RIGL
REVISTA IBEROAMERICANA DE GOBIERNO LOCAL

LA MINERÍA DE TEXTO COMO MÉTODO DE ANÁLISIS: UNA MIRADA A LA ADMINISTRACIÓN LOCAL EN LA PRENSA ESPAÑOLA.

Ignacio Jesús Serrano Contreras
PhD Candidate en la Universidad de Granada
España

Resumen:
El propósito de este artículo es el de mostrar una ampliación inter y multidisciplinar de las Ciencias Políticas y, en especial, de la Comunicación Política. Este trabajo expone un breve acercamiento a algunas de las posibilidades que ofrece la computación como método de extracción, cuantificación, clasificación y generación de datos. Pretendemos poner en valor a la minería de texto como medio para la comprensión de un gran número de observaciones de una forma rápida y expeditiva. A través de las tres principales cabeceras de la prensa en España (ABC, El Mundo y El País), se lleva a cabo un modelo para la búsqueda de patrones y marcos en noticias de ámbito local.

Palabras clave:
Algoritmos, Minería de Texto, Comunicación Política, Medios de Comunicación, Administración Local.

Abstract:
The goal of this paper is to show an interdisciplinary and multidisciplinary extension of Political Science and, in particular, of Political Communication. This study presents a brief approach to some of the possibilities offered by computing as a method of data extraction, quantification, classification and generation. We intend to highlight text mining as a means of understanding a large number of observations in a rapid and expeditious manner. Through the main newspapers in Spain (ABC, El Mundo and El País), we implement a model to search style patterns about coverage local management.

Key words:
Algorithms, Text Mining, Political Communication, Mass Media, Local Management.

RIGL
REVISTA IBEROAMERICANA DE GOBIERNO LOCAL

Introducción

Las ciencias sociales, al igual que el resto de ramas del conocimiento, han sufrido cambios en los que sus conceptos se iban difuminando a la par que sus espectros se abrían hacia otras ciencias colindantes. En este particular, la Comunicación Política emergió anidando sus estudios en los métodos y análisis de la Ciencia Política. Esta ligazón se expande con el cambio de paradigma que sufren los medios de masas, anclándose en el establecimiento de la prensa escrita como contrapoder (Curran, 2005) y fiscalizador de las democracias liberales. En torno a este nuevo marco de relaciones surgen estudios que resaltan la capacidad de los medios para fijar temáticas, guiar la opinión pública o servir de muleta para el desarrollo e implementación de determinadas políticas públicas (McCombs y Shaw, 1972).

Mientras que por un lado los estudios comenzaban a fraguarse fijando la mirada en las relaciones de poder, otra batida de estos se centraba en los efectos que, desde la prensa al cine (Blumer, 1933), tenían sobre el público. Esta dicotomía sobre los efectos y el poder, origina unos nexos para encontrar un paradigma común. Se ve reflejado, por ejemplo, en el influjo que provoca el contexto (Lippmann, 1922) o en los mecanismos para seleccionar la información (Katz y Lazarsfeld, 1955), su método y su medio.

De este modo, se generan nuevos significantes que propugnan por la comprensión de las relaciones emergidas entre comunicación, comportamiento y poder. Se orquestan nuevas metodologías y se congregan epistemologías que buscan operacionalizar significantes presentes en múltiples áreas de la investigación. Surgen nuevos arquetipos como *public choice*, esfera púbica, *agenda setting*, *framing* o *priming*, construidos de forma multidisciplinar. Son definidos para ampliar la comprensión de los nuevos entresijos sociales que bullen durante el siglo XX. Supone una aproximación entre ciencias; desde la economía (teoría de juegos) a la psicología (comportamiento), hasta incluir otras más lejanas como las neurociencias (Luengo, 2016).

Este discurrir de los avances hace que la nueva era que hoy día tenemos presente, la digital, produzca nuevos enfoques que no solo desbordan al ámbito académico, sino que, además, agregan un cariz sintomático para comprender la sociedad actual. Aportaciones como el *big data*, el m*achine learning* o la inteligencia artificial, se han ido colando en el acervo común. Esto permite que sus aportaciones se vayan diseminando más allá de las barreras de su ámbito original de estudio. A su vez, esta serie de actividades ayudan a solventar disyuntivas que se creían insoslayables, como la subjetividad (Arce y Menéndez, 2018), aportando factores objetivos.

La inclusión de estos avances dentro de la Ciencia Política permite que la comprensión de actividades, hasta ahora poco abordables por su magnitud, se convierta en una realidad. Los procesos de máquina, expuestos por la algoritmia a través de la computación, hacen que tareas complejas se vuelvan rápidas y sencillas. Por ello, asumir nuevos rumbos, como la minería de texto, nos permite alcanzar escenarios difíciles de abordar sin la automatización. Esa descodificación del lenguaje humano, en aras de una comprensión de máquina, propicia esa traslación de lo cualitativo a lo cuantitativo. Esto posibilita que el texto y el contexto se abran para la búsqueda e interpretación de patrones (Justicia, 2017), variables y dejes que los humanos ejecutamos cuando nos comunicamos.

Este trabajo tiene como misión indagar en la inclusión de nuevos métodos, anclados en la computación, dentro de la comunicación política. La necesidad que subyace es la de cimentar la búsqueda de marcos imperceptibles para el procesamiento humano. El objetivo de este estudio será implementar algunas técnicas de minería de texto, para analizar la cobertura y estilografía que la prensa española lleva a cabo sobre el ámbito de la administración local.

Minería de texto

La minería de texto se empieza a estructurar de forma sólida a principios de los 90. Cobra un sentido más preeminente con la llegada de Internet y la web 2.0 y se vuelve un baluarte con el surgimiento de la digitalización, y su enorme cantidad de datos inasumibles para el desempeño humano. Inmersos por la eficiencia y eficacia, las empresas se lanzan a concretar modelos que produzcan un conocimiento más profundo de sus clientes, sus competidores, sus cuentas… Surge así, no solo un interés por los balances, sino también por los emails, los comentarios y las correspondencias.

La idea emana de una traslación de datos cualitativos, en este caso palabras, a otros eminentemente cuantitativos. Siguiendo a Justicia de la Torre (2017), la pieza angular del proceso se conforma con la obtención explícita de información, a través de datos implícitos y desestructurados. Según Don R. Swanson (1991), se pretende una búsqueda y definición de patrones en colecciones de texto (Contreras, 2016), con la intención de producir nuevo conocimiento. Este será empleado para alcanzar nuevas relaciones y conclusiones que permitan armar de forma pragmática, una combinación entre hombre y máquina (Eíto y Senso, 2004).

Esta simbiosis se construye en fases que fluctúan en función de nuestros intereses: obtención, cribado, limpieza –y depuración-, procesamiento, resultados... Todas tienen como fin capitalizar el conocimiento que se alberga en nuestros corpus. Siendo esencial para esta reinterpretación, el Procesamiento Natural del Lenguaje –PNL-, que es lo que provoca que la comunicación humana se vuelva legible para la computadora. De ahí se podrán obtener distintas aplicaciones que van desde las clasificaciones o proyecciones, a generaciones automatizadas de texto. El proceso se pone en relieve desde nuestro buscador online, pasando por nuestros filtros de *spam* en el correo, hasta la escritura automatizada de un ensayo. Todas, en mayor o menor medida, han de minar las muestras textuales a través del empleo de modelos pertinentes según nuestro interés.

En este estudio consideramos que esa composición es desarrollada por el análisis documental a través de la teoría del framing. Los preceptos sugieren la presencia de distintos marcos (Goffman, 1974) dentro de una cobertura mediática. Estos se conjugan a través de las distintas visiones de un mismo suceso; bien por factores adheridos a los posicionamientos ideológicos, bien por ciertos síntomas de enunciación presentes en la forma de confeccionar una noticia. La cuestión a responder después sería cuán de concreto es el interés (Reig, 2009), y cómo esto, repercutiría en la capacidad cognitiva del espectador. De este modo, la minería de texto podría aportar una automatización del análisis del contenido pertinente para la búsqueda de estos patrones; además de computarlos desde una objetividad numérica, encontrando por ejemplo: búsqueda de palabras clave, relaciones entre términos, evoluciones de significados, frecuencia de usos.

Estudio y muestra

Para desarrollar los objetivos marcados por nuestro estudio, se va a implementar un análisis semántico y estilográfico. A continuación, y mediante el uso de la minería de texto, se pretende cuantificar preceptos cualitativos presentes en la construcción de piezas periodísticas. Trabajos como los propuestos por Justicia (2017), Wilkerson y Casas (2017) o García-Marín et al., (2018), nos sirven de puntal para, con estos y otros mimbres, obtener una mayor compresión del ámbito social.

El objetivo principal es encontrar marcos presentes en los medios de comunicación. A través de las tres principales cabeceras españolas (ABC, El Mundo y El País) -versión papel-, proponemos un análisis de la temática de la administración local. Así establecemos que:

H1. La temática de la administración local no reporta diferentes encuadres en la prensa española.

El corpus del estudio ha sido tomado de la base de datos *My news* mediante la combinación de las técnicas de búsqueda parametrizada y booleana. La palabra clave ("administraciones locales")[1] ha comprendido el periodo 2009 a 2019. A la hora de realizar la búsqueda, el topic podría estar tanto en el titular como en el cuerpo de la noticia. Cribado y eliminado el ruido de la muestra, se obtuvieron un total de 541 observaciones: ABC (180), El Mundo (127) y El País (234).

Obtenidas las piezas se procedió a la limpieza, depuración y estructuración pertinentes de este tipo de estudios. Para realizar estas tareas se empleó el lenguaje de programación *R*. En primer lugar, se creó una función de limpieza con la que transformar todo el texto; conversión a minúsculas, eliminación de valores numéricos, signos de puntuación, caracteres extraños y espacios en blanco análogos que pudieran crearse. Una vez el corpus se estandarizó, se procedió a su *tokenización*; división del texto por palabras. Posteriormente se filtraron también las *stopwords*, o palabras vacías, aquellas que no aportaban significado y podían hacer que el procesamiento del lenguaje no captara el valor íntegro del texto.

Métodos y resultados.

A continuación se presentan los distintos análisis que, mediante minería de texto, se han llevado a cabo. Se presentan, de forma ordinal, los procesos establecidos para la cuantificación y estructuración de la información pertinente y relevante de los textos analizados.

La tokenización de los textos arrojó los siguientes datos:

- ABC (93143 palabras)
- El Mundo (78110 palabras)
- El País (159058 palabras)

[1] Se ha procedido al uso del plural ya que podría arrojar unos resultados más nítidos a nuestra investigación.

La Figura 1 nos muestra tanto la longitud media de las piezas, como su desviación típica:

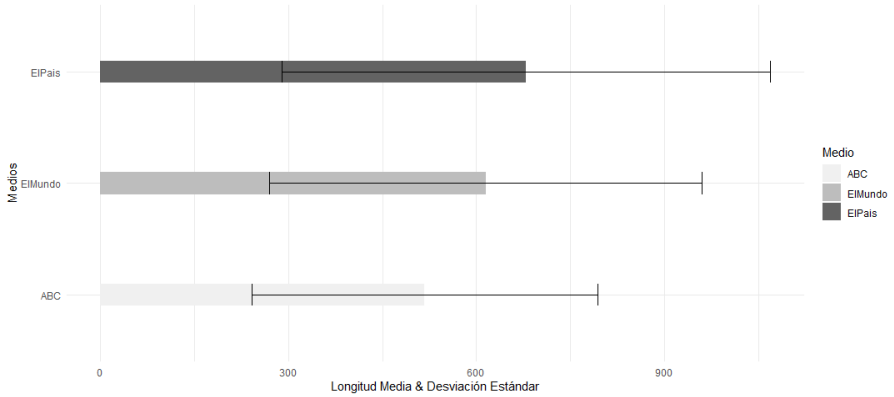

Figura 1. Longitud media y desviación estándar por pieza

En cuanto a la Figura 2, nos ofrece cuáles fueron las 10 palabras más usadas por los medios:

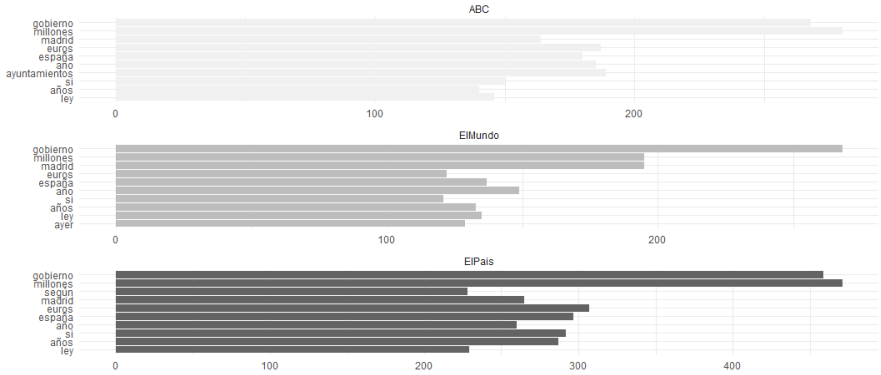

Figura 2. Top 10 de palabras más usadas por medio

Una vez obtenidas las cantidades y las palabras más utilizadas, se procedió a un análisis de correlación para comprobar los usos. Mediante una correlación de Pearson se obtuvo:

- El País & ABC
  0.9254892
- ABC & El Mundo
  0.9192969
- El Mundo & El País
  0. 9146469

Esta correlación de usos también puede observarse de forma individual y en base a la palabra empleada. Así, podemos comprobar que los medios tienden a usar una cantidad muy similar de palabras en todos sus textos (Figura 3). A su vez, estas palabras nos permiten testear cómo la mayoría de esos tokens, se distribuyen entre los medios en similar evolución Frecuencia/Ranking. La representación de la Figura 4 está producida según la *Ley de Zipf*. George Zipf (1932) establece una relación matemática donde la frecuencia de uso irá decreciendo proporcionalmente a su ranking de clasificación

(Montemurro, 2001). Se puede comprobar desde distintas acciones; atendiendo a los párrafos como forma de división, hasta en base a los idiomas. La figura lo que demuestra es que el empleo de palabras se torna muy similar.

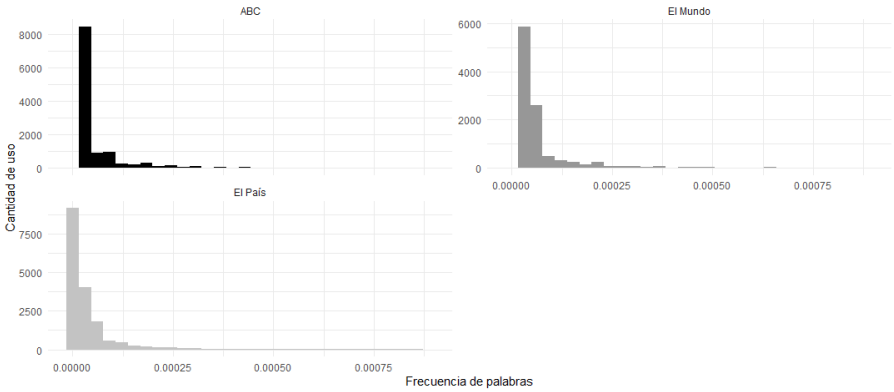

Figura 3. Rango frecuencias/usos de palabras

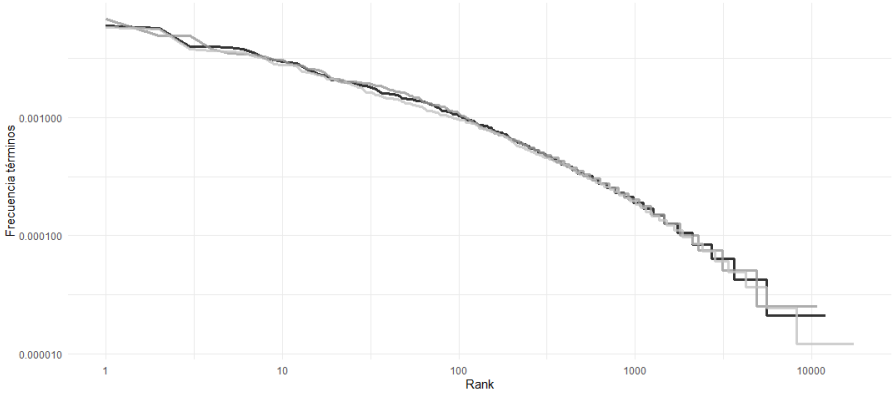

Figura 4. Ranking uso con la ley de Zipf

Posteriormente se realizó un análisis diferencial con la pretensión de ver qué palabras eran las más contrapuestas. Para ello se realizó un *log of odds ratio* que sirve para medir las distancias en el empleo de palabras. Un odds mide la probabilidad de ocurrencia de un suceso (Cerda; Vera; Rada; 2013) así como la probabilidad de que este hecho no suceda. Asimismo, este ratio se emplearía a través de la concurrencia entre dos odds con el fin de comprobar de forma cruzada las relaciones causales entre dos fenómenos (Aedo, Pavlov, Clavero, 2010). En nuestro caso, las divisiones ofrecerían de nuevo tres conjuntos distintos de comparación, conjugando así la siguiente representación de las 30 palabras por presencia/ausencia entre medios:

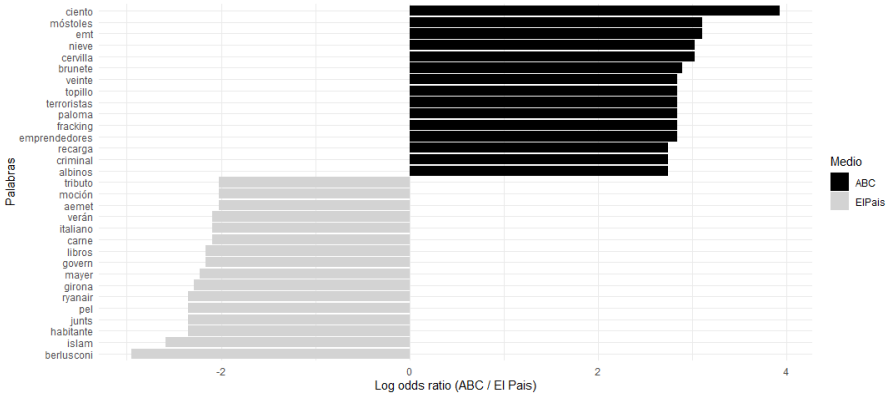

Figura 3. Odds ratio (ABC/El País)

RIGL
REVISTA IBEROAMERICANA
DE GOBIERNO LOCAL

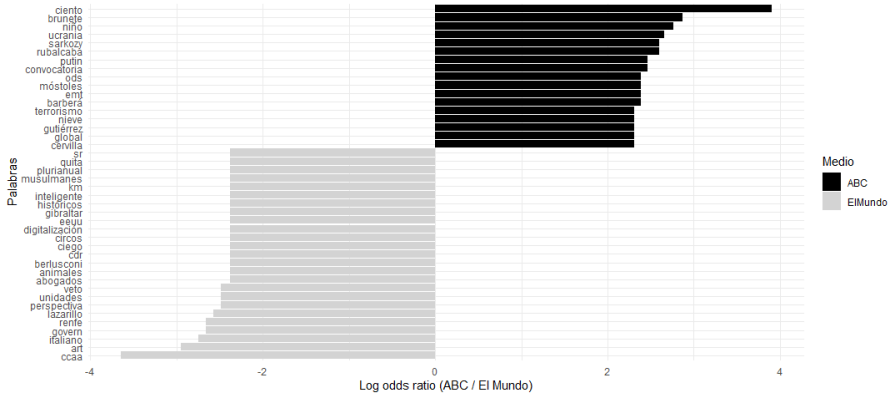

Figura 4. Odds ratio (ABC/El Mundo)

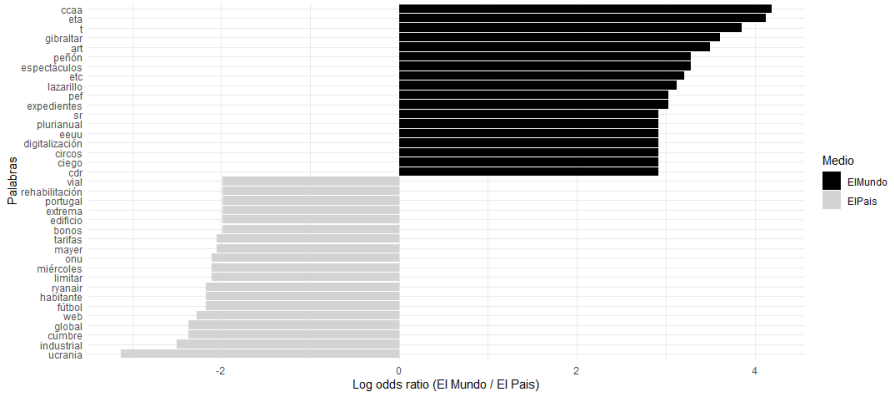

Figura 5. Odds ratio (El Mundo/El País)

Dentro del abanico de posibilidades de la minería de texto, podemos desarrollar un modelo matemático que ofrezca las co-ocurrencias por partículas semánticas, en este caso, piezas periodísticas. Dejando a un lado los preceptos de una *Cadena de Markov*, donde el término relacionado solo puede conocerse en base al término conocido, este que hemos empleado nos permite conocer las relaciones nodales por partículas. Así, podríamos seleccionar algunas palabras clave para establecer las probabilidades. En base al *Coeficiente de Phi*, se marca una relación dicotómica de pares, midiendo la probabilidad de un valor con respecto al conjunto de sus subconjuntos. De este modo, que un suceso ocurra, ha de depender del conjunto de ítems analizados, tal y como señala el modelo de Rasch (1960) (Heine, 2020), empleado en el paquete *pairwise* de R. Nuestro estudio seleccionó "*ayuntamientos*" y "*diputaciones*" para comprobar sus pares y con ello las relaciones entre medios. Como se comprueba en las gráficas, hay cierta disonancia en el empleo de palabras en base a un ítem de partida. Para completar el análisis, se añade el grado de correlación entre medios en función de la probabilidad de usar determinadas palabras:

Abc – El País	0.3124085
El País – El Mundo	0.2402973
Abc – El Mundo	0.2358577

Tabla 1. Correlaciones de probabilidades "ayuntamientos" y "diputaciones" por medio

RIGL
REVISTA IBEROAMERICANA
DE GOBIERNO LOCAL

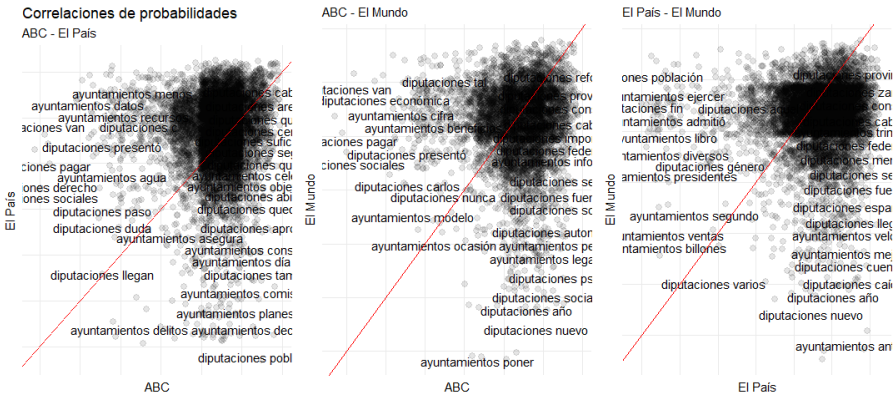

Figura 6. Correlaciones de probabilidades "ayuntamientos" y "diputaciones"

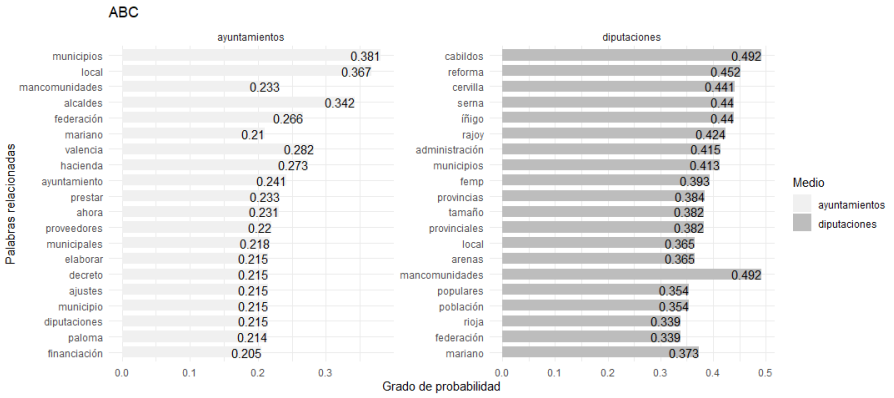

Figura 7. Relaciones de pares (*ABC*)

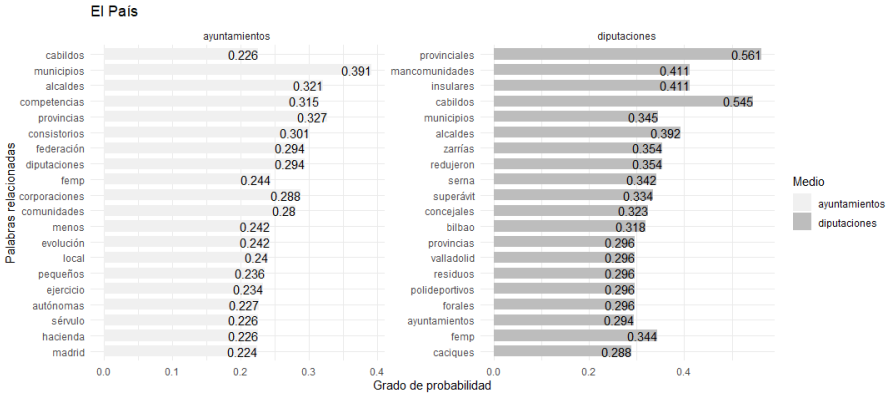

Figura 8. Relaciones de pares (*El País*)

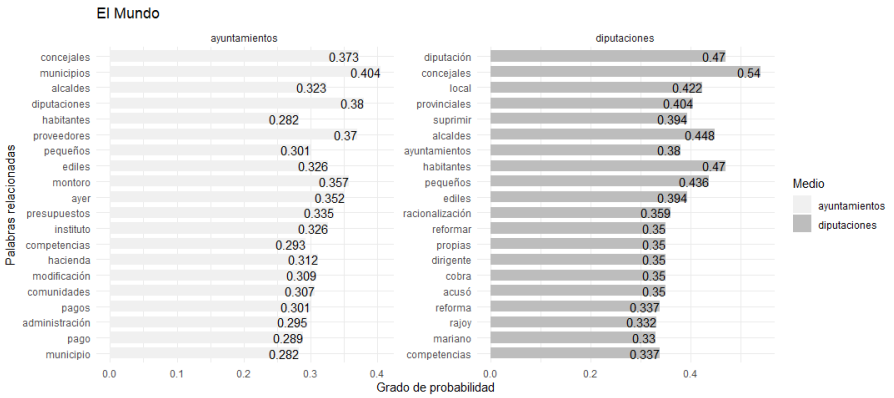

Figura 9. Relaciones de pares (*El Mundo*)

En último lugar, se realizó un análisis de sentimientos temporal, en el que atestiguar una posible evolución de la muestra. Esta técnica se implementó mediante el

RIGL
REVISTA IBEROAMERICANA DE GOBIERNO LOCAL

diccionario ML-Senticon (Cruz et al., 2014), encargado de puntuar la tokenización de nuestro corpus, para después establecer una media por pieza:

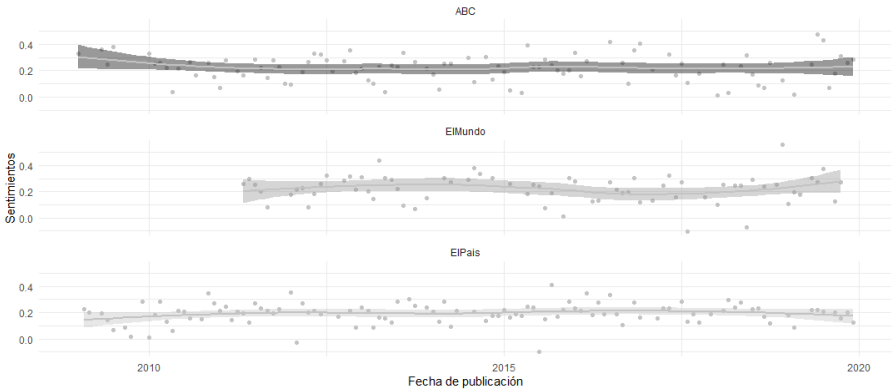

Figura 10. Análisis de sentimientos

Conclusiones

La primera conclusión que arroja el estudio evidencia las posibilidades que ofrece la ciencia de la computación y, en particular, la minería de texto. A pesar de ello, los diferentes procesos empleados no serían suficientes para validar nuestra hipótesis, ya que los detalles que en su mayoría se recogen representan factores estilográficos, no pudiendo mostrar diferencias ideológicas y editoriales de gran calado.

Hay varios factores que se manifiestan indispensables para entender tales resultados: tamaño de la muestra, tópic de búsqueda genérico, temática poco polarizante… Tales condicionantes no permiten a los datos mostrar una diferenciación sólida entre periódicos. A su vez, y como se ha mencionado, este conjunto de métodos lo que realizan es una conversión de variables subjetivas a otras objetivas. Se ha de desarrollar otra serie de modelos que, con un carácter automatizado y semiautomatizado, nos permitan evolucionar el procesamiento natural del lenguaje y su comprensión contextual. Sin obviar su capacidad, la metodología expuesta nos sugiere una compresión acelerada de la realidad y que, una vez ampliada y supervisada por el conocimiento humano, puede ofrecer grandes resultados. Si bien no es sencilla la conjunción de tales empréstitos, sí podemos augurar que la legibilidad puede aumentar el entendimiento de la comunicación humana. Procesos como los probabilísticos para conocer la presencia/ausencia de relaciones por contexto, nos abren la puerta a estructuras más complejas de machine learning. Métodos como los LSI o LDA producen una conjunción más elaborada de, por ejemplo, las temáticas que se manifiestan en las piezas. Así, gracias a una clasificación, ya sea lineal (e.g., García-Marín y Calatrava, 2018) o de clúster, poder encontrar esos marcos y patrones presentes en los textos.

En referencia a los resultados, si bien la base de datos no aportó un periodo concreto del diario El Mundo, el que corresponde de 2009 a 2011, esto no entorpeció el análisis y simplemente explica la variación en el número de palabras. El tamaño de los artículos se torna similar, aunque ABC muestra una cantidad media de tokens algo menor con respecto a los otros medios. Esto puede deberse básicamente a la composición que este periódico suele hacer de sus páginas impresas. Asimismo, aunque las desviaciones típicas del número de palabras por artículo dilucidan alguna variabilidad, estas no ofrecerían una especial importancia más allá de las distancias

medias citadas. Ocurriría lo mismo con las diez palabras más usadas por medio, coincidiendo entre los tres en nueve de ellas. A su vez, en el análisis de correlación de palabras, no muestra atisbo de dudas, con unos promedios superiores a 0.9, por lo que la correlación de usos es casi total, como se ve también la Figura 4, de la *Ley de Zipf*. En lo que respecta al pivotaje del ratio de odds, sí cabría algún comentario por las diferencias que refleja ABC. Estas se evidencian principalmente cuando se contrapone con El País, que destaca por el uso de topónimos en catalán, como es el caso de "girona". Algo que también ocurriría con el empleo de "govern" tanto por parte de El Mundo como de El País, en contraposición con ABC. Se pone así de manifiesto un claro posicionamiento de ABC en el uso del castellano con respecto a otras líneas editoriales, que prefieren utilizar palabras en catalán para citar determinadas ciudades o instituciones.

La parte más interesante del análisis estaría en las asociaciones de palabras. A pesar de que nuestro estudio está circunscrito a la palabra clave "administraciones locales", quisimos ahondar en qué supone para los medios, dentro de este conjunto, "ayuntamientos" y "diputaciones". Esta elección se debe a que la presencia de nuestro topic, por su magnitud, podría desviar los resultados. Dentro del proceso llevado a cabo, sí queremos dejar claro que los ratios son bajos, tanto de las muestras, como se ve en las Tablas 2 y 3, como de las correlaciones, al menos las que puedan ser significantes para cariz ideológico. El estudio lo que permite, comprobando las Figuras 6 a 9, y la Tabla 1, es que hay nexos de unión, pero no tan fuertes como en el de palabras más usadas; véase Figura 4. Sí se manifiesta una temática, la económica. Esta se torna relevante, esencialmente por el periodo, que abarca la práctica totalidad de la crisis de 2008. Si bien los contextos giran en torno a las competencias y reducción de gasto, lo que parecen mostrar los datos es una predilección de ABC y El Mundo, por una preponderancia al tema de los pagos, la financiación y las reformas. No es de extrañar que se produzca esto, ya que el gobierno de aquel periodo, el PP, podría estar más próximo a la línea editorial de estos medios. Aún así, si se quisiera ahondar en la concreción de temáticas y líneas editoriales, se debería emplear un trabajo más conciso y con factores metodológicos que arrojasen una comprensión más nítida de las posibles discrepancias entre medios.

Referencias

Arce, Sergio & Menéndez Menéndez, María Isabel. (2018). Aplicaciones de la estadística al framing y la minería de texto en estudios de comunicación. *Información, Cultura y Sociedad*. 39. 61-70.

Aedo, Sócrates; Pavlov, Stefanía., & Clavero, Francisca (2010). Riesgo relativo y Odds ratio ¿Qué son y cómo se interpretan? *Rev. Obstet. Ginecol*. - Hosp. Santiago Oriente Dr. Luis Tisné Brousse. 2010; VOL 5 (1): 51-54.

Blumer, Herbert. (1993). Movies and conduct. Nueva York (*Payne Fund Studies*).
Cerda, Jaime., Vera, Claudio., & Rada, Gabriel. (2013). Odds ratio: aspectos teóricos y prácticos. *Revista médica de Chile*, 141(10), 1329-1335. https://doi.org/10.4067/s0034-98872013001000014

Contreras Barrera, M. (2016). Minería de texto en la clasificación de documentos digitales. Biblios: Journal of Librarianship and Information Science, (64), 33-43. https://doi.org/10.5195/biblios.2016.309.

Cruz, Fermín. L., Troyano, Jóse. A., Pontes, Beatriz., & Ortega, Francisco. J. (2014). ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, 53, 113-120.

Curran, James. (2005). Medios de comunicación y poder en una sociedad democrática. Madrid, España: Alianza Editorial.

Eíto Brun, Ricardo., & Senso, Jose A. (2004) "Minería textual". En: *El profesional de la información*, enero-febrero, v. 13, n. 1, pp. 11-27.

García Marín, Javier & Calatrava, Adolfo & Luengo, Oscar. (2018). Debates electorales y conflicto. Un análisis con máquinas de soporte virtual (SVM) de la cobertura mediática de los debates en España desde 2008. *El Profesional de la Información*. 27. 10.3145/epi.2018.may.15.

García Marín, Javier & Calatrava, Adolfo. (2018). The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press. *Comunicación y Sociedad*. 31. 10.15581/003.31.3.175-188.

Goffman, Erving. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, USA: Harvard University Press.

Heine, Joerg-Henrik. (2020). *Rasch Model Parameters by Pairwise Algorithm*. *Package 'pairwise'*. CRAN.

Justicia de la Torre, María del Consuelo. (2017). *Nuevas técnicas de minería de textos: Aplicaciones*. [Doctoral dissertation, Universidad de Granada]. Universidad de Granada.
Katz, Elihu., & Lazarsfeld, Paul F. (1955). *Personal Influence. The Part Played by People in the Flow of Mass Communication*.

Lippmann, Walter. (2003). *La opinión pública*. -, España: Langre.

Luengo, Ó. G. (2016). Comunicación política: de la propaganda a las neurociencias. En Colino, César et. al. (comp.) Ciencia política: Una aventura Vital (pp.721-740). Valencia: Tirant Lo Blanch.

McCombs, Maxwell., & Shaw, Donald. (1972). *The agenda setting function of the media*, en Public Opinion Quarterly, vol. XXXVI, págs. 176-187.

Montemurro, Marcelo. A. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications, 300(3-4), 567-578. https://doi.org/10.1016/s0378-4371(01)00355-7

Reig, Ramón. (2009). Bases teóricas y documentales para el estudio de la Estructura de la Información y el análisis estructural de los mensajes. Estudios sobre el Mensaje Periodístico 2009, 15 385-407.

Swanson, Don. R. (1991). Complementary structures in disjoint science literatures. Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '91. https://doi.org/10.1145/122860.122889.

Wilkerson, John., & Casas, Andreu. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.

Anexos

Medio	n
ABC	78
El Mundo	45
El País	88

Tabla 2. "*ayuntamientos*"

Medio	n
ABC	19
El Mundo	12
El País	22

Tabla 3. "*diputaciones*"