ISSN (electrónico): 2014-2242 / www.cac.cat

QUADERNS DEL CAC

El reto de medir el sesgo ideológico en los medios escritos digitales

Ana S. Cardenal
Profesora agregada de la Universitat Oberta de Catalunya (UOC)

acardenal@uoc.edu
Código ORCID: orcid.org/0000-0002-1540-8004.

Carol Galais
Investigadora posdoctoral de la UOC

cgalais@uoc.edu
Código ORCID: orcid.org/0000-0003-2726-2193.

Joaquim Moré
Doctorado del Programa de Sociedad de la Información (UOC)

qimore@gmail.com
Código ORCID: orcid.org/0000-0001-5432-0657.

Camilo Cristancho
Investigador posdoctoral de la Universitat de Barcelona (UB)

camilo.cristancho@ub.edu
Código ORCID: orcid.org/0000-0003-1794-4457.

Silvia Majó-Vázquez
Investigadora posdoctoral del Reuters Institute for the Study of Journalism de la University of Oxford

silvia.majo-vazquez@politics.ox.ac.uk
Código ORCID: orcid.org/0000-0002-2312-7907.

Artículo recibido el 16/04/18 y aceptado el 19/06/18

Resumen
Este trabajo establece una propuesta para medir el sesgo ideológico de los medios digitales que se basa en el aprendizaje automatizado de contenidos. Utilizamos una estrategia sustentada en el uso de textos para identificar palabras cargadas ideológicamente, que estudios de ciencia política también emplean para medir las posiciones de los partidos y los candidatos. Nuestra propuesta presenta dos rasgos diferenciales respecto a estudios previos: usa el concepto frame como unidad de análisis para identificar el sesgo ideológico de los medios, y utiliza los tuiteos de los políticos en Twitter como texto de referencia para identificar grupos de palabras conectadas ideológicamente, i. e., los frames.

Palabras clave
Medios digitales, sesgo ideológico, aprendizaje automatizado, algoritmos, análisis de contenido.

Abstract
This paper makes a proposal to measure the ideological bias of digital media that is based on machine learning. We use a strategy based on the use of texts to identify ideologically charged words, which studies of political science also use to measure the positions of parties and candidates. Our proposal presents two differential features with respect to previous studies: it uses the concept of a frame as unit of analysis to identify ideological bias and it relies on the tweets of politicians as the reference text for identifying ideologically connected groups of word – i.e., frames.

Keywords
Digital media, media bias, machine learning, algorithms, content analysis.

1. Introducción. Por qué estudiar el sesgo de los medios digitales

En nuestro territorio, al igual que en todo Occidente, la esfera de los medios de comunicación digital está en ascendiente expansión. Solo en el Estado español, en 2015 se crearon 579 nuevos medios, la mayoría de ellos únicamente con versiones en línea (APM 2015). Esta creciente diversidad en la oferta de medios de comunicación dibuja un panorama fragmentado y significa un reto para los investigadores en comunicación política. Desconocemos cuál es el grado de pluralidad de nuestros medios digitales, es decir, su diversidad desde un punto de vista ideológico. Además, para saber cuál es el posible efecto de los medios en la opinión pública, es preciso conocer en primer lugar cuál es su inclinación política.

El grado de pluralidad del sistema mediático de un país constituye un criterio de valoración positiva de ese sistema de comunicación, según el Consejo de Europa (1994). Por lo tanto, identificar el sesgo ideológico de los múltiples medios digitales debe permitirnos, por un lado, evaluar la diversidad de

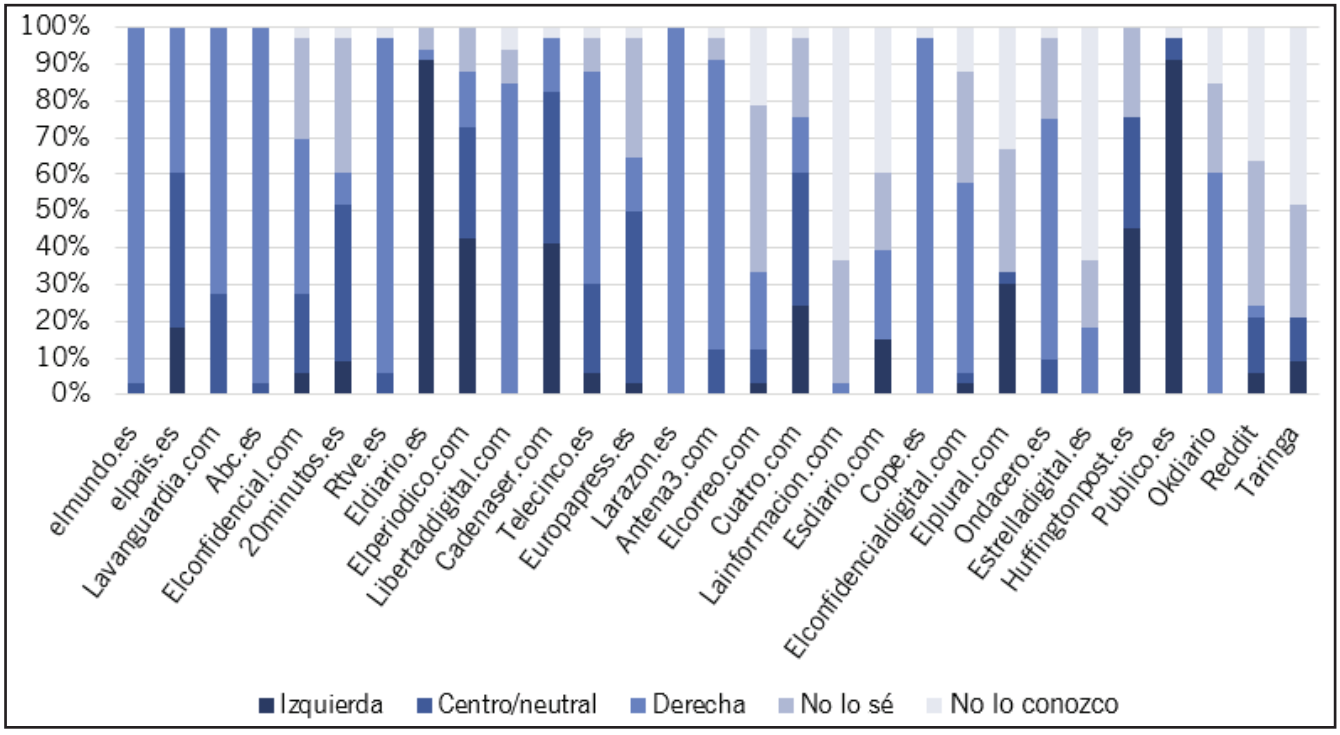
un sistema mediático y, en definitiva, su contribución al proceso democrático, y por el otro, responder a si ciertamente la creciente oferta de medios de comunicación conlleva que estos sean cada vez más partidistas y estén más polarizados (Stroud 2011). Además, proveer a la audiencia con información sobre el sesgo de los nuevos medios contribuiría a su alfabetización mediática (Buckingham 2007; Gilster 1997) y, de este modo, repercutiría positivamente en sus competencias cívicas, en la detección de noticias falsas y, por último, en un control más efectivo de los gobernantes.[1]

En cuanto a los efectos de los medios sobre la opinión pública, la investigación ha demostrado que su influencia está limitada por el sesgo de confirmación y la exposición selectiva, por los cuales los individuos buscan información coherente con aquello que creen previamente (Lazarsfeld, Berelson y Gaudet 1944; Nickerson 1998) y evitan exponerse a información contraria a sus actitudes o creencias, puesto que este contraste genera incomodidad (Festinger 1962; Olson y Stone 2014). Sin embargo, la multiplicación de la oferta informativa en línea dificulta que los usuarios se hagan una idea precisa del sesgo ideológico de cada nuevo medio digital y, por lo tanto, de la congruencia entre estos y sus propias actitudes. Así, los ciudadanos se estarían exponiendo ahora, en internet, a estímulos e ideas más diversas porque no pueden identificar el sesgo de todos los medios digitales existentes. Queda por saber cuál es el sentido de su influencia.

Dentro de las fronteras del Estado, solo algunos estudios han tratado este tema. Entre las excepciones más destacadas encontramos los trabajos de Almiron, quien ha analizado la estructura de propiedad y sus líneas editoriales para los medios tradicionales (2009) y para los diarios digitales sin referente impreso (2006). En una aproximación más reciente, la autora también se ha ocupado de la diversidad ideológica de estos diarios y ha analizado en qué términos se refieren a las ideologías más tradicionales, pero sin atribuir a cada medio un sesgo o una etiqueta ideológica concreta, sino representando el panorama conjunto que estos medios ofrecen (Pineda y Almiron 2013). Pero continuamos sin tener una brújula comúnmente aceptada a la que referirnos cuando hablamos de los sesgos ideológicos de los nuevos medios digitales.

Podemos intentar una primera aproximación al fenómeno de la ideología de los medios digitales analizando las percepciones de la ciudadanía. A través de tres encuestas distintas, nos hemos acercado a la percepción de la ciudadanía española sobre la ideología de algunos de los principales medios digitales estatales.[2] Lo más destacable es el porcentaje de individuos que no saben clasificar los medios. Así, entre el 23 y el 33% de las personas no sabe cuál es la ideología del Huffington Post o de 20 Minutos, a pesar de conocerlos. Casi un tercio de la población española no sabe cuál es la ideología de medios como eldiario.es o El Confidencial. Si preguntamos a estudiantes universitarios, casi la mitad no sabe ubicar eldiario.es, El Confidencial o el Huffington Post. Una estrategia alternativa consiste en preguntar a expertos. La figura 1 muestra los resultados de una encuesta realizada en septiembre de 2017 a 33 expertos en las áreas de ciencia política y ciencias de la información en España. Se les preguntó por la ideología de los 30 medios más visitados en el año anterior, según Alexa.

Figura 1. Percepción de distintos medios digitales. Encuesta a expertos. Septiembre de 2017. N=33



Fuente: Elaboración propia.

Si excluimos las versiones digitales de medios tradicionales como *El Mundo*, *ABC*, etc., encontramos un porcentaje sorprendentemente alto de "No lo sé" y "No lo conozco", que llega a ser de más del 50% para La Información (3,6% de la audiencia digital, según ComScore). Podemos concluir, pues, que situar estos medios en un mapa mental de ideologías resulta una tarea complicada, incluso para expertos en medios y política.

La presente búsqueda tiene por objeto clasificar los principales medios digitales en el territorio español en función de su sesgo ideológico, apostando por el análisis de contenido automatizado y, por lo tanto, eficiente y objetivo. Esta información será de utilidad no solo en el ámbito académico para los debates ya presentados sobre exposición selectiva, sino que también tendrá una importancia política vital para evaluar la pluralidad de los medios y mejorar el grado de alfabetización digital de la ciudadanía, lo que a su vez se considera positivo para la calidad democrática del sistema político.

2. Marco teórico. Medir el sesgo en los medios

2.1 Definiciones y conceptos básicos

El sesgo ideológico no implica un intento deshonesto ni deliberado de tergiversar la realidad, sino una forma de describir la realidad que es significativa y sistemáticamente distorsionada (Groeling 2013: 130). A su vez, la ideología se ha definido como la distorsión de una realidad objetiva que refleja construcciones mentales subjetivas y colectivas (Benabou 2008:1). Uno de los autores seminales en este debate, Converse, define la ideología como las partes (o los subconjuntos) de un sistema de creencias, como "una configuración de ideas y actitudes cuyos elementos están unidos por algún tipo de constricción o interdependencia funcional" (Converse 1964: 207).

La idea que propone Converse (1964) implica que, cuanta más dependencia funcional exista entre los elementos de un sistema de creencias, menos recursos cognitivos harán falta para ser descrito o comprendido. Desde este punto de vista, una de las dimensiones de juicio que más ha servido para simplificar los acontecimientos en la política ha sido la dimensión izquierda-derecha. Sobre esta dimensión se ubica a partidos, líderes, políticas y otros objetos de la política (Converse 1964: 214). La interdependencia entre los elementos que caracteriza un sistema de creencias también explicaría, según Converse, que la difusión social de las ideologías tienda a hacerse por "paquetes".[3] Esto afecta a la interpretación de las propias ideologías. Los partidos, por ejemplo, votan sobre diferentes temas de forma conectada (Benoit y Laver 2006, 2007) y presentan paquetes de alternativas a los electores (Downs 1957). Los electores emplean la dimensión izquierda-derecha para dar sentido a la decisión del voto y tomar decisiones sobre los paquetes de alternativas presentadas.

Los medios de comunicación también difunden las ideologías políticas a través de paquetes, en este caso de conjuntos de palabras o términos que evocan otros conceptos conectados ideológicamente. Con estas construcciones apelan a los diferentes sistemas de creencias y conceptos que los definen.

2.2 Limitaciones de los estudios previos sobre el sesgo de los medios

Estudios previos sobre el sesgo ideológico de los medios han usado básicamente dos aproximaciones para medirlo: la primera, basada en la caracterización de la audiencia, y la segunda, en el contenido publicado (véase también Budak *et al.* 2016). La primera aproximación ha utilizado el perfil ideológico de la audiencia de un medio para atribuirle una ideología. Por ejemplo, la literatura sobre exposición selectiva a la información (Freedman y Sears 1965) asume que la audiencia sigue medios afines ideológicamente. Así, conociendo la ideología de su audiencia se puede atribuir una ideología a los medios (Bakshy, Messing y Adamic 2015; Gentzkow y Shapiro 2011; Newman, Fletcher, Kalogeropoulos, Levy y Nielsen 2017; Barberá y Sood 2014).

Esta aproximación es parsimoniosa y relativamente simple. Aun así, la proliferación de medios dificulta cada vez más que la audiencia conozca el sesgo ideológico de los medios. Otro inconveniente es que proporciona medidas relativas y no objetivas de este sesgo. Si tenemos en cuenta que los movimientos de la audiencia pueden ser muy sensibles a pequeñas diferencias en el sesgo entre los medios, este método no nos permitiría valorar bien las diferencias existentes (Budak *et al.* 2016).

La segunda aproximación utilizada en la literatura para identificar el sesgo de los medios se basa en el contenido que estos elaboran. Pero la mayoría de los medios no toman posiciones explícitas sobre los temas que cubren, y esto constituye una dificultad (Barberá y Sood 2016). Ante esta limitación, los trabajos existentes han seguido tres grandes estrategias.

La primera consiste en limitar el análisis a un conjunto reducido pero altamente informativo del contenido publicado. Se trata del contenido editorial, que sí incorpora explícitamente el posicionamiento de los medios sobre los hechos de actualidad. Sin embargo, se ha criticado a los estudios que hacen uso de los editoriales, porque miden únicamente el sesgo de una parte muy pequeña del contenido, lo que puede exagerar el sesgo de la globalidad del diario (Barberá y Sood 2014).

La segunda estrategia se basa en el aprendizaje automatizado para detectar patrones (lingüísticos) en un conjunto amplio e indiscriminado de noticias. Se parte de la identificación de un conjunto de documentos (por ejemplo, programas de partidos) a partir de los cuales se detectan palabras ideológicamente cargadas. Posteriormente, se asigna una puntuación a cada una de estas palabras, se cuentan y se utilizan para estimar la ideología del medio (Gentzkow y Shapiro 2010; Wihbey, Coleman, Joseph y Lazer 2017). Aun así, las palabras cargadas ideológicamente representan un porcentaje todavía muy pequeño del contenido total publicado por los medios y, por lo

tanto, trabajar con este material produce un volumen elevado de ruido (Gentzkow y Shapiro 2010). Además, las palabras o frases asociadas a una ideología a menudo son utilizadas por medios de ideología opuesta en registros como el humor, la ironía o el sarcasmo para criticar a adversarios políticos. Claramente, este uso dificulta la clasificación de los medios (Barberá y Sood 2014: 4).

Finalmente, la tercera estrategia se basa en una combinación de aprendizaje automatizado y codificación humana (o *crowdsourcing*) para superar algunas de las limitaciones asociadas a la estrategia basada únicamente en el aprendizaje automatizado. La codificación humana permite identificar la ironía y la broma y corregir falsos positivos (Budak *et al.* 2016).

2.3 Una nueva dirección

En el presente trabajo apostamos por la segunda estrategia, totalmente basada en el uso del aprendizaje automatizado, para identificar o estimar la ideología de una muestra estratégica de medios. Sin embargo, nuestra propuesta presenta algunas novedades.

La primera es que aquí vamos algo más allá de los estudios previos, y no basamos nuestro análisis en palabras (o frases cortas) cargadas ideológicamente, sino en un conjunto de sintagmas nominales conectados. De este modo nos aseguramos de que los términos por los que empezamos tengan significado por sí mismos. La segunda novedad es que no nos centraremos tanto en una lista de términos propios de la derecha o la izquierda sino en los discursos en los que aparecen (*frames*). La tercera es que usaremos tuiteos en Twitter de políticos como texto de referencia para identificar la ideología en vez de programas electorales o los discursos parlamentarios.

Algunos estudios utilizan las cuentas de Twitter de los usuarios de los medios para deducir su ideología y, en última instancia, atribuirla a los medios (Barberá y Sood 2014), pero ningún estudio, que sepamos, ha utilizado las cuentas de Twitter de políticos para detectar qué términos y discursos son los típicos de una ideología. Creemos que puede ser una estrategia eficiente, porque internet ha contribuido a la polarización de los debates en línea. Así, en Twitter se emplearía un lenguaje con mayor carga ideológica que en otros medios (Toff y Kim 2013), aunque bastante parecida a la de los diarios digitales (Mullainathan y Shleifer 2005).. En segundo lugar, conceptualizaciones recientes de los partidos políticos los presentan como coaliciones laxas compuestas por actores que comparten una agenda y unos objetivos comunes (Bawn *et al*. 2012). En estas redes, el uso de las palabras por parte de los profesionales de la comunicación para la construcción de un relato cobra importancia (Toff y Kim 2013). El contexto o escenario en el que esta coalición de intereses que son los partidos pondría a prueba este lenguaje no serían los programas electorales, que poca gente lee y son bastante neutros, sino las redes sociales: un espacio mucho más dinámico y en fase de expansión (Newman *et al*. 2017).

3. Metodología

Para clasificar los medios digitales según su ideología, hemos seguido tres fases que a continuación veremos en detalle.

3.1 Fase 1: Identificación del corpus para detectar discursos ideologizados

A la hora de elegir el corpus de referencia para identificar contenidos ideológicos, nos inclinamos por los tuiteos de los políticos en Twitter, puesto que es una herramienta que se caracteriza por su inmediatez, brevedad y coloquialismo, y esto permite utilizar concepciones y recursos retóricos parecidos a los titulares de los diarios.[4] En concreto, elegimos como corpus de referencia las cuentas de Twitter de 296 parlamentarios españoles en la XII legislatura.[5]

Para explotar el máximo nivel de contraste, y optimizar la tarea de atribución de ideología a los diputados, en esta búsqueda nos hemos limitado a los dos partidos con una ideología más extrema y clara en el eje izquierda/derecha: la coalición Unidos Podemos (o simplemente, Podemos) y Partido Popular (PP), respectivamente. Estos son los dos partidos políticos de ámbito estatal (PAES) con representación parlamentaria que los españoles sitúan más en los extremos del eje izquierda/derecha (fuente: 8.ª oleada del panel DEC/UAB, diciembre de 2015).

El conjunto de datos analizados consiste en casi medio millón de tuiteos de los diputados de Podemos y PP en el Congreso de los Diputados.[6] La distribución del número de tuiteos por partido se presenta en la tabla 1.

3.2 Fase 2: Identificación de las relaciones semánticas que son características de un discurso ideológico (*frames*)

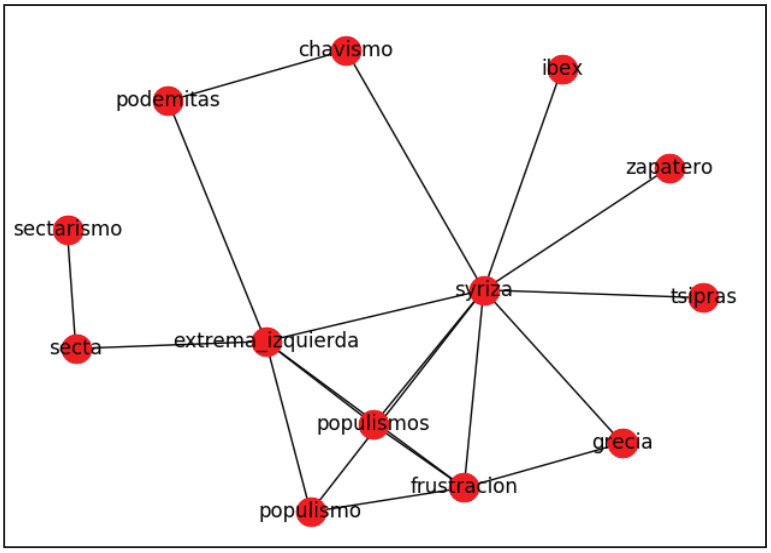
Metodológicamente, un *frame* es una relación de proximidad semántica entre un término *IT* (*ideology term*, que también podemos entender como palabra clave) del discurso y unos términos *t* del propio discurso.[7] La conjunción de un término *IT*

Tabla 1. Distribución en el tiempo de los tuiteos de los diputados españoles en la XII legislatura desde el inicio de su actividad en la red social

Año	Usuarios PP	Usuarios Podemos	Tuiteos PP	Tuiteos Podemos
2009	6	7	1.214	270
2010	15	10	2.993	1.492
2011	35	17	21.324	7.377
2012	38	19	48.498	20.362
2013	50	25	77.700	27.010
2014	60	32	94.667	35.147
2015	76	48	166.789	77.927
2016	88	56	203.838	156.512
2017	102	62	173.722	298.474

Fuente: Elaboración propia.

Figura 2. Representación gráfica del *frame* surgido del *IT* "populismos" para el PP



Fuente: Elaboración propia.

con una serie de términos *t* indica una determinada visión de las cosas por parte del término *IT*. Así, para el PP el *IT* "populismos" lleva asociados los términos *t* "populismo, frustración, Syriza, extrema_izquierda, Grecia". A su vez, el término *IT* "Syriza" presenta asociados los términos "Zapatero, frustración, Grecia". Por lo tanto, durante el periodo en el que se publicaron los tuiteos, los representantes del PP relacionaban Grecia con el populismo y la frustración, etc. La figura 2 representa una red que relaciona los términos en torno al *IT* "populismos". Así, no nos sorprendería encontrar un tuiteo o el titular de un editorial que dijera que Zapatero es un populista y que ha sido el "Tsipras" de España. El tuiteo o titular concentra una tesis, un mensaje y unos valores del partido expresados con unos determinados términos que conforman un discurso, que es lo que recogen los *frames*.

Estas relaciones evocan el concepto de *frame* de Lakoff (2004), en el que los conceptos tienen una estructura. Por ejemplo, la palabra "elefante" es un *frame* que evoca la imagen de un elefante y todo lo que conocemos sobre los elefantes. De forma parecida, nuestros *frames* quieren capturar la estructura de relaciones que una sola palabra como "populismo" o "Grecia" tiene en el discurso de un partido político o de un grupo de una determinada ideología.

Para detectar los *frames*, primero identificamos los sintagmas nominales de los tuiteos de los representantes de una determinada ideología. Para esta tarea hemos usado la herramienta Parse Tree del paquete pattern.es del proyecto CLiPS.[8] Una vez obtenidos los sintagmas nominales, se buscan sus términos *t*; es decir, los términos semánticamente más cercanos en el conjunto de todos los tuiteos. Para obtenerlos aplicamos el método Word2vec[9] mediante un módulo de Python, el cual indica que dos sintagmas nominales *p* y *p'* son cercanos si aparecen en contextos similares.[10]

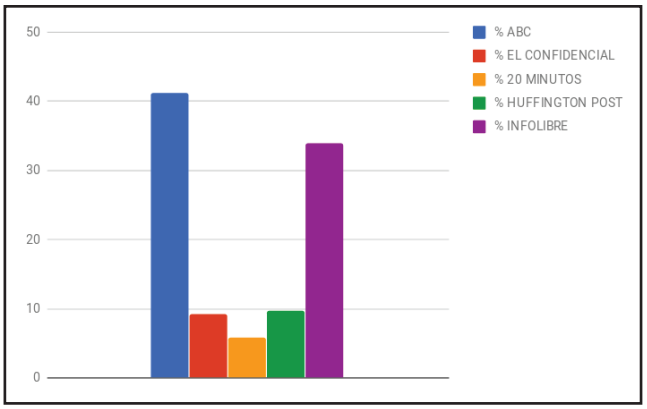
Es decir, las palabras que suelen estar alrededor de *p* también suelen estar alrededor de *p'*. Aplicado a la detección de los términos *t*, la explicación de que "populismos" y "extrema_izquierda" son cercanos es que las palabras que rodean a "populismos" suelen aparecer también cerca de "extrema_izquierda".

Seguidamente, establecimos criterios para identificar cuáles de entre todos los sintagmas nominales son *IT*. En primer lugar, el sintagma nominal tiene que aparecer tanto en los tuiteos del PP como en los de Podemos. Sin esta condición no podemos decidir si hay discrepancia en los *frames* entre los dos partidos (dado que solo uno lo usa). En segundo lugar, el *IT* tiene que aparecer con más frecuencia en los tuiteos de un partido que en los del otro. Consideramos, aquí, que un criterio razonable es que un término "propio" de un partido tiene que aparecer en los tuiteos de sus diputados más del doble de veces que en el corpus de referencia del otro partido. En tercer lugar, los *frames* de los partidos (esto es, los términos *t* asociados al *IT*) deben ser diferentes. Es decir, el vector que se genera con los tuiteos de un partido debe tener una distancia considerable respecto al vector para el mismo término generado con los tuiteos del partido opuesto. Una vez que se crean los vectores de los sintagmas nominales del PP y de Podemos, se calcula la distancia (*cosine similarity*) para cada vector. Nos quedaremos como candidatos a *IT* los que tengan una *cosine similarity* inferior a 0.1, apuntando, por lo tanto, a una gran diferencia.

3.3 Fase 3. Comprobación de las correspondencias entre los *frames* de un discurso político de una determinada ideología y las noticias de los diarios

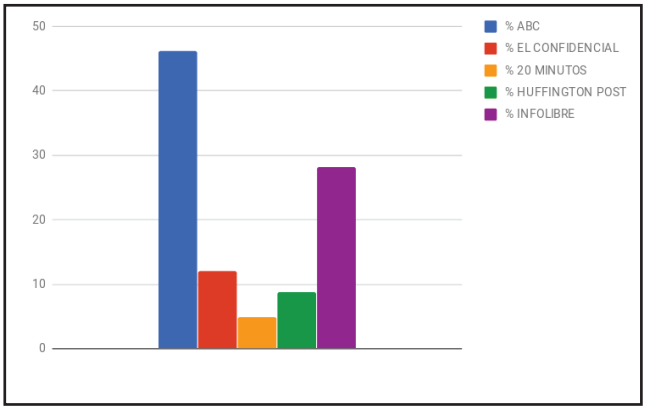
A la hora de aplicar el método, hemos decidido centrarnos en algunos de los medios que se ha detectado (véase la introducción) que generaban mayor confusión en la audiencia: el Huffington Post, El Confidencial, infoLibre y 20 Minutos. Además, hemos incluido el ABC por ser el medio más claramente situado a la derecha en todas las encuestas analizadas, lo que puede servirnos como punto de referencia.

Figura 3. Porcentaje de *IT-PP* para los diarios analizados



Fuente: Elaboración propia.

Figura 4. Porcentaje de *IT-Podemos* para los diarios analizados



Fuente: Elaboración propia.

Hemos obtenido los textos de la base de datos de prensa FACTIVA, y hemos acotado la búsqueda entre principios de diciembre de 2016 (precampaña elecciones generales 2016) y finales de junio de 2017 (elecciones del 26 de junio de 2017 e inicio de la XII legislatura).

Para realizar la comprobación de las correspondencias, hemos considerado distintas opciones:

- Contar la frecuencia de los *IT* de una determinada ideología en cada diario. Así, un diario más afín al PP usará más *IT-PP* que un diario con una línea ideológica de izquierdas.
- Determinar si los vectores que describen los *IT* en los tuiteos y los vectores que describen los *frames* de estos *IT* en los diarios son parecidos.
- Centrarnos en el número de términos *t* que acompañan a un *IT* para cada partido que aparecen en los distintos diarios.

En el siguiente apartado explicamos los resultados obtenidos por los diferentes métodos y las posibilidades de mejora de estos.

4. Resultados

4.1 *IT* característicos del PP y de Podemos

Hemos obtenido 327 *IT* característicos del PP (*IT-PP*) y 113 de Podemos (*IT-Podemos*). Son, pues, sintagmas nominales presentes en los discursos del partido opuesto, con una frecuencia superior al doble que en los tuiteos del partido ideológicamente opuesto, y con un vector de t_s que tiene una distancia (*cosine similarity*) inferior a 0.1 respecto al vector del mismo sintagma nominal generado con los tuiteos del partido opuesto (es decir, generan marcos interpretativos muy distintos).

Por ejemplo, tanto el PP como Podemos hablan del "proceso independentista", pero el PP habla sobre esta idea más del doble de veces que Podemos. Los términos *t* con los que se refieren a ella son extremadamente distintos (valor de la *cosine distance* entre el vector "proceso independentista" del PP respecto al vector generado para el mismo término *IT* de Podemos = 0.0978). Por lo tanto, este *IT* es divisorio: tiene un *frame* del PP (derecha) y un *frame* de Podemos (izquierda), a pesar de ser más característico del PP. Ahora bien, llama la atención la presencia de *IT* como "populismo", "proetarras" o "coleta" entre los *IT* propios de Podemos, puesto que son términos que la derecha utiliza para desacreditarlos. Esto apunta a que los tuiteos de Podemos tienen una gran carga de referencialidad del discurso del partido ideológicamente contrario.

4.2 Correspondencia entre tuiteos y diarios según la frecuencia de los *IT*

La primera opción para comprobar la correspondencia entre los tuiteos y los diarios fue verificar la frecuencia de los *IT* de una determinada ideología en los diarios. Así, un diario afín al PP usará más *IT-PP* que otro diario.

En la figura 3 vemos el porcentaje de *IT-PP* distribuidos por diarios. Algo más del 40% de las apariciones de *IT-PP* se producen en el ABC. Lo siguen infoLibre y El Confidencial. Así, el diario más afín al PP sería el ABC, mientras que 20 Minutos sería el más alejado. Ahora bien, ¿qué pasa si observamos la correspondencia entre los *IT-Podemos* y los mismos diarios?

En la figura 4 observamos que el ABC también es el diario en el que se concentran más *IT-Podemos*, aunque menos acusadamente que en el ejemplo anterior. La distribución relativa del resto de los diarios es muy similar al ejemplo anterior. Estos resultados se alejan demasiado del criterio de los ciudadanos y de los expertos como para fiarnos de ellos. Por lo tanto, no parece que la distribución por frecuencia de los *IT* según la ideología sirva para detectar alineamientos claros entre los tuiteos de los políticos y los diarios. La apropiación por parte de Podemos de *frames* derivados de *IT* originariamente de derechas (y más presentes en los diarios presumiblemente más de derechas) podría estar detrás de estos resultados tan contraintuitivos.

4.3 Correspondencia entre tuiteos y diarios según la similitud de frames

El siguiente paso fue comprobar si los vectores que describen los *IT* en los tuiteos y los vectores que describen los *frames* de estos *IT* en los diarios son parecidos. Por ejemplo, queríamos

Tabla 2. Vectores de coocurrencia de términos *t* para una serie de *IT* del PP en los diarios analizados

IT	ABC	El Confidencial	20 Minutos	Huffington Post	InfoLibre
Centralidad	19	17	0	13	0
Abismo	8	3	0	0	0
Coleta	1	0	0	2	0
Populismo	2	1	0	0	1

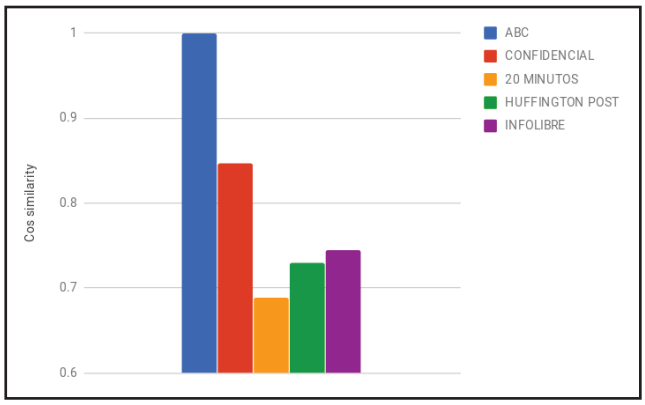
Fuente: Elaboración propia.

Tabla 3. Vector de coocurrencia de los tres términos *t* asociados al *IT* de Podemos "proetarras"

IT	ABC	El Confidencial	20 Minutos	Huffington Post	InfoLibre
Proetarras	3	0	0	0	0

Fuente: Elaboración propia.

Figura 5. Proximidad de los diarios respecto al ABC en cuanto a los *frames* del PP



Fuente: Elaboración propia.

comprobar si los diarios afines a Podemos suelen vincular más a menudo la UE a la austeridad y a Angela Merkel que los diarios afines al PP.

Igual que habíamos hecho con los tuiteos de los diputados, convertimos los sintagmas de cada diario en vectores, cuyas dimensiones eran los términos *t*; es decir, los términos más relacionados semánticamente, obtenidos con Word2vec. Los vectores de los *IT-PP* e *IT-Podemos* se compararon –vía *cosine similarity*– con los vectores de los mismos sintagmas nominales de los diarios. Comprobamos que la referencialidad a los *IT* del partido ideológicamente opuesto también era una característica de los diarios, por lo cual obtuvimos resultados parecidos a los de la frecuencia de *IT*.

4.4 Correspondencia entre tuiteos y diarios según los focos en los *t*

La última opción explorada se centraba en los términos *t* y su capacidad de relacionarse con *IT* de ideología diferente. En términos de *frames* significa que, dado un *IT*, los diarios afines a un partido coinciden al hablar de los mismos *t*.

Para comprobarlo, recogimos los términos *t* relacionados semánticamente con los *IT* de los tuiteos del PP y de Podemos. Después comprobamos cuántos *t* propios de cada partido aparecían en las noticias de un diario y creamos, para cada *IT*, un vector con el número de *t* del PP y de Podemos coocurrentes para cada diario.[11] La tabla 2 ilustra estos vectores con los t_{pp} relacionados con "centralidad", "abismo", "coleta" y "populismo". Por ejemplo, "centralidad" y "coleta" tienen 19 y 1 t_{pp} coocurrentes en el diario ABC, respectivamente, pero ningún t_{pp} en infoLibre. "Populismo" tiene 2 t_{pp} en el ABC y 1 en El Confidencial, pero ninguno en 20 Minutos ni en el Huffington Post.

Creados los vectores para cada diario, tomamos el diario donde más aparecen los *frames* del PP como referencia: el ABC. La incidencia de términos *t* del PP en el resto de los diarios se representa en relación con este diario, que toma el valor 1.

Como puede verse en la figura 5, El Confidencial es el diario más cercano al ABC si tenemos en cuenta la frecuencia de aparición de los *IT* con los *t* del PP. 20 Minutos, Huffington Post e infoLibre están más alejados, siendo 20 Minutos el que más lo está. Con este sistema –vectores de coocurrencia en los diarios– podríamos, en principio, encontrar "falsos" *IT* de izquierdas. Por ejemplo, la tabla 3 representa el vector correspondiente a "proetarras" –un *IT*, recordémoslo, más frecuentemente utilizado en el discurso de Podemos que en el del PP–.

"Proetarras" tiene 3 términos *t* con los que coaparece en un solo diario, que es el ABC. Teniendo en cuenta que estos *t* son "Otegui", "Bildu" y "ETA", habría que reflexionar sobre si la coaparición de un *IT* con unos *t* determinados en un diario ya alineado ideológicamente (como hemos hecho con el ABC) es un criterio para (re)clasificar –ideológicamente– un *IT*, a pesar de que sea muy utilizado por el partido de ideología contraria. En cualquier caso, este procedimiento podría servir para hacer "limpieza" de *IT* clasificados erróneamente como de izquierdas o de derechas, y parece ser una posible solución al problema de la apropiación de *frames* por parte del partido contrario como herramienta para avivar el conflicto, señalar paradojas en los contrarios, etc.

5. Conclusiones

Medir el sesgo de los medios digitales escritos es necesario porque necesitamos saber cuál es el alcance y el sentido de su efecto, a fin de evaluar la pluralidad del panorama informativo y mejorar el grado de alfabetización digital de la ciudadanía.

La revisión de la literatura existente sobre la medida del sesgo ideológico de los medios ha revelado que los diferentes medios empleados hasta la actualidad presentan una serie de limitaciones. Atribuir a cada medio la ideología de su audiencia asume que los ciudadanos conocen el sesgo de los medios y se exponen a ellos selectivamente; pero ni el primer supuesto ni el segundo son siempre verdad. La segunda aproximación utiliza el contenido publicado, siguiendo tres posibles variantes. La primera es limitarse a una pequeña cantidad de texto muy indicativa del contenido (editoriales), la segunda consiste en detectar automáticamente patrones lingüísticos, y la última, en combinar estos procedimientos automatizados con codificación humana. La utilización de editoriales tiende a presentar una ideología más extremista de la que tiene realmente el medio, y la última estrategia es muy costosa en términos de recursos. Hemos adoptado, pues, la segunda.

Ahora bien, nuestra perspectiva incluye tres novedades. Primero, nuestra unidad de análisis no es una lista de palabras o frases cargadas ideológicamente, sino un conjunto de sintagmas nominales conectados y cargados ideológicamente. Segundo, la medida utilizada para asignar una ideología no se establece únicamente a partir de la frecuencia de uso de dichas cadenas de palabras sino, sobre todo, a partir de la discrepancia entre ellas. El último aspecto innovador radica en el cuerpo de texto que usamos como referente para identificar *frames* ideológicos: utilizamos tuiteos de líderes políticos en Twitter, y no programas electorales o discursos parlamentarios.

Para identificar contenido con carga ideológica, nos hemos centrado en los *frames* (conjuntos de palabras semánticamente cercanas alrededor de un término *IT*) propios de los dos partidos de ámbito estatal más polarizados de acuerdo con las percepciones de la opinión pública española: PP y Podemos. Hemos detectado una serie de términos comunes a los dos partidos, pero más presentes en los tuiteos de los diputados de un partido que en los del otro. Hemos verificado que los términos *t* de los que se acompañan sean bastante diferentes antes de identificar los *frames*.

Durante este proceso nos hemos encontrado con diferentes vías muertas. Una de ellas ha sido contar las correspondencias de los *frames* de cada partido en los diarios, debido probablemente a la apropiación por parte de Podemos de *frames* críticos con ellos surgidos desde la derecha. De forma similar, comparar la distancia entre *frames* de partidos y de diarios nos lleva al mismo punto: los resultados parecen tener sentido solo si atendemos a los *frames* propios del PP y a las similitudes entre este partido y los medios, pero esto no se aplica a Podemos.

Próximos desarrollos deberán intentar solucionar el problema de las referencias irónicas a los marcos de interpretación del adversario. Este se ha apuntado anteriormente como uno de los principales problemas del análisis del contenido a través del aprendizaje automatizado para atribuir una ideología a los medios (Barberá y Sood 2014). Nuestros datos confirman que Podemos referencia las críticas surgidas desde la derecha a sus actitudes y argumentaciones "populistas", haciendo burla de ellas, lo que imposibilita identificar de forma automática su intencionalidad. Otra posibilidad sería incluir una dimensión temporal para dar mayor peso a los términos que aparezcan primero en el tiempo como elementos identificadores del *frame* de un partido. Por otra parte, se podría diluir este error de medida ampliando el corpus de referencia al resto de los partidos de ámbito estatal. Así, este fenómeno típico de Podemos quedaría diluido entre los tuiteos del PSOE. Por último, se podría combinar el aprendizaje automatizado con la codificación humana. Esta estrategia, a pesar de ser más costosa, nos permitiría descartar términos utilizados con ironía o sarcasmo.

Notas

1. La alfabetización mediática es el desarrollo de una comprensión razonada y crítica de la naturaleza de los medios de comunicación y sus efectos, de cómo crean significado y de cómo organizan su propia realidad (Gilster 1997; Aparici 1996).
2. Estas encuestas se realizaron entre 2015 y 2016. La primera es una encuesta del grupo de investigación eGovernança: Administración y Democracia Electrónica (GADE) de la Universitat Oberta de Catalunya (UOC), realizada para el proyecto Opinionet. La segunda es una encuesta del grupo de investigación Democracia, Elecciones y Ciudadanía (DEC) de la Universitat Autònoma de Barcelona (UAB). La tercera también es una encuesta del grupo GADE a la que han respondido los estudiantes de la UOC.
3. Este formato en paquetes se corresponde casi perfectamente con la noción de *frames* o marcos de interpretación propio de los análisis semánticos.
4. Tras varias pruebas exploratorias, se han desestimado los discursos parlamentarios porque no era posible construir un corpus de texto lo bastante grande como para extraer términos o conjuntos de términos cargados ideológicamente. En este sentido, también se ha optado por prescindir de los programas electorales, porque en los análisis preliminares efectuados no se han detectado discrepancias significativas en los *frames* de los diferentes partidos a partir de los programas electorales. Además, los programas electorales (y las codificaciones propuestas por el proyecto Party Manifiesto) ya no se usan para estimar las posicionas ideológicas de los partidos (Benoit y Laver 2006, 2007). Por último, los programas de los partidos utilizan un lenguaje muy formal que se aleja del lenguaje más informal y con carga ideológica que sí se usa en los medios.

5. De entre los 350 diputados, solo 296 tienen una cuenta activa de Twitter.

6. Algunas personas, sobre todo del PP, eran diputadas en 2009; pero dentro de la coalición Unidos Podemos, solo algunas personas pertenecientes a Izquierda Unida lo eran antes de 2016. Sin embargo, entendemos que con sus tuiteos antes de esta fecha están difundiendo mensajes y valores en consonancia con este partido.

7. Entendemos aquí *proximidad semántica* como coocurrencia, o aparecer en posiciones adyacentes en el mismo texto. Se trata de un concepto propio del análisis cuantitativo de textos. El algoritmo utilizado para determinarla (Word2vec) recoge esta proximidad física de las palabras manteniendo las propiedades gramaticales de los textos de los que se extraen.

8. <https://www.clips.uantwerpen.be/pages/pattern-es>. El sintagma nominal es –junto con los verbos– el elemento básico que estructura una oración, la principal sede del significado léxico y, en definitiva, la manera en que se denominan los conceptos. Así, podemos recoger denominaciones como "Tribunal Superior de Justicia", en vez del bigrama "Tribunal Superior" o de los unigramas "Tribunal", "Superior" y "Justicia".

9. Word2vec es un método representativo de la tendencia más reciente en aprendizaje automático que se llama *deep learning*, con una estructura de redes neuronales (Dikolov *et al.* 2013). Es un método que se está aplicando con mucho éxito a la traducción automática (Mikolov, Quoc y Sutskever 2013), al análisis del sentimiento (Acosta *et al.* 2017) y a la clasificación de documentos (Lilleberg, Zhu y Zhang 2015). Incluso la abstracción de la idea de contexto, definido en un espacio vectorial, ha fomentado la aparición de otras aplicaciones como los recomendadores (Ozsoy 2016).

10. Word2vec utiliza un algoritmo que calcula, por cada sintagma nominal, los sintagmas nominales más cercanos. La proximidad es un valor que va del 0 al 1 (de menos cercano a más cercano). Para este proyecto hemos considerado como términos *t* los que superan el valor de la mediana (0.5).

11. La métrica utilizada para medir la coocurrencia ha sido la Normalized Google Distance (NGD), con un rango de valores entre el 0 (ninguna proximidad) y el 1 (máxima proximidad). Es una medida de distancia semántica según el grado de coaparición de dos términos, en nuestro caso, entre el *IT* y su *t*, en el titular y en el cuerpo de la noticia.

Referencias

Acosta, J.; Lamaute, N.; Luo, M.; Finkelstein, E.; Cotoranu, A. *Proceedings of Student-Faculty Research Day*. CSIS, Pace University, 5 de mayo de 2017.

Almiron, N. "Pluralismo en Internet: el caso de los diarios digitales españoles de información general sin referente impreso". Ámbitos, 15 (2006).

Almiron, N. "Grupos privados propietarios de medios de comunicación en España: principales datos estructurales y financieros". *Comunicación y Sociedad*, 22 (2009), 1.

Aparici, R. *La revolución de los medios audiovisuales: educación y nuevas tecnologías*. Madrid: Ediciones de la Torre, 1996. ISBN: 84-7960-132-9.

Asociación de Periodistas de Madrid. *Informe anual de la profesión periodística*. Madrid: APM, 2015.

Bakshy, E.; Messing, S.; Adamic, L. A. "Exposure to ideologically diverse news and opinion on Facebook". *Science*, 348 (2015), 6239, 1130-1132.

Barberá, P.; Sood, G. "Follow Your Ideology: A Measure of Ideological Location of Media Sources". Manuscrito no publicado, 2016.

Bawn, K.; Cohen, M.; Karol, D.; Masket, S.; Noel, H.; Zaller, J. "A theory of political parties: Groups, policy demands and nominations in American politics". *Perspectives on Politics*, 10 (2012), 3, 571-597.

Benabou, R. "Ideology". *NBER Working Paper Series*, 13907 (2008).

Benoit, K.; Laver, M. *Party Policy in Modern Democracies*. Londres: Routledge, 2006. ISBN: 978-0415499798.

Benoit, K.; Laver, M. "Estimating party policy positions: Comparing expert surveys and hand-coded content analysis". *Electoral Studies*, 26 (2007), 1, 90-107.

Budak, C., Goel, S., & Rao, J. M. "Fair and balanced? Quantifying media bias through crowdsourced content analysis". *Public Opinion Quarterly*, 80 (2016), 1, 250-271.

Consejo de Europa. "4ème Conférence ministérielle Européenne sur la politique des communications de masse. Les média dans une société démocratique". Praga, 7-8 de diciembre. *Rapport d'activité du Comité d'experts sur les concentrations des media et le pluralisme*. MCM (94)5. Estrasburgo: Consejo de Europa, 1994, p. 8.

Converse, P. E. "The nature of mass opinion beliefs". En: Apter. D. (ed.). *Ideology and Discontent*. Nueva York: The Free Press of Glencoe, 1964. ISBN: 9780029007600.

Downs, A. *An Economic Theory of Democracy*. Nueva York: Harper and Row, 1957. ISBN: 9780060417505.

Festinger, L. *A Theory of Cognitive Dissonance*. Vol. 2. California: Stanford University Press, 1962. ISBN: 9780804701310.

Freedman, J. L.; Sears, D. O. "Selective exposure". En: Berkowitz L. (ed.). *Advances in Experimental Social Psychology*. Vol. 2. Nueva York: Academic Press, 1965, p. 58-97.

Gentzkow, M.; Shapiro, J. M. "What drives media slant? Evidence from US daily newspapers". *Econometrica*, 78 (2010), 1, 35-71.

Gentzkow, M.; Shapiro, J. M. "Ideological segregation online and offline". *The Quarterly Journal of Economics*, 126 (2011), 4, 1799-1839.

Gilster, P. *Digital Literacy*. Nova Jersey: John Wiley & Sons, 1997.

Groeling, T. "Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news". *Annual Review of Political Science*, 16 (2013).

Iyengar, S.; Hahn, K. S. "Red media, blue media: Evidence of ideological selectivity in media use". *Journal of Communication*, 59 (2009), 1, 19-39.

Kalogeropoulos, A.; Newman, N. (2017). "'I saw the news on Facebook': Brand attribution when accessing news from distributed environments". *Digital News Project 2017*. Oxford: Reuters Institute for the Study of Journalism, University of Oxford, 2017. <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-07/Brand%20attributions%20report.pdf>.

Lakoff, G. *Don't Think of an Elephant: Know your Values and Frame the Debate*. Vermont [Estados Unidos]: Chelsea Green Publishing, 2004.

Lazarsfeld, P. F.; Berelson, B.; Gaudet, H. *The People's Choice: How the Voter Makes up his Mind in a Presidential Election*. Nova York: Duell, Sloan and Pearce, 1944. ISBN: 978-0231085830.

Lilleberg, J., Zhu, Y., Zhang, Y. "Support vector machines and Word2vec for text classification with semantic features". IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing, julio de 2015.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. "Efficient estimation of word representations in vector space". En: *Proceedings of Workshop at ICLR, 2013.*

Mikolov, T.; Le Quoc, V.; Sutskever, I. "Exploiting similarities among languages in machine translation". arXiv preprint arXiv:1309.4168, 2013.

Mullainathan, S.; Shleifer, A. "The market for news". *American Economic Review*, 95(1), (2005), 1031-1053

Newman, N.; Fletcher, R.; Kalogeropoulos, A.; Levy, D. A.; Nielsen, R. K. *Digital News Report 2017*. Oxford: Reuters Institute for the Study of Journalism, University of Oxford, 2017. <http://www.digitalnewsreport.org/>.

Nickerson, R. S. "Confirmation bias: A ubiquitous phenomenon in many guises". *Review of general psychology*, 2 (1998), 2, 175.

Olson, J. M.; Stone, J. "The influence of behavior". *The handbook of attitudes*, 223 (2014).

Gulcin Ozsoy, M. "From word embeddings to item recommendation". arXiv preprint arXiv:1601.01356, 2016.

Pineda, A.; Almiron, N. "Ideology, politics, and opinion journalism: A content analysis of Spanish online-only newspapers". *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 11 (2013), 2, 558-574.

Stroud, N. J. *Niche News: The Politics of News Choice*. Oxford: Oxford University Press on Demand, 2011. ISBN: 9780199755509.

Toff, B. J.; Kim, Y. M. "Words That Matter: Twitter and Partisan Polarization". UW Madison's Political Behavior Research Group meeting. Madison, Wisconsin, 13 de noviembre de 2013.

Wihbey, J.; Coleman, T. D.; Joseph, K.; Lazer, D. "Exploring the Ideological Nature of Journalists' Social Networks on Twitter and Associations with News Story Content". *DS+J*, 2017. <https://arxiv.org/pdf/1708.0627.pdf>.