PP list include 'victory' or 're-election', and her opposite number, Pablo Iglesias, is usually close to 'death' or 'abandoned' (Figure 4).

Figure 4

Binomial of terms with maximum probability of appearing together, for each medium



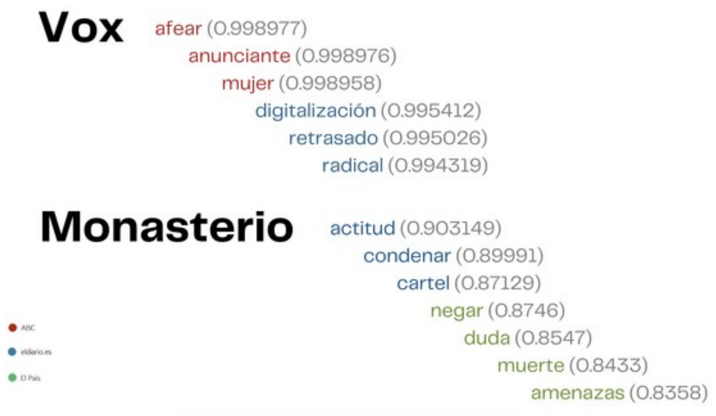
	Ayuso	Iglesias	Vox
ABC	presidenta **(0.983)**	violencia (0.938)	afear (0.998)
eldiario.es	reelección (0.883)	abandonado (8.55)	radical **(0.994)**
El País	justificar **(0.91)**	antidisturbios (0.878)	negar (0.874)

Among the terms closest to Vox or its list leader Rocío Monasterio are ‘to spoil’, ‘radical’ or ‘doubt’, for ABC, eldiario.es and El País, respectively.

In a third phase, it is used the dimensional reductor tool UMAP, with which it is observed how the candidates are positioned as similar terms, fulfilling similar syntactic functions, used in the same contexts and with a certain link to each other (Figure 5). On the other hand, the political proposals are also positioned in proximity (‘freelances’, ‘companies’, ‘taxes’; ‘sustainability’, ‘mobility’; ‘pandemic’, ‘health care’, ‘nursing homes’...). Other terms, more loaded with sentiment, are usually associated with the same actors or political measures (Iglesias, close to ‘death’ and ‘threats’, and those related to Vox, Monasterio and Abascal, with ‘extreme’ and ‘falsehoods’).

Figure 5

Projection of terms with UMAP from the media eldiario.es



It is interesting to explore these links, which would have been difficult to extract with other techniques. The relationship of closeness or remoteness between terms should be studied and opens up other lines of research: why for a media outlet 'corruption' seems more closely linked to 'investment', 'education' or 'health care' rather than 'transparency' or 'taxation', for example, or how certain candidates and political groupings are closely linked in the discourse without necessarily being ideologically so.

Based on the projection obtained by UMAP, apparently close terms have been re-subjected to word2vec to ensure numerically the probability of appearing related in the text. Once again, the proximity of parties or candidates with qualifiers - usually negative - and always more linked to events than to political proposals, is clarified. As for the latter, for example, it is visible how eldiario.es links the discourse related to 'employment' and 'transparency' proposals with Partido Popular, while for El País this party is more closely linked to 'tourism' and 'health care' measures during the pandemic, and for ABC, with 'taxation' and 'mobility' proposals.

4. Discussion

This study provides an exploratory starting point and offers a different approach to the text analysis, in that the study of the data allows the researcher to question why the words are distributed in this way, why some pairs of terms are related or why some are much more frequent than others.

This study has been compared with research that applies various methods similar and equal to these: TFIDF (Qaiser and Ali, 2018), support vector machines (Cervantes et al., 2020), latent Dirichlet allocation (Kim et al, 2018), or word2vec (Jang et al., 2019) among many other techniques (Kowsari et al., 2019), although their use is approvedly useful, but employed in their case for classification and not for exploration. Other new proposals (Campos et al., 2020) will be very useful as an alternative to TF and TFIDF for keyword extraction prior to the application of neural networks and projection, or as complementary research (Thavareesan and Mahesan, 2020). In any case, we also wanted to compare the results of this media analysis with other previous research in similar contexts - not methodologies. We found that Labio-Bernal (2018) and Sánchez Gutiérrez (2016) both carried out discourse analyses, from which they extracted that the political party Podemos was systematically treated negatively by the media, personalized in its candidate Pablo Iglesias. In these cases, the sample was smaller than in this study, which, although it does not draw conclusions of this weight, allows us to glimpse the relationship between this political party and terms of negative connotation, sometimes pejorative and even violent.

Mancera-Rueda and Villar-Hernández (2020) carried out a discourse analysis of previous elections in which they found that the Vox party was associated with the campaign-issues and not with electoral proposals. Similar results are obtained in the present research, highlighting the probability that the party's acronym or its head of list are linked to key terms of campaign events (*rally*, *debate*, etc) and others related to specific events (such as immigrants or unaccompanied minors, which generated some conflict during the campaign), but not to electoral measures (taxation, education, health care, employment, gender violence...). Other qualitative studies - in this case of television debates, carried out by García-Marín (2015) and García-Marín et al. (2018) - glimpsed a trend towards infotainment by detecting the personalization of proposals and ideas in well-known or charismatic characters and words more typical of the language of entertainment, "even pugilistic" (García-Martín, 2015), than of information. Similar ideas can be observed with the methodology applied in this study, which detects numerous adjectives, terms loaded with connotation, pejorative and personified in the political candidate. Fenoll and Rodríguez-Ballesteros (2017) had previously detected, with text analysis software, that among the 50 words most used by the media to cover electoral information, many were typical of competitive language: leader, first, second, last, win... As has also been verified in this research, Fenoll and Rodríguez-Ballesteros agreed that the media omit topics in favour of content or campaign events, and even a "simplification of the information to the winner/loser binomial", which coincides with the results obtained here in the polarization of coverage, and which "deprives the audience of the necessary elements of judgement for an informed voting choice" (2017).

The methodological combination used here is simple to implement and can be extrapolated to other text analyses, especially when working with large samples, long texts, and complex wording, different from that commonly found in social media, for example. The automation of the proposed techniques is useful for comparison, as it can be applied repeatedly on different cases to highlight the differences and similarities between the sources studied. It is also a starting point for other research, both qualitative and mixed, which can explore in depth the analysis of the media or themes that have emerged from this methodological application. Based on grounded theory, it allows for an analysis of the data and a subsequent inductive analysis. The flexibility of the method allows experimentation with different groups of words, taking into account any concept dealt with in the media or any other documentary source. Not only is it compatible, but it would be very useful to combine the results obtained with these methods with other methods for sentiment analysis, for example.

However, this methodology has certain limitations. The technology applied reveals some starting points that need to be treated carefully. A relationship between terms does not imply that it is straightforward, nor does it allow conclusions to be drawn per se. It is a quick, simple and scalable method that does not replace in-depth discourse analysis, but is presented as a complementary methodology, prior to more in-depth content studies.

5. References

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*

Conflict of interest: the authors declare that there is no conflict of interest.

English translation: provided by the authors.

HOW TO CITE (APA 7ª)

Castillo-Campos, M., Becerra-Alonso, D., & Varona-Aramburu, D. (2022). Natural Language Processing Methods Applied to the Study of Media Coverage. *Communication & Methods - Comunicación y Métodos, 4*(2), 85-99. https://doi.org/10.35951/v4i2.171