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

QUADERNS DEL CAC

El repte de mesurar el biaix ideològic en els mitjans escrits digitals

Ana S. Cardenal
Professora agregada de la Universitat Oberta de Catalunya (UOC)
acardenal@uoc.edu
Codi ORCID: orcid.org/0000-0002-1540-8004.

Carol Galais
Investigadora postdoctoral de la UOC
cgalais@uoc.edu
Codi ORCID: orcid.org/0000-0003-2726-2193.

Joaquim Moré
Doctorat del Programa de Societat de la Informació de la UOC
qimore@gmail.com
Codi ORCID: orcid.org/0000-0001-5432-0657.

Camilo Cristancho
Investigador postdoctoral de la Universitat de Barcelona (UB)
camilo.cristancho@ub.edu
Codi ORCID: orcid.org/0000-0003-1794-4457.

Silvia Majó-Vázquez
Investigadora postdoctoral del Reuters Institute for the Study of Journalism de la University of Oxford
silvia.majo-vazquez@politics.ox.ac.uk
Codi ORCID: orcid.org/0000-0002-2312-7907.

Article rebut el 16/04/18 i acceptat el 19/06/18

Resum
Aquest treball fa una proposta per mesurar el biaix ideològic dels mitjans digitals que es basa en l'aprenentatge automatitzat de continguts. Fem servir una estratègia sustentada en l'ús de textos per identificar paraules carregades ideològicament, que estudis de ciència política també utilitzen per mesurar les posicions dels partits i els candidats. La nostra proposta presenta dos trets diferencials respecte a estudis previs: fa servir el concepte de frame *com a unitat d'anàlisi per identificar el biaix ideològic dels mitjans, i utilitza les piulades dels polítics a Twitter com a text de referència per identificar grups de paraules connectades ideològicament, i. e., els* frames.

Paraules clau
Mitjans digitals, biaix ideològic, aprenentatge automatitzat, algoritmes, anàlisi de contingut.

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. Introducció. Per què estudiar el biaix dels mitjans digitals

Al nostre territori, com arreu a Occident, l'esfera dels mitjans de comunicació digital està en expansió ascendent. Només a l'Estat espanyol, l'any 2015 es van crear 579 nous mitjans, la majoria dels quals només amb versions en línia (APM 2015). Aquesta diversitat creixent en l'oferta de mitjans de comunicació dibuixa un panorama fragmentat i representa un repte per als investigadors en comunicació política. Desconeixem quin és el grau de pluralitat dels nostres mitjans digitals, és a dir, la seva diversitat des d'un punt de vista ideològic. A més, per saber quin és el possible efecte dels mitjans en l'opinió pública, és necessari conèixer en primer lloc quina és la seva inclinació política.

El grau de pluralitat del sistema mediàtic d'un país constitueix un criteri de valoració positiva d'aquell sistema de comunicació, segons el Consell d'Europa (1994). Per tant, identificar el biaix ideològic dels múltiples mitjans digitals ens ha de permetre, d'una banda, avaluar la diversitat d'un sistema mediàtic i, en definitiva, la seva contribució al procés democràtic, i de l'altra, esbrinar si certament l'oferta creixent de mitjans de comunicació comporta que aquests siguin cada vegada més partidistes i polaritzats (Stroud 2011). A més, proveir l'audiència

d'informació sobre el biaix dels nous mitjans contribuiria a la seva alfabetització mediàtica (Buckingham 2007; Gilster 1997) i, per tant, repercutiria positivament en les seves competències cíviques, en la detecció de notícies falses i, finalment, en un control més efectiu dels governants.[1]

Pel que fa als efectes dels mitjans sobre l'opinió pública, la recerca ha demostrat que la seva influència està limitada pel biaix de confirmació i l'exposició selectiva, pels quals els individus cerquen informació coherent amb allò que creuen prèviament (Lazarsfeld, Berelson i Gaudet 1944; Nickerson 1998) i eviten exposar-se a informació contrària a les seves actituds o creences, ja que aquest contrast genera incomoditat (Festinger 1962; Olson i Stone 2014). Tanmateix, la multiplicació de l'oferta informativa en línia fa difícil que els usuaris es facin una idea precisa del biaix ideològic de cada nou mitjà digital i, per tant, de la congruència entre aquests i les seves pròpies actituds. Així, els ciutadans s'estarien exposant ara, a internet, a estímuls i idees més diverses perquè no poden identificar el biaix de tots els mitjans digitals existents. Queda per saber quin és el sentit de la seva influència.

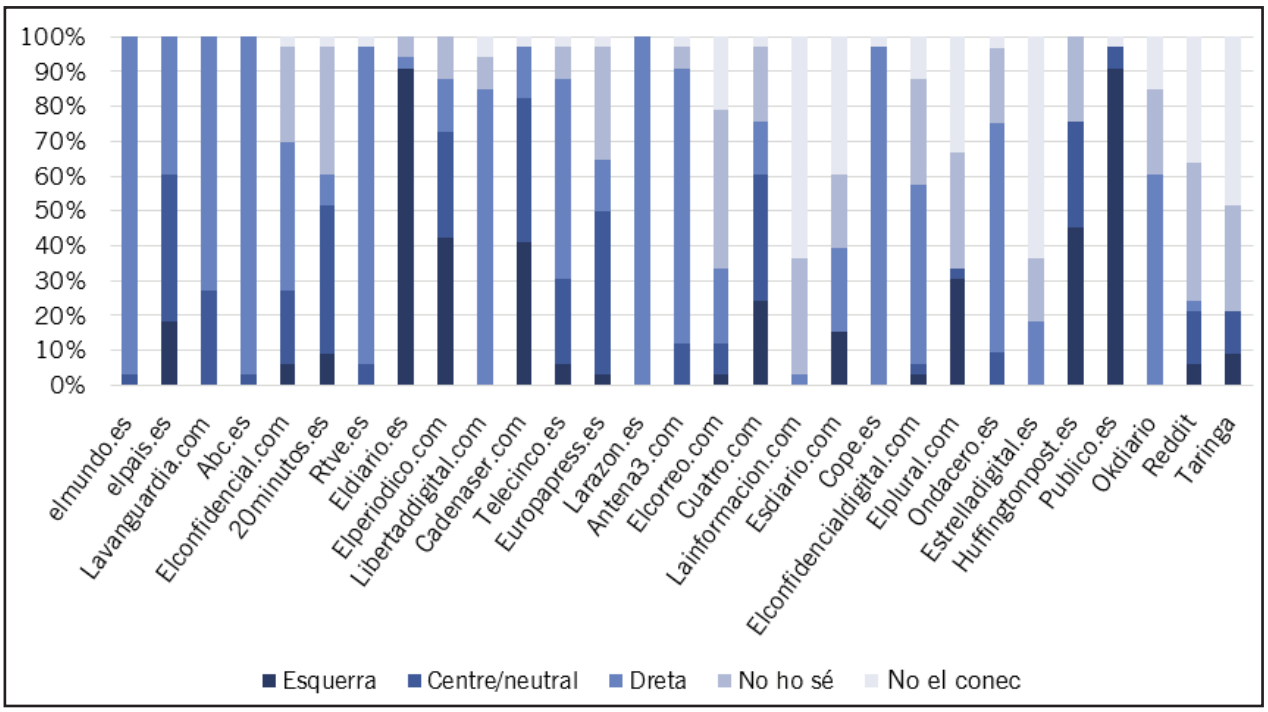
Dins les fronteres de l'Estat, només alguns estudis han tractat aquest tema. Entre les excepcions més destacades trobem els treballs d'Almiron, que ha analitzat l'estructura de propietat i les seves línies editorials per als mitjans tradicionals (2009) i per als diaris digitals sense referent imprès (2006). En una aproximació més recent, l'autora també s'ha ocupat de la diversitat ideològica d'aquests diaris i ha analitzat en quins termes es refereixen a les ideologies més tradicionals, però sense atribuir a cada mitjà un biaix o una etiqueta ideològica concreta, sinó representant el panorama conjunt que aquests mitjans ofereixen (Pineda i Almiron 2013). Però continuem sense tenir una brúixola comunament acceptada a la qual referir-nos quan parlem dels biaixos ideològics dels nous mitjans digitals.

Podem intentar una primera aproximació al fenomen de la ideologia dels mitjans digitals analitzant les percepcions de la ciutadania. Mitjançant tres enquestes diferents, ens hem apropat a la percepció de la ciutadania espanyola sobre la ideologia d'alguns dels principals mitjans digitals estatals.[2] El més destacable és el percentatge d'individus que no saben classificar els mitjans. Així, entre el 23 i el 33% de persones no saben quina és la ideologia del Huffington Post o de 20 Minutos, tot i conèixer-los. Gairebé un terç de la població espanyola no coneix quina és la ideologia de mitjans com eldiario.es o El Confidencial. Si preguntem a estudiants universitaris, gairebé la meitat no sap ubicar eldiario.es, El Confidencial o el Huffington Post. Una estratègia alternativa consisteix a preguntar a experts. La figura 1 palesa els resultats d'una enquesta realitzada el setembre de 2017 a 33 experts en les àrees de ciència política i ciències de la informació a Espanya. Se'ls va preguntar per la ideologia dels 30 mitjans més visitats l'any anterior, segons Alexa.

Si excloem les versions digitals de mitjans tradicionals com *El Mundo*, *ABC*, etc., trobem un percentatge sorprenentment alt de "No ho sé" i "No el conec", que arriba a ser de més del 50% per a La Información (3,6% de l'audiència digital, segons ComScore). Podem concloure, doncs, que situar aquests mitjans en un mapa mental d'ideologies resulta una tasca complicada, fins i tot per a experts en mitjans i política.

La recerca té l'objectiu de classificar els principals mitjans digitals al territori espanyol en funció del seu biaix ideològic, apostant per l'anàlisi de contingut automatitzat i, per tant, eficient i objectiu. Aquesta informació serà d'utilitat no només en

Figura 1. Percepció de diferents mitjans digitals. Enquesta a experts. Setembre de 2017. N=33



Font: Elaboració pròpia.

l'àmbit acadèmic per als debats ja presentats sobre exposició selectiva, sinó que també tindrà una importància política vital per avaluar la pluralitat dels mitjans i per millorar el grau d'alfabetització digital de la ciutadania, cosa que alhora es considera positiva per a la qualitat democràtica del sistema polític.

2. Marc teòric. Mesurar el biaix en els mitjans

2.1 Definicions i conceptes bàsics

El biaix ideològic no implica un intent deshonest ni deliberat de tergiversar la realitat, sinó una forma de descriure la realitat que és significativament i sistemàticament distorsionada (Groeling 2013: 130). Al seu torn, la ideologia s'ha definit com la distorsió d'una realitat objectiva que reflecteix construccions mentals subjectives i col·lectives (Benabou 2008:1). Un dels autors seminals en aquest debat, Converse, defineix la ideologia com les parts (o els subconjunts) d'un sistema de creences, com "una configuració d'idees i actituds en què els elements estan lligats a través d'alguna forma de constricció o interdependència funcional" (Converse 1964: 207).

La idea que proposa Converse (1964) implica que, com més dependència funcional hi hagi entre els elements d'un sistema de creences, menys recursos cognitius caldran per ser descrit o comprès. Des d'aquest punt de vista, una de les dimensions de judici que més útil ha estat per simplificar els esdeveniments en la política ha estat la dimensió esquerra-dreta. Sobre aquesta dimensió hom ubica partits, líders, polítiques i altres objectes de la política (Converse 1964: 214). La interdependència entre els elements que caracteritza un sistema de creences també explicaria, segons Converse, que la difusió social de les ideologies tendeixi a fer-se per "paquets".[3] Això afecta la interpretació de les mateixes ideologies. Els partits, per exemple, voten sobre diferents temes de manera connectada (Benoit i Laver 2006, 2007) i presenten paquets d'alternatives als electors (Downs 1957). Els electors utilitzen la dimensió esquerra-dreta per donar sentit a la decisió del vot i per prendre decisions sobre els paquets d'alternatives presentades.

Els mitjans de comunicació també difonen les ideologies polítiques a través de paquets, en aquest cas de conjunts de paraules o termes que evoquen altres conceptes connectats ideològicament. Amb aquestes construccions apel·len als diferents sistemes de creences i conceptes que els defineixen.

2.2 Limitacions dels estudis previs sobre el biaix dels mitjans

Estudis previs sobre el biaix ideològic dels mitjans han fet servir bàsicament dues aproximacions per mesurar-lo: la primera, basada en la caracterització de l'audiència, i la segona, en el contingut publicat (vegeu també Budak *et al*. 2016). La primera aproximació ha utilitzat el perfil ideològic de l'audiència d'un mitjà per atribuir-li una ideologia. Per exemple, la literatura sobre exposició selectiva a la informació (Freedman i Sears 1965) assumeix que l'audiència segueix mitjans afins ideològicament. Així, coneixent la ideologia de la seva audiència es pot atribuir una ideologia als mitjans (Bakshy, Messing i Adamic 2015; Gentzkow i Shapiro 2011; Newman, Fletcher, Kalogeropoulos, Levy i Nielsen 2017; Barberá i Sood 2014).

Aquesta aproximació és parsimoniosa i relativament simple. Tanmateix, la proliferació de mitjans fa cada cop més difícil per a l'audiència conèixer el biaix ideològic dels mitjans. Un altre inconvenient és que proporciona mesures relatives i no objectives d'aquest biaix. Si tenim en compte que els moviments de l'audiència poden ser molt sensibles a petites diferències en el biaix entre els mitjans, aquest mètode no ens permetria valorar bé les diferències existents (Budak *et al*. 2016).

La segona aproximació utilitzada a la literatura per identificar el biaix dels mitjans es basa en el contingut que aquests elaboren. Però la majoria de mitjans no prenen posicions explícites sobre els temes que cobreixen, i això constitueix una dificultat (Barberá i Sood 2016). Davant aquesta limitació, els treballs existents han seguit tres grans estratègies.

La primera consisteix a limitar l'anàlisi a un conjunt reduït però altament informatiu del contingut publicat. Es tracta del contingut editorial, que sí que incorpora explícitament el posicionament dels mitjans sobre els fets d'actualitat. Tanmateix, s'ha criticat els estudis que fan ús dels editorials, perquè mesuren només el biaix d'una part molt petita del contingut, que pot exagerar el biaix de la globalitat del diari (Barberá i Sood 2014).

La segona estratègia es basa en l'aprenentatge automatitzat per detectar patrons (lingüístics) en un conjunt ampli i indiscriminat de notícies. Es parteix de la identificació d'un conjunt de documents (per exemple, programes de partits) a partir dels quals es detecten paraules carregades ideològicament. Posteriorment, s'assigna una puntuació a cadascuna d'aquestes paraules, es recompten i s'utilitzen per estimar la ideologia del mitjà (Gentzkow i Shapiro 2010; Wihbey, Coleman, Joseph i Lazer 2017). Les paraules carregades ideològicament, però, representen un percentatge encara molt petit del contingut total publicat pels mitjans i, per tant, treballar amb aquest material produeix un volum elevat de soroll (Gentzkow i Shapiro 2010). A més, les paraules o frases associades amb una ideologia sovint són utilitzades per mitjans d'ideologia oposada en registres com l'humor, la ironia o el sarcasme per criticar adversaris polítics. Clarament, aquest ús dificulta la classificació dels mitjans (Barberá i Sood 2014: 4).

Finalment, la tercera estratègia es basa en una combinació d'aprenentatge automatitzat i codificació humana (o *crowdsourcing*) per superar algunes de les limitacions associades a l'estratègia basada únicament en l'aprenentatge automatitzat. La codificació humana permet identificar la ironia i la broma i corregir falsos positius (Budak *et al*. 2016).

2.3 Una nova direcció

En aquest treball apostem per la segona estratègia, basada totalment en l'ús de l'aprenentatge automatitzat, per identificar o estimar la ideologia d'una mostra estratègica de mitjans. La nostra proposta, però, presenta algunes novetats.

La primera és que aquí anem una mica més enllà dels estudis previs, i no basem la nostra anàlisi en paraules (o frases curtes) carregades ideològicament, sinó en un conjunt de sintagmes nominals connectats. D'aquesta manera ens assegurem que els termes pels quals comencem tinguin significat per si mateixos. La segona novetat és que no ens centrarem tant en una llista de termes propis de la dreta o l'esquerra sinó en els discursos en què apareixen (*frames*). La tercera és que farem servir piulades a Twitter de polítics com a text de referència per identificar la ideologia en lloc de programes electorals o els discursos parlamentaris.

Alguns estudis utilitzen els comptes de Twitter dels usuaris dels mitjans per deduir-ne la ideologia i, en última instància, atribuir-la als mitjans (Barberá i Sood 2014), però cap estudi, que sapiguem, ha utilitzat els comptes de Twitter de polítics per detectar quins termes i discursos són els típics d'una ideologia. Creiem que pot ser una estratègia eficient, perquè internet ha contribuït a la polarització dels debats en línia. Així doncs, a Twitter es faria servir un llenguatge amb més càrrega ideològica que en altres mitjans (Toff i Kim 2013), tot i que prou semblant a la dels diaris digitals (Mullainathan i Shleifer 2005). En segon lloc, conceptualitzacions recents dels partits polítics els presenten com a coalicions laxes formades per actors que comparteixen una agenda i uns objectius comuns (Bawn *et al*. 2012). En aquestes xarxes, l'ús de les paraules per part dels professionals de la comunicació per a la construcció d'un relat guanya importància (Toff i Kim 2013). El context o escenari en què aquesta coalició d'interessos que són els partits posaria a prova aquest llenguatge no serien els programes electorals, que poca gent llegeix i són bastant neutres, sinó les xarxes socials: un espai molt més dinàmic i en fase d'expansió (Newman *et al*. 2017).

3. Metodologia

Per classificar els mitjans digitals segons la seva ideologia, hem seguit tres fases que tot seguit veurem en detall.

3.1 Fase 1: Identificació del corpus per detectar discursos ideologitzats

A l'hora de triar el corpus de referència per identificar-hi continguts ideològics, ens vam decantar per les piulades dels polítics a Twitter, ja que és una eina que es caracteritza per la immediatesa, la brevetat i el col·loquialisme, i això permet utilitzar concepcions i recursos retòrics semblants als titulars dels diaris.[4] En concret, vam triar com a corpus de referència els comptes de Twitter de 296 parlamentaris espanyols en la XII legislatura.[5]

Per tal d'explotar el màxim nivell de contrast, i optimitzar la tasca d'atribució d'ideologia als diputats, en aquesta recerca ens hem limitat als dos partits amb una ideologia més extrema i clara en l'eix esquerra/dreta: la coalició Unidos Podemos (o simplement, Podemos) i Partido Popular (PP), respectivament.

Taula 1. Distribució en el temps de les piulades dels diputats espanyols en la XII legislatura des de l'inici de la seva activitat a la xarxa social

Any	Usuaris PP	Usuaris Podemos	Piulades PP	Piulades 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

Font: Elaboració pròpia.

Aquests són els dos partits polítics d'àmbit estatal (PAES) amb representació parlamentària que els espanyols situen més als extrems de l'eix esquerra/dreta (font: 8a onada del panel DEC/UAB, desembre de 2015).

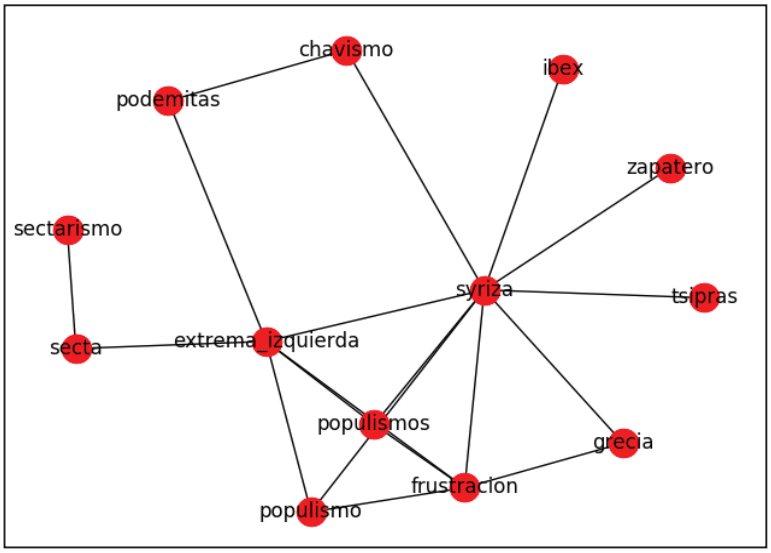
El conjunt de dades analitzades consisteix en gairebé mig milió de piulades dels diputats de Podemos i PP al Congrés dels Diputats.[6] La distribució del nombre de les piulades per partit es presenta a la taula 1.

3.2 Fase 2: Identificació de les relacions semàntiques que són característiques d'un discurs ideològic (*frames*)

Metodològicament, un *frame* és una relació de proximitat semàntica entre un terme *IT* (*ideology term*, que també podem entendre com a paraula clau) del discurs i uns termes *t* del mateix discurs.[7] La conjunció d'un terme *IT* amb una sèrie de termes *t* indica una visió determinada de les coses per part del terme *IT*. Així, per al PP l'*IT* "populismos" té associats els termes *t* "populismo, frustración, Syriza, extrema_izquierda, Grecia". Alhora, el terme *IT* "Syriza" presenta associats els termes "Zapatero, frustración, Grecia". Per tant, durant el període en què es van publicar les piulades, els representants del PP relacionaven Grècia amb el populisme i la frustració, etc. La figura 2 representa una xarxa que relaciona els termes al voltant de l'*IT* "populismos". Així doncs, no ens sorprendria trobar una piulada o el titular d'un editorial que digués que Zapatero és un populista i que ha estat el "Tsipras" d'Espanya. La piulada o el titular concentra una tesi, un missatge i uns valors del partit expressats amb uns termes determinats que conformen un discurs, que és el que recullen els frames.

Aquestes relacions evoquen el concepte de *frame* de Lakoff (2004), en què els conceptes tenen una estructura. Per exemple, la paraula "elefant" és un *frame* que evoca la imatge d'un elefant i tot el que coneixem sobre els elefants. De manera similar, els nostres *frames* volen capturar l'estructura de relacions que una

Figura 2. Representació gràfica del frame sorgit de l'*IT* "populismos" per al PP



Font: Elaboració pròpia.

sola paraula com "populismo" o "Grecia" té en el discurs d'un partit polític o d'un grup d'una ideologia determinada.

Per detectar els *frames*, primer identifiquem els sintagmes nominals de les piulades dels representants d'una ideologia determinada. Per a aquesta tasca hem fet servir l'eina Parse Tree del paquet pattern.es del projecte CLiPS.[8] Un cop obtinguts els sintagmes nominals, es busquen els seus termes *t*; és a dir, els termes més propers semànticament en el conjunt de totes les piulades. Per obtenir-los apliquem el mètode Word2vec[9] mitjançant un mòdul de Python, el qual indica que dos sintagmes nominals *p* i *p'* són propers si apareixen en contextos similars.[10]

És a dir, les paraules que solen estar al voltant de *p* també solen estar al voltant de *p'*. Aplicat a la detecció dels termes *t*, l'explicació que "populismos" i "extrema_izquierda" són propers és que les paraules que envolten "populismos" solen aparèixer també properes a "extrema_izquierda".

Seguidament, vam establir criteris per identificar quins d'entre tots els sintagmes nominals són *IT*. En primer lloc, el sintagma nominal ha d'aparèixer tant a les piulades del PP com a les de Podemos. Sense aquesta condició no podem decidir si hi ha discrepància en els *frames* entre els dos partits (atès que només un el fa servir). En segon lloc, l'*IT* ha d'aparèixer amb més freqüència en les piulades d'un partit que en les de l'altre. Considerem, aquí, que un criteri raonable és que un terme "propi" d'un partit ha d'aparèixer a les piulades dels seus diputats més del doble de vegades que al corpus de referència de l'altre partit. En tercer lloc, els *frames* dels partits (és a dir, els termes *t* associats a l'*IT*) han de ser diferents. És a dir, el vector que es genera amb les piulades d'un partit ha de tenir una distància considerable respecte al vector per al mateix terme generat amb les piulades del partit oposat. Un cop es creen els vectors dels sintagmes nominals del PP i de Podemos, es calcula la distància (*cosine similarity*) per a cada vector. Ens quedarem com a candidats a *IT* els que tinguin una *cosine similarity* inferior a 0.1, apuntant, per tant, a una gran diferència.

3.3 Fase 3. Comprovació de les correspondències entre els *frames* d'un discurs polític d'una ideologia determinada i les notícies dels diaris

A l'hora d'aplicar el mètode, hem decidit centrar-nos en alguns dels mitjans que s'ha detectat (vegeu la introducció) que generaven més confusió en l'audiència: el Huffington Post, El Confidencial, infoLibre i 20 Minutos. A més, hem inclòs l'ABC pel fet de ser el mitjà més clarament situat a la dreta en totes les enquestes analitzades, cosa que ens pot servir de punt de referència.

Hem obtingut els textos de la base de dades de premsa FACTIVA, i hem acotat la recerca entre inicis de desembre de 2016 (precampanya eleccions generals 2016) i finals de juny de 2017 (eleccions del 26 de juny de 2017 i inici de la XII legislatura).

Per comprovar les correspondències hem considerat diferents opcions:

a. Comptar la freqüència dels *IT* d'una ideologia determinada en cada diari. Així, un diari més afí al PP farà servir més *IT-PP* que un diari de línia ideològica d'esquerres.
b. Determinar si els vectors que descriuen els *IT* en les piulades i els vectors que descriuen els *frames* d'aquests *IT* en els diaris són semblants.
c. Centrar-nos en el nombre de termes *t* que acompanyen un *IT* per a cada partit que apareixen en els diferents diaris.

A l'apartat següent expliquem els resultats obtinguts pels diferents mètodes i les seves possibilitats de millora.

4. Resultats

4.1 *IT* característics del PP i de Podemos

Hem obtingut 327 *IT* característics del PP (*IT-PP*) i 113 de Podemos (*IT-Podemos*). Són, doncs, sintagmes nominals presents en els discursos del partit oposat, amb una freqüència superior al doble que en les piulades del partit ideològicament oposat, i amb un vector de t_s que té una distància (*cosine similarity*) inferior a 0.1 respecte al vector del mateix sintagma nominal generat amb les piulades del partit oposat (és a dir, generen marcs interpretatius molt diferents).

Per exemple, tant el PP com Podemos parlen del "proceso independentista", però el PP parla més del doble de vegades que Podemos d'aquesta idea. Els termes *t* amb què s'hi refereixen són extremament diferents (valor de la *cosine distance* entre el vector "proceso independentista" del PP respecte al vector generat per al mateix terme *IT* de Podemos = 0.0978). Per tant, aquest *IT* és divisori: té un *frame* del PP (dreta) i un *frame* de Podemos (esquerra), tot i ser més característic del PP. Ara bé, crida l'atenció la presència d'*IT* com "populismo", "proetarras" o "coleta" entre els *IT* propis de Podemos, ja que són termes que la dreta utilitza per desacreditar-los. Això apunta que les piulades de Podemos tenen una bona càrrega de referencialitat del discurs del partit contrari ideològicament.

4.2 Correspondència entre piulades i diaris segons la freqüència dels *IT*

La primera opció per comprovar la correspondència entre les piulades i els diaris va ser verificar la freqüència dels *IT* d'una ideologia determinada en els diaris. Així, un diari afí al PP farà servir més *IT-PP* que un altre diari.

A la figura 3 veiem el percentatge d'IT-PP distribuïts per diaris. Una mica més del 40% de les aparicions d'IT-PP es produeixen a l'ABC. El segueixen infoLibre i El Confidencial. Així, el diari més afí al PP seria l'ABC, mentre que 20 Minutos seria el més allunyat. Què passa, però, si observem la correspondència entre els *IT-Podemos* i els mateixos diaris?

A la figura 4 observem que l'ABC també és el diari en què es concentren més *IT-Podemos*, encara que menys acusadament que a l'exemple anterior. La distribució relativa de la resta de diaris és molt similar a l'exemple anterior. Aquests resultats s'allunyen massa del criteri dels ciutadans i dels experts perquè puguem refiar-nos-en. Per tant, no sembla que la distribució per freqüència dels *IT* segons la ideologia serveixi per detectar alineaments clars entre les piulades dels polítics i els diaris. L'apropiació per part de Podemos de *frames* derivats d'*IT* originàriament de dretes (i més presents en els diaris presumiblement més de dretes) podria estar al darrere d'aquests resultats tan contraintuïtius.

4.3 Correspondència entre piulades i diaris segons la similitud de *frames*

El pas següent va ser comprovar si els vectors que descriuen els *IT* en les piulades i els vectors que descriuen els frames d'aquests *IT* en els diaris són semblants. Per exemple, volíem comprovar si els diaris afins a Podemos solen vincular més sovint la UE amb l'austeritat i amb Angela Merkel que no pas els diaris afins al PP.

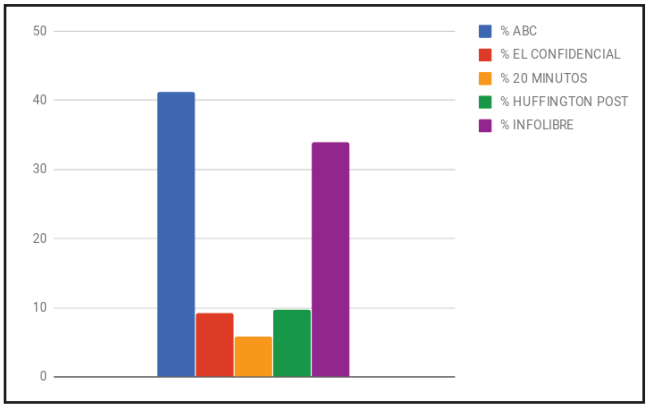
De la mateixa manera que havíem fet amb les piulades dels diputats, vam convertir els sintagmes de cada diari en vectors, les dimensions dels quals eren els termes *t*; és a dir, els termes més relacionats semànticament, obtinguts amb Word2vec. Els vectors dels *IT-PP* i *IT-Podemos* es van comparar –via *cosine similarity*– amb els vectors dels mateixos sintagmes nominals dels diaris. Vam comprovar que la referencialitat als *IT* del partit oposat ideològicament també era una característica dels diaris, per la qual cosa vam obtenir resultats semblants als de la freqüència d'*IT*.

4.4 Correspondència entre piulades i diaris segons els focus en els *t*

L'última opció explorada se centrava en els termes *t* i la seva capacitat de relacionar-se amb *IT* d'ideologia diferent. En termes de *frames* significa que, donat un *IT*, els diaris afins a un partit coincideixen en parlar dels mateixos *t*.

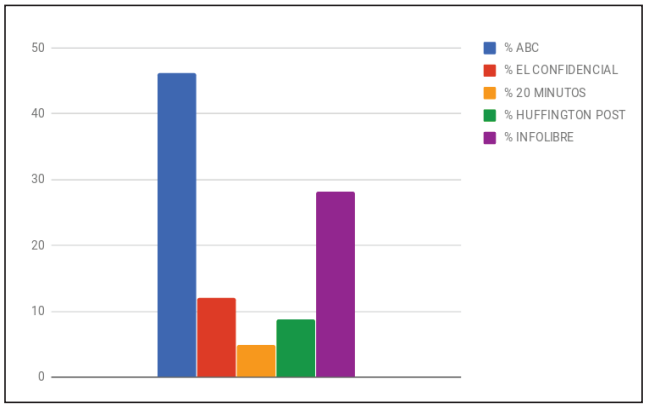
Per comprovar-ho, vam recollir els termes *t* relacionats semànticament amb els *IT* de les piulades del PP i de Podemos. Després vam comprovar quants *t* propis de cada partit

Figura 3. Percentatge d'*IT-PP* per als diaris analitzats



Font: Elaboració pròpia.

Figura 4. Percentatge d'*IT-Podemos* per als diaris analitzats



Font: Elaboració pròpia.

Taula 2. Vectors de coocurrència de termes *t* per a una sèrie d'*IT* del PP en els diaris analitzats

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

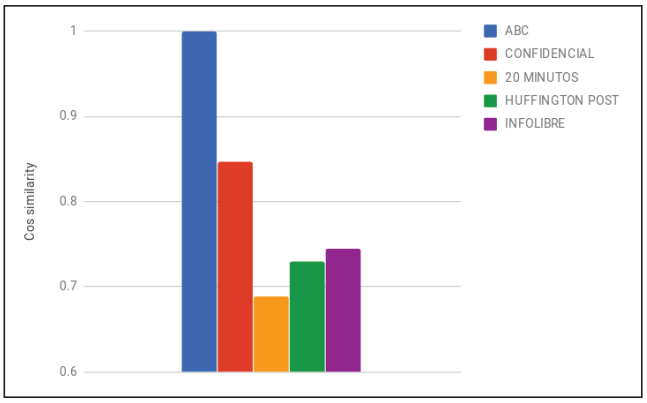
Font: Elaboració pròpia.

Taula 3. Vector de coocurrència dels tres termes *t* associats a l'*IT* de Podemos "proetarras"

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

Font: Elaboració pròpia.

Figura 5. Proximitat dels diaris respecte a l'ABC quant als *frames* del PP



Font: Elaboració pròpia.

apareixien a les notícies d'un diari i vam crear, per a cada *IT*, un vector amb el nombre de t del *PP* i de Podemos coocurrents per a cada diari.[11] La taula 2 il·lustra aquests vectors amb els t_{pp} relacionats amb "centralidad", "abismo", "coleta" i "populismo". Per exemple, "centralidad" i "coleta" tenen 19 i 1 t_{pp} coocurrents al diari ABC, respectivament, però cap t_{pp} a infoLibre. "Populismo" té 2 t_{pp} a l'ABC i 1 a El Confidencial, però cap a 20 Minutos ni al Huffington Post.

Creats els vectors per a cada diari, agafem el diari en què més apareixen els *frames* del PP com a referència: l'ABC. La incidència de termes *t* del PP en la resta de diaris es representa en relació amb aquest diari, que pren el valor 1.

Com es pot veure a la figura 5, El Confidencial és el diari més proper a l'ABC si tenim en compte la freqüència d'aparició dels *IT* amb els *t* del PP. 20 Minutos, Huffington Post i infoLibre n'estan més allunyats, amb 20 Minutos com el que ho està més. Amb aquest sistema –vectors de coocurrència en els diaris– podríem, en principi, trobar "falsos" *IT* d'esquerres. Per exemple, la taula 3 representa el vector corresponent a "proetarras" –un *IT*, recordem-ho, més freqüentment utilitzat en el discurs de Podemos que en el del PP.

"Proetarras" té tres termes *t* amb els quals coapareix en un sol diari, que és l'ABC. Tenint en compte que aquests t són "Otegui", "Bildu" i "ETA", caldria reflexionar sobre si la coaparició d'un *IT* amb uns *t* determinats en un diari ja alineat ideològicament (com hem fet amb l'ABC) és un criteri per (re)classificar –ideològicament– un *IT*, malgrat que sigui molt utilitzat pel partit contrari ideològicament. En tot cas, aquest procediment podria servir per fer "neteja" d'*IT* classificats erròniament com d'esquerres o de dretes, i sembla una possible solució al problema de l'apropiació de *frames* per part del partit contrari com a eina per atiar el conflicte, assenyalar paradoxes en els contraris, etc.

5. Conclusions

Mesurar el biaix dels mitjans digitals escrits és necessari perquè necessitem saber quin és l'abast i el sentit del seu efecte, a fi d'avaluar la pluralitat del panorama informatiu i millorar el grau d'alfabetització digital de la ciutadania.

La revisió de la literatura existent sobre la mesura del biaix ideològic dels mitjans ha revelat que els diferents mitjans emprats fins a l'actualitat presenten un seguit de limitacions. Atribuir a cada mitjà la ideologia de la seva audiència assumeix que els ciutadans coneixen el biaix dels mitjans i s'hi exposen selectivament; però ni el primer supòsit ni el segon són sempre veritat. La segona aproximació utilitza el contingut publicat, seguint tres possibles variants. La primera és limitar-se a una petita quantitat de text molt indicativa del contingut (editorials), la segona consisteix a detectar automàticament patrons lingüístics i, la darrera, a combinar aquests procediments automatitzats amb codificació humana. La utilització d'editorials tendeix a presentar una ideologia més extremista de la que té realment el mitjà, i la darrera estratègia és força costosa en termes de recursos. Hem adoptat, doncs, la segona.

La nostra perspectiva, però, inclou tres novetats. Primer, la nostra unitat d'anàlisi no és una llista de paraules o frases carregades ideològicament, sinó un conjunt de sintagmes nominals connectats i carregats ideològicament. Segon, la mesura utilitzada per assignar una ideologia no s'estableix

únicament a partir de la freqüència d'ús d'aquestes cadenes de paraules sinó, sobretot, a partir de la discrepància entre aquestes. El darrer aspecte innovador rau en el cos de text que fem servir com a referent per identificar frames ideològics: utilitzem piulades de líders polítics a Twitter, i no programes electorals o discursos parlamentaris.

Per tal d'identificar contingut amb càrrega ideològica, ens hem centrat en els frames (conjunts de paraules semànticament properes al voltant d'un terme *IT*) propis dels dos partits d'àmbit estatal més polaritzats d'acord amb les percepcions de l'opinió pública espanyola: el PP i Podemos. Hem detectat un seguit de termes comuns a tots dos partits, però més presents en les piulades dels diputats d'un partit que en les de l'altre. Hem verificat que els termes *t* de què s'acompanyen siguin força diferents abans d'identificar els *frames*.

Durant aquest procés hem topat amb diferents vies mortes. Una d'aquestes ha estat comptar les correspondències dels frames de cada partit en els diaris, degut, probablement, a l'apropiació per part de Podemos de *frames* crítics amb ells sorgits a la dreta. De manera similar, comparar la distància entre frames de partits i de diaris ens duu al mateix punt: sembla que els resultats només tenen sentit si atenem els frames propis del PP i les similituds entre aquest partit i els mitjans, però això no s'aplica a Podemos.

Propers desenvolupaments hauran d'intentar solucionar el problema de les referències iròniques als marcs d'interpretació de l'adversari. Aquest s'ha apuntat anteriorment com un dels principals problemes de l'anàlisi del contingut a través de l'aprenentatge automatitzat per atribuir una ideologia als mitjans (Barberá i Sood 2014). Les nostres dades confirmen que Podemos referencia les crítiques sorgides des de la dreta a les seves actituds i argumentacions "populistes", fent-ne befa, cosa que impossibilita identificar-ne de manera automàtica la intencionalitat. Una altra possibilitat seria incloure una dimensió temporal per donar més pes als termes que apareguin primer en el temps com a elements identificadors del frame d'un partit. D'altra banda, es podria diluir aquest error de mesura ampliant el corpus de referència a la resta de partits d'àmbit estatal. Així, aquest fenomen típic de Podemos quedaria diluït entre les piulades del PSOE. Finalment, es podria combinar l'aprenentatge automatitzat amb la codificació humana. Aquesta estratègia, tot i ser més costosa, ens permetria descartar termes utilitzats amb ironia o sarcasme.

Notes

1. L'alfabetització mediàtica és el desenvolupament d'una comprensió raonada i crítica de la naturalesa dels mitjans de comunicació i els seus efectes, de com creen significat i de com organitzen la seva pròpia realitat (Gilster 1997; Aparici 1996).
2. Aquestes enquestes es van efectuar entre 2015 i 2016. La primera és una enquesta del grup de recerca eGovernança: Administració i Democràcia Electrònica (GADE) de la Universitat Oberta de Catalunya (UOC), realitzada per al projecte Opinionet. La segona és una enquesta del grup de recerca Democràcia, Eleccions i Ciutadania (DEC) de la Universitat Autònoma de Barcelona (UAB). La tercera també és una enquesta del grup GADE a la qual han respost els estudiants de la UOC.
3. Aquest format en paquets es correspon gairebé perfectament amb la noció de frames o marcs d'interpretació propi de les anàlisis semàntiques.
4. Després de diferents proves exploratòries, s'han desestimat els discursos parlamentaris perquè no era possible construir un corpus de text prou gran per extreure'n termes o conjunts de termes carregats ideològicament. En aquest sentit, també s'ha optat per prescindir dels programes electorals, perquè en les anàlisis preliminars efectuades no s'han detectat discrepàncies significatives en els *frames* dels diferents partits a partir dels programes electorals. A més, els programes electorals (i les codificacions proposades pel projecte Party Manifesto) ja no es fan servir per estimar les posiciones ideològiques dels partits (Benoit i Laver 2006, 2007). Finalment, els programes dels partits utilitzen un llenguatge molt formal que s'allunya del llenguatge més informal i amb càrrega ideològica que sí que es fa servir en els mitjans.
5. D'entre els 350 diputats, només 296 tenen un compte actiu de Twitter.
6. Algunes persones, sobretot del PP, eren diputades el 2009; però dins de la coalició Unidos Podemos, només algunes persones pertanyents a Izquierda Unida ho eren abans de 2016. Entenem, però, que amb les seves piulades abans d'aquesta data estan difonent missatges i valors consonants amb aquest partit.
7. Entenem aquí *proximitat semàntica* com a coocurrència, o aparèixer en posicions adjacents al mateix text. Es tracta d'un concepte propi de l'anàlisi quantitativa de textos. L'algoritme utilitzat per determinar-la (Word2vec) recull aquesta proximitat física de les paraules mantenint les propietats gramaticals dels textos dels quals s'extreuen.
8. <https://www.clips.uantwerpen.be/pages/pattern-es>. El sintagma nominal és –juntament amb els verbs– l'element bàsic que estructura una oració, la seu principal del significat lèxic i, en definitiva, la manera com es denominen els conceptes. Així, podem recollir denominacions com "Tribunal Superior de Justicia", en comptes del bigrama "Tribunal Superior" o dels unigrames "Tribunal", "Superior" i "Justicia".

9. Word2vec és un mètode representatiu de la tendència més recent en aprenentatge automàtic que es diu *deep learning,* amb una estructura de xarxes neuronals (Dikolov *et al*. 2013). És un mètode que s'està aplicant amb molt d'èxit a la traducció automàtica (Mikolov, Quoc i Sutskever 2013), a l'anàlisi del sentiment (Acosta *et al*. 2017) i a la classificació de documents (Lilleberg, Zhu i Zhang 2015). Fins i tot l'abstracció de la idea de context, definit en un espai vectorial, ha fomentat l'aparició d'altres aplicacions com els recomanadors (Ozsoy 2016).

10. Word2vec utilitza un algoritme que calcula, per cada sintagma nominal, els sintagmes nominals més propers. La proximitat és un valor que va del 0 a l'1 (de menys proper a més proper). Per a aquest projecte hem considerat com a termes t els que superen el valor de la mediana (0.5).

11. La mètrica utilitzada per mesurar la coocurrència ha estat la Normalized Google Distance (NGD), amb un rang de valors entre el 0 (cap proximitat) i l'1 (màxima proximitat). És una mesura de distància semàntica segons el grau de coaparició de dos termes, en el nostre cas, entre l'*IT* i el seu *t,* en el titular i en el cos de la notícia.

Referències

Acosta, J.; Lamaute, N.; Luo, M.; Finkelstein, E.; Cotoranu, A. *Proceedings of Student-Faculty Research Day*. CSIS, Pace University, 5 de maig 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". Manuscrit no publicat, 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.

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

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

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

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

Freedman, J. L.; Sears, D. O. "Selective exposure". A: Berkowitz L. (ed.). *Advances in Experimental Social Psychology*. Vol. 2. Nova 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 [Estats Units]: 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, juliol de 2015.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. "Efficient estimation of word representations in vector space". A: *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, p. 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 novembre 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>.