# The challenge of measuring ideological bias in written digital media

**Ana S. Cardenal**
*Associate Lecturer at the Universitat Oberta de Catalunya (UOC)*

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

**Carol Galais**
*Postdoctoral Researcher at the UOC*

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

**Joaquim Moré**
*PhD Candidate on the UOC Information Society Programme*

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

**Camilo Cristancho**
*Postdoctoral Researcher at the Universitat de Barcelona (UB)*

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

**Sílvia Majó-Vázquez**
*Postdoctoral Researcher at the Reuters Institute for the Study of Journalism at the University of Oxford*

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

***Article received on 16/04/18 and accepted on 19/06/18***

***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.*

***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.*

## 1. Introduction. Why study bias in digital media?

In our country, as in the rest of the western world, digital media are growing. In Spain alone, 579 new media outlets were set up in 2015, most of them only with online versions (APM 2015). This mounting media diversity paints a fragmented picture and is a challenge for researchers in political communication. We do not know the degree of plurality of our digital media, i.e. their diversity from an ideological point of view. Furthermore, in order to learn the possible impact of the media on public opinion, we first need to know what their political leaning is.

The Council of Europe (1994) argues that the degree of plurality of a country's media system is a positive factor for such system. Accordingly, identifying the ideological bias of the numerous digital media outlets should enable us firstly to evaluate a media system's diversity and ultimately its input into the democratic process, and secondly take action if the rising media offering does indeed mean that the media are increasingly partisan and polarised (Stroud 2011). In addition, providing the audience with information about the bias of new media would add to their media literacy (Buckingham 2007; Gilster 1997) and consequently have a positive impact on their civic skills, on the identification of fake news and, at the end of the day, on more effective control of rulers.[1]

As for the media's impact on public opinion, research has shown that their influence is limited by confirmation bias and selective exposure, whereby individuals seek out information which is consistent with the views they already hold (Lazarsfeld, Berelson and Gaudet 1944; Nickerson 1998) and avoid exposing themselves to any that conflicts with their attitudes or beliefs since this comparison generates discomfort (Festinger 1962; Olson and Stone 2014). However, the burgeoning of the range of online information makes it difficult for users to get an accurate idea of the ideological bias of each new digital media outlet and therefore of the congruence between such media and their own attitudes. Hence the public would now be exposing themselves to more diverse stimuli and ideas online because they are unable to identify the bias of all the digital media outlets now available. It remains to be seen which way their influence will go.

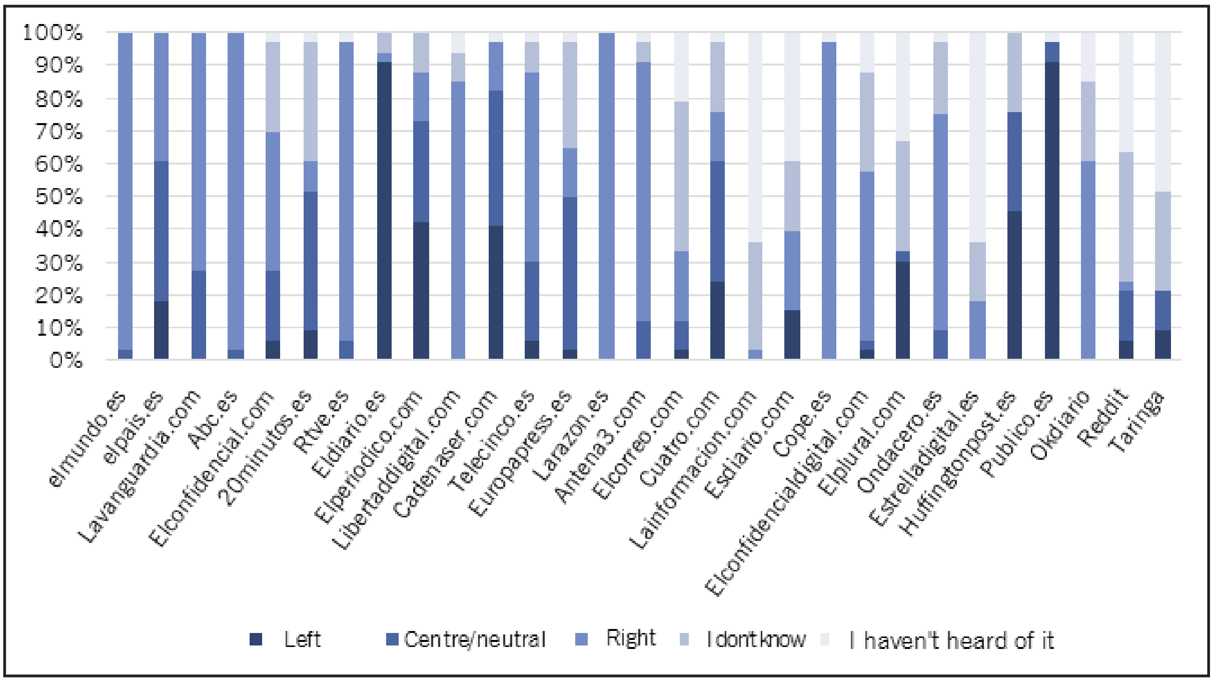
There are only a few studies which have addressed this issue in Spain. The most notable exceptions include the papers by Almiron, who has analysed ownership structure and editorial lines for traditional media (2009) and for digital newspapers without a print version (2006). In a more recent study the author has also tackled the ideological diversity of these newspapers by examining the terms they use to refer to the most traditional ideologies, albeit without attributing a specific ideological bias or label to each media outlet and instead depicting the overall landscape presented by these media (Pineda and Almiron 2013). Nonetheless, we still do not have a commonly accepted compass to refer to when we discuss the ideological biases of the new digital media.

We can initially approach digital media's ideology by analysing public perceptions. We have used three different surveys to examine Spaniards' perception of the ideology of some of the main national digital media outlets.[2] The most remarkable thing is the percentage of individuals who are unable to classify the media. Thus between 23 and 33% of people do not know what the ideology of the Huffington Post or 20 Minutos is, even though they are aware they exist. Almost a third of Spaniards do not know what the ideology of media outlets such as eldiario.es or El Confidencial is. If we ask university students, almost half are unable to place eldiario.es, El Confidencial or the Huffington Post on the ideological spectrum. An alternative strategy is to ask the experts. Figure 1 shows the results of a survey conducted in September 2017 with 33 political science and information science experts in Spain. These experts were asked about the ideology of the 30 media sites most visited in the previous year according to Alexa.

If we exclude digital versions of traditional media such as *El Mundo*, *ABC*, etc., we find surprisingly high "I don't know" and "I haven't heard of it" percentages which stand at more than 50% for La Información (3.6% of the digital audience according to ComScore). We can thus conclude that placing these media on a mental map of ideologies is tricky even for media and politics experts.

The purpose of this research is to classify the main digital media in Spain by their ideological bias using machine and consequently efficient and objective content analysis. This information will be useful not only in academia for the debates noted above about selective exposure, but also of vital political importance for evaluating media plurality and improving the public's digital literacy which at the same time is seen as constructive for the political system's democratic quality.

**Figure 1. Perception of digital media outlets. Experts' survey. September 2017. N=33**



Source: authors.

## 2. Theoretical framework. Measuring bias in the media

### 2.1 Definitions and key concepts

Ideological bias does not mean a dishonest and deliberate attempt to twist reality but rather a portrayal of it which is significantly and systematically distorted (Groeling 2013: 130). In turn, ideology has been defined as the distortion of an objective reality that reflects subjective and collective mental constructions (Benabou 2008:1). One of the seminal authors in this debate, Converse, defines ideology as the parts (or subsets) of a belief system, as "a configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence" (Converse 1964, 207).

The idea put forward by Converse (1964) suggests that the more functional interdependence there is between the components of a belief system, the fewer cognitive resources will be needed to describe or grasp it. From this standpoint, one of the dimensions of judgment that has been most useful in simplifying events in politics has been the left-right one. Parties, leaders, policies and other political objects are placed along this dimension (Converse 1964, 214). Converse further argues that the interdependence between the components making up a belief system would also explain the fact that ideologies tend to be socially diffused in 'packages'.[3] This impacts the interpretation of the ideologies themselves. Parties, for example, vote on different issues in a connected way (Benoit i Laver 2006, 2007) and present alternative packages to voters (Downs 1957). Voters use the left-right dimension to give meaning to their voting choice and to make decisions about the packages of alternatives on offer.

The media also disseminate political ideologies through packages, in this case a set of words or terms which call to mind other ideologically connected concepts. They use these constructions to appeal to the various belief systems and concepts that define them.

### 2.2 Limitations of previous media bias studies

Previous studies about the ideological bias of the media have essentially used two approaches to measure it: the first is based on describing the audience and the second on the published content (see also Budak *et al*. 2016). The first approach has used the ideological profile of a media outlet's audience to attribute an ideology to it. For example, the literature on selective exposure to information (Freedman and Sears 1965) assumes that the audience follows ideologically related media. Thus knowledge of the ideology of the media's audience enables us to attribute an ideology to them (Bakshy, Messing and Adamic 2015; Gentzkow and Shapiro 2011; Newman, Fletcher, Kalogeropoulos, Levy and Nielsen 2017; Barberá and Sood 2014).

This approach is frugal and relatively simple. However, the proliferation of media makes it increasingly difficult for the audience to become aware of their ideological bias. Another drawback is that it provides relative and non-objective measurements of this bias. Bearing in mind that audience shifts can be very sensitive to small differences in bias between media outlets, this method would not enable us to evaluate the differences properly (Budak *et al*. 2016).

The second approach used in the literature to identify media bias draws on the content they produce. However, most media outlets do not take up explicit stances on the issues they cover, which is something of a problem (Barbera and Sood 2016). Given this limitation, existing papers have used three major strategies.

The first is to restrict the analysis to a small but highly informative set of published output, namely editorial content, which does plainly set out the media's positioning on current affairs. However, studies using editorials have been criticised because they measure only the bias of a very small part of the newspaper's output which may exaggerate its overall bias (Barberá and Sood 2014).

The second strategy leverages machine learning to detect (linguistic) patterns in a broad and indiscriminate set of news items. It is based on identifying a set of documents (for example, party programmes) which are used to detect ideologically charged words. Subsequently each of these words is given a score and they are counted and used to assess the media outlet's ideology (Gentzkow and Shapiro 2010; Wihbey, Coleman, Joseph, and Lazer 2017). However, ideologically charged words account for a still very small percentage of the total content published by the media and hence working with this material produces a high volume of noise (Gentzkow and Shapiro 2010). In addition, the words or phrases associated with an ideology are frequently used by opposing ideologies in registers such as humour, irony or sarcasm to criticise political adversaries. Clearly such use makes it difficult to classify the media (Barberá and Sood 2014, 4).

Finally, the third strategy is based on a combination of machine learning and human coding (or crowdsourcing) to overcome some of the limitations associated with the strategy based solely on machine learning. Human coding makes it possible to identify irony and joking and correct false positives (Budak *et al*. 2016).

### 2.3 A new direction

In this paper we opt for the second strategy based entirely on the use of machine learning to identify or assess the ideology of a strategic sample of media outlets. Nonetheless, our approach does have some new features.

The first is that here we go a little beyond the previous studies and we do not base our analysis on ideologically charged words (or short phrases) but rather on a set of connected noun phrases. This means we can make sure that the terms we begin with have meaning in themselves. The second innovation is that we focus not so much on a list of terms typical of the right or the left but on the discourses in which they appear (frames). The third is that we use politicians' tweets as a reference text for identifying ideology instead of electoral programmes or parliamentary speeches.

Some studies use the Twitter accounts of media outlet users to figure out their ideology and ultimately attribute it to the media outlet (Barberá and Sood 2014). However, no study that we know of has used the Twitter accounts of politicians to identify which terms and discourses are typical of an ideology. We believe that this may well be an effective strategy because the Internet has helped to polarise online debates. Hence more ideologically charged language would be used on Twitter than in other media (Toff and Kim 2013), albeit quite similar to what can be found in digital newspapers (Mullainathan and Shleifer 2005). Secondly, recent portrayals of political parties present them as loose coalitions made up of actors who share a common agenda and objectives (Bawn *et al*. 2012). The use of words by communication professionals to build a narrative is gaining importance on these sites (Toff and Kim 2013). The context or scenario where this coalition of interests, which is what parties are, would test out this language would not be electoral programmes, which few people read and are quite neutral, but rather social media which are a much more vibrant and expanding venue (Newman *et al*. 2017).

## 3. Methodology

We have classified digital media by ideology in three stages which we will see in detail below.

### 3.1 Stage 1: Identification of the corpus to detect ideological discourses

We opted for the tweets of politicians on Twitter as the reference corpus for identifying ideological content because it is a tool characterised by immediacy, brevity and colloquial language which allows the use of concepts and rhetorical resources similar to newspaper headlines.[4] Specifically, our reference corpus was the Twitter accounts of 296 Spanish MPs in the 12th Parliament.[5]

In order to mine the highest level of contrast and optimise attribution of ideology to the MPs, we have restricted ourselves in this research to the two parties with a more extreme and clear ideology on the left/right axis: the Unidos Podemos (or simply Podemos) coalition and the Partido Popular (PP), respectively. These are the two state-wide political parties with parliamentary representation that Spaniards place most at the ends of the left/right axis (source: 8th wave of the DEC/UAB panel, December 2015).

The dataset analysed consists of almost half a million tweets by Podemos and PP MPs. [6] Table 1 shows the distribution of tweets per party.

**Table 1. Distribution over time of tweets by Spanish MPs in the 12th Parliament from when they joined the social media site**

| Year | PP users | Podemos users | PP tweets | Podemos tweets |
|---|---:|---:|---:|---:|
| 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 |

Source: authors.

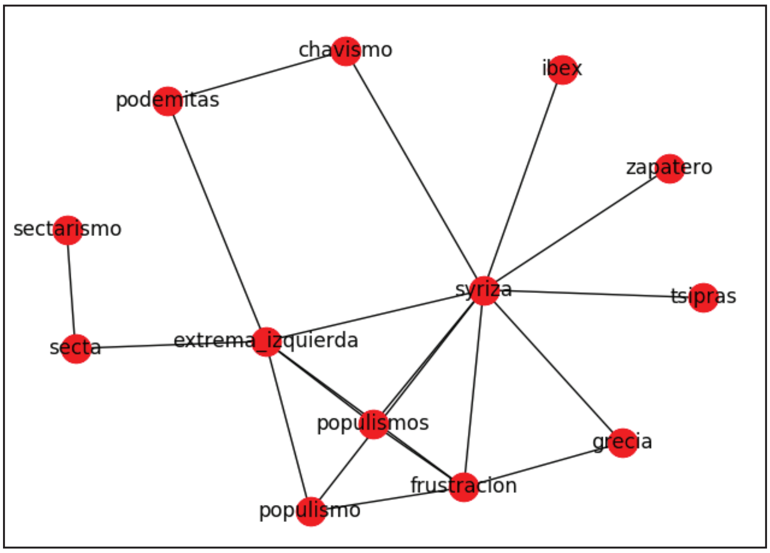
### 3.2 Stage 2: identifying the semantic relationships which are characteristic of an ideological discourse (frames)

Methodologically speaking, a frame is a semantic proximity relationship between an *IT* (ideology term, which we could also see as a keyword) in the discourse and some *t* terms in the same discourse.[7] The conjunction of an *IT* with a series of *t* terms indicates a certain view of things by the *IT*. While for the PP the *IT* “populisms” has associated the *t* terms “populism, frustration, Syriza, extreme_left, Greece”, the *IT* “Syriza” is associated with the terms “Zapatero, frustrations, Greece”. Thus during the period when the tweets were posted, PP MPs related Greece with populism and frustration, etc. Figure 2 shows a network that relates the terms around the *IT* “populisms”. Hence we would not be surprised to find a tweet or the headline of an editorial which said that Zapatero is a populist and has been the Spanish Tsipras. The tweet or the headline concentrates a thesis, a message and some values of the party expressed with particular terms which make up a discourse which in turn is what the frames collate.

These relationships call to mind the frame concept in Lakoff (2004) where concepts have a structure. For example, the word *elephant* is a frame that evokes the image of an elephant and everything we know about elephants. In a similar way our frames seek to capture the structure of relationships that a single word like *populism* or *Greece* has in the discourse of a political party or a group with a particular ideology.

To detect the frames we first identified the noun phrases in the tweets by the representatives of a given ideology. We did this using the Parse Tree tool in the pattern.es package from the CLiPS project.[8] Once we had obtained the noun phrases we then searched for their *t* terms, i.e. the terms that are semantically closest to the set of all the tweets. To get them we used the Word2vec[9] method with a Python module which indicates that

**Figure 2. Graphic representation of the frame coming out of the *IT* "populisms" for the PP**



Source: authors.

two noun phrases *p* and *p'* are close if they appear in similar contexts.[10]

That is to say, the words that are usually around *p* are also usually around *p'*. When applied to identifying *t* terms, the explanation why "populisms" and "extreme_left" are close is that the words surrounding "populisms" usually also appear close to "extreme_left".

Next we set criteria to identify which of all the noun phrases are *IT*s (ideology terms). In the first place, the noun phrase has to appear in both the PP and Podemos tweets. Without this condition we cannot decide if there is a discrepancy in the frames between the two parties (since only one uses it). Secondly, the *IT* should appear more frequently in the tweets of one party than the other. We consider that a reasonable criterion here is that a term "typical" of a party must appear in its MPs' tweets more than twice as often as in the reference corpus of the other party. Thirdly, the frames of the parties (that is to say, the *t* terms associated with the *IT*) must be different. In other words, the vector generated with the tweets of one party must be a considerable distance from the vector for the same term generated with the tweets of the opposite party. Once the vectors are created by the noun phrases of the PP and Podemos, the distance (cosine similarity) is calculated for each vector. Our candidates to be *IT* will be the ones which have a cosine similarity less than 0.1, thus indicating a big difference.

### 3.3 Stage 3. Checking correspondences between the frames of a political discourse of a particular ideology and the news items in newspapers

When applying the method we decided to focus on some of the media outlets where there has been greatest audience confusion (see introduction): the Huffington Post, El Confidencial, infoLibre and 20 Minutos. We have also included *ABC* as the most clearly right-wing media outlet in all the surveys analysed which will be our point of reference.

We obtained the texts from the FACTIVA press database and restricted our search to the time from the beginning of December 2016 (pre-campaign period for the 2016 general election) and the end of June 2017 (26 June 2017 general election and the start of the 12th Parliament).

We have considered a number of options to check the correspondence:

Counting the frequency of the *IT*s of a particular ideology in each newspaper. Thus a newspaper closer to the PP will use more *IT-PP* than an ideologically left-wing newspaper.

Determining whether the vectors that describe the *IT* in the tweets and the vectors that describe the frames of these *IT*s in the newspapers are similar.

Focussing on the number of *t* terms that go with an *IT* for each party which appear in the various newspapers.

In the next section we set out the results obtained by the different methods and how they might be improved.

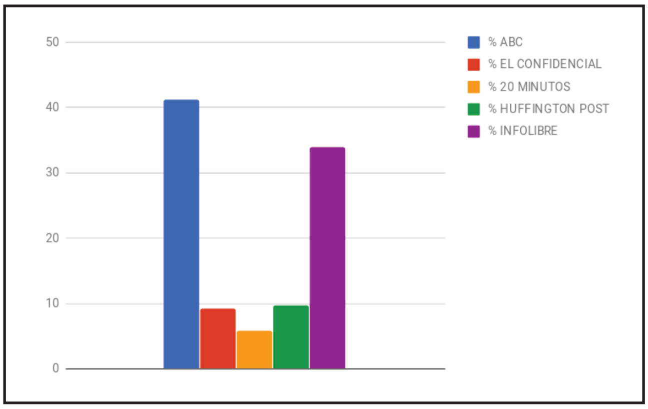
## 4. Results

### 4.1 *IT*s characteristic of the PP and Podemos

We have obtained 327 *IT*s characteristic of the PP (*IT-PP*) and 113 for Podemos (*IT-Podemos*). They are, then, noun phrases present in the discourses of the opposite party at a frequency higher than double than in the tweets of the ideologically opposed party and with a *t* vector with a distance (cosine similarity) of less than 0.1 with respect to the vector of the same noun phrase generated with the tweets of the opposite party (i.e. they generate very different interpretative frames).

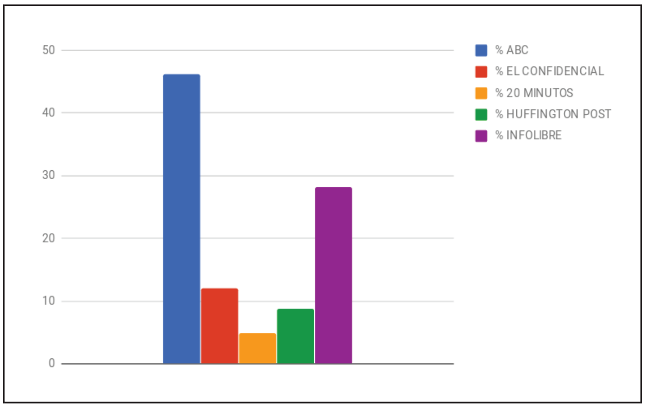
For example, both the PP and Podemos talk about the "independence process", but the PP mentions it twice as often as Podemos. The *t* terms they use to refer to it are extremely different (value of the cosine distance between the PP's "independence process" vector compared to the vector

**Figure 3. Percentage of *IT-PP* by newspapers analysed**



Source: authors.

**Figure 4. Percentage of *IT-Podemos* by newspapers analysed**



Source: authors.

generated by the same IT term for Podemos = 0.0978). Therefore, this *IT* is divisive: it has a PP frame (right) and a Podemos frame (left), in spite of being more characteristic of the PP. However, the presence of *IT*s such as *populism*, *pro-ETA* and *ponytail* among the *IT*s typical of Podemos is striking because they are terms that the right uses to discredit it. This suggests that Podemos's tweets have a considerable referential charge to the ideologically opposed party's discourse.

### 4.2 Correspondence between tweets and newspapers by *IT* frequency

The first option for verifying the correspondence between tweets and newspapers was to verify the frequency of the *IT*s of a particular ideology in the newspapers. Thus a newspaper close to the PP will use more *IT-PP* than another newspaper.

Figure 3 shows the percentage of *IT-PP* distributed by newspapers. A little more than 40% of the appearances of *IT-PP* occur in *ABC*. It is followed by the newspapers infoLibre and El Confidencial. Thus the newspaper closest to the PP would be *ABC*, while 20 Minutos would be the one furthest away. But what happens if we look at the correspondence between *IT-Podemos* and the same newspapers?

Figure 4 shows that *ABC* is also the newspaper with more *IT-Podemos*, albeit less acutely than in the previous example. The relative distribution of the rest of the newspapers is very similar to the previous example. These results are overly far from the assessment of the public and experts to be reliable. Hence it does not seem that the frequency distribution of the *IT* by ideology makes it possible to identify clear alignments between politicians' tweets and newspapers. The appropriation by Podemos of frames derived from *IT*s originally from the right (and more present in newspapers which are presumably more right-wing) could be behind such counter-intuitive results.

### 4.3 Correspondence between tweets and newspapers by similarity of frames

The next step was to check whether the vectors for the *IT*s in the tweets and the vectors for the frames of these *IT*s are similar. For example, we wanted to see if the newspapers closer to Podemos tended to link the EU and Angela Merkel with austerity more often than the newspapers closer to the PP.

As we had done with the MPs' tweets, we converted each newspaper's noun phrases into vectors whose dimensions were the *t* terms; i.e. the most semantically related terms obtained with Word2vec. We compared the *IT-PP* and *IT-Podemos* vectors via cosine similarity with the vectors of the same noun phrases of the newspapers. We found that the referentiality to the *IT*s of the ideologically opposite party was also a characteristic of the newspapers, so we obtained results similar to those for *IT* frequency.

### 4.4 Correspondence between tweets and newspapers by focus on the *t*

The last option explored focused on the *t* terms and their ability to interact with the *IT*s of a different ideology. In terms of frames, this means that with a given *IT*, newspapers which are close to a party will coincide when talking about the same *t* terms.

To verify this we gathered the *t* terms semantically related to the *IT* of the PP and Podemos tweets. We then checked how many *t* terms of each party appeared in the news items of a newspaper and for each *IT* we created a vector with the number of *t* terms of the PP and Podemos co-occurring for each newspaper.[11] Table 2 illustrates these vectors with the $t_{pp}$ related to *centrality*, *abyss*, *ponytail* and *populism*. For example, 'centrality' and 'ponytail' has 19 and 1 $t_{pp}$ co-occurring in the newspaper *ABC* respectively, but no $t_{pp}$ in infoLibre. "Populism" has two $t_{pp}$ in *ABC* and one in El Confidencial, but none in 20 Minutos or Huffington Post.

Once the vectors for each newspaper had been created, we took the newspaper in which PP frames appear most as the benchmark: *ABC*. The incidence of PP *t* terms in the rest of newspapers is represented in relation to this newspaper, which has the value 1.

As can be seen in Figure 5, El Confidencial is the newspaper closest to *ABC* in terms of the frequency of appearance of *IT* with PP *t* terms. 20 Minutos, the Huffington Post and infoLibre are further away, with 20 Minutos the most distant. With this

**Table 2. Co-occurrence for *t* term vectors for a series of *IT-PP* in the newspapers analysed**

| *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 |

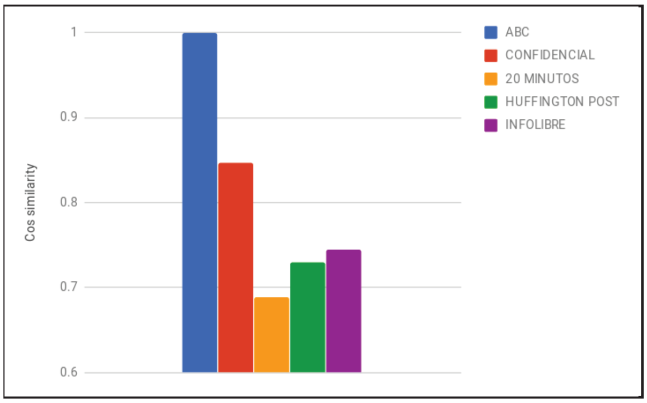
Source: authors.

**Table 3. Co-occurrence vector of the 3 *t* terms associated with the Podemos pro-ETA IT**

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

Source: authors.

**Figure 5. Proximity of newspapers with respect to *ABC* in terms of PP frames**



Source: authors.

system - vectors of co-occurrence in the newspapers - we might in principle find “fake” left-wing ITs. For example, Table 3 shows the vector for *pro-ETA*, an IT which it should be recalled is used more frequently in Podemos’s discourse than in the PP’s.

*Pro-ETA* has 3 *t* terms with which it jointly appears in a single newspaper, namely *ABC*. Bearing in mind that these are *Otegui*, *Bildu* and *ETA*, it should be considered whether the co-appearance of an IT with some particular *t* terms in a newspaper already ideologically aligned (as we have already done with *ABC*) is a criterion for ideologically (re)classifying an IT even though it is widely used by the ideologically opposed party. At all events this procedure could be used to “clean up” ITs wrongly classified as left or right and seems a possible solution to the problem of the appropriation of frames by the opposing party as a tool to stir up conflict, point out paradoxes in its opponents, etc.

## 5. Conclusions

Measuring the bias of written digital media is essential because we need to know the scope and meaning of its impact in order to assess the plurality of the information landscape and improve the public’s digital literacy.

Our review of the literature about measuring the ideological bias of the media has shown that the various methods used to date have a number of limitations. Attributing the ideology of its audience to each media outlet assumes that the public are aware of the media’s bias and selectively expose themselves, yet neither the first nor the second hypothesis are always true. The second approach uses published content in three possible variants. The first is limited to a small amount of text which is highly indicative of content (editorials), the second is to detect linguistic patterns by machine and the last consists of combining these machine procedures with human coding. Using editorials tends to present a more extreme ideology than the one the media outlet really has while the last strategy is quite expensive in terms of resources. We have thus chosen the second one.

However, our approach includes three new features. Firstly, our unit of analysis is not a list of ideologically charged words or phrases but rather a set of connected ideologically charged noun phrases. Secondly, the measurement we use to assign an ideology is based not only on the frequency of use of these word chains but most of all on the discrepancy between them. The last innovative aspect lies in the text corpus we use as a reference to identify ideological frames, namely tweets from political leaders on Twitter and not electoral programmes or parliamentary speeches.

To identify ideologically charged content we have focused on the frames (sets of semantically close words around an IT) which are typical of the two most polarised state-wide parties according to the perceptions of Spanish public opinion: the PP and Podemos. We have identified a series of terms common to both parties but more present in the tweets of the MPs of one party than the other. We have verified that the *t* terms accompanying them are quite different before identifying frames.

During this process we encountered several dead ends. One of them was counting the correspondences of the frames of each party with newspapers, probably due to the appropriation by

Podemos of frames used by the right to criticise them. Similarly, comparing the distance between party and newspaper frames leads us to the same point: the results seem to make sense if we look only at the PP's frames and the similarities between this party and the media, but this is not the case with Podemos. Forthcoming developments should seek to solve the problem of ironic references to the opponent's interpretation frameworks. This has been previously mentioned as one of the main problems of content analysis using machine learning to attribute an ideology to the media (Barberá and Sood 2014). Our data confirm that Podemos references right-wing criticisms of its "populist" attitudes and arguments to make fun of them which means it is impossible to identify their intentionality by machine. Another option would be to add a time dimension to give more weight to the terms which appear first in time as factors that identify a party's frame. Otherwise, this measurement error could be tempered by expanding the reference corpus to the rest of state-wide parties. Thus this typical Podemos phenomenon would be diluted among Socialist party tweets. Finally, machine learning could be combined with human coding. Although more expensive, this strategy would enable us to discard terms used ironically or sarcastically.

## Notes

1. Media literacy means the development of reasoned and critical understanding of the nature of the media and their effects, how they create meaning and how they organise their own reality (Gilster 1997, Aparici 1996).
2. These surveys were conducted between 2015 and 2016. The first is by the eGovernance Research Group: electronic government and democracy (GADE) at the Universitat Oberta de Catalunya (UOC) carried out by the Opinionet project. The second is a survey by the Democracy, Elections and Citizenship (DEC) research group at the Universitat Autònoma de Barcelona (UAB). The third is also a survey by the GADE group which was answered by UOC students.
3. This package format corresponds almost perfectly with the typical interpretation frames in semantic analysis.
4. After a number of exploratory tests, we rejected parliamentary speeches as it was not possible to build a large enough text corpus to extract ideologically charged terms or sets of terms. We also decided to dispense with electoral programmes because our preliminary analysis did not identify any significant discrepancies in the parties' frames based on their electoral programmes. Furthermore, electoral programmes (and the coding proposed by the Party Manifesto Project) are no longer used to assess the ideological positions of the parties (Benoit and Laver 2006; 2007). Finally, the parties' programmes use very formal language that is somewhat removed from the more informal and ideologically charged language employed in the media.
5. Only 296 of the 350 MPs have an active Twitter account.
6. Some people, especially in the PP, were MPs in 2009, but only a few members of Izquierda Unida who became part of the Unidos Podemos coalition had been before 2016. However, we think that they were sending messages and values consonant with this party in their tweets before this date.
7. Here we see *semantic proximity* as co-occurrence, or appearing in positions adjacent to the text. It is a concept in quantitative text analysis. The algorithm used to determine it (Word2vec) collates this physical proximity of words while maintaining the grammatical properties of the texts from which they are drawn.
8. <https://www.clips.uantwerpen.be/pages/pattern-es> Together with verbs the noun phrase is the basic element that structures a sentence, the main seat of lexical meaning and, in a nutshell, the way in which concepts are named. Thus we can gather names such as the *High Court of Justice* instead of the bigram "High Court" or the monograms "Court", "High" and "Justice".
9. Word2vec is a method representative of the latest trend in machine learning called Deep Learning with a structure of neural networks (Dikolov *et al*. 2013). It is a method that is being used with great success in machine translation (Mikolov, Quoc, Sutskever 2013), feeling analysis (Acosta, et al., 2017) and document classification (Lilleberg, Zhu, Zhang 2015). Even the abstraction of the idea of context, defined in a vector space, has encouraged the appearance of other applications as recommenders (Ozsoy 2016).
10. Word2vec uses an algorithm which calculates the closest nominal syntax for each noun phrase. Proximity is a value that ranges from 0 to 1 (from furthest away to nearest). In this project we have considered as *t* terms ones that exceed the value of the median (0.5).
11. We used Normalized Google Distance (NGD) to measure co-occurrence with a range of values between 0 (no proximity) and 1 (maximum proximity). It is a measure of semantic distance according to the degree of co-occurrence of two terms, in our case between the IT and its *t*, the headline and the body of the news item.

## References

Acosta, J.; Lamaute, N.; Luo, M.; Finkelstein, E.; Cotoranu, A. *Proceedings of Student-Faculty Research Day*, CSIS, Pace University, 5 May, 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". Unpublished manuscript, 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*. London: 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.

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

Council of Europe: "4ème Conférence ministérelle Europeenne sur la politique des communications de masse. Les media dans une société démocratique". Prague, 7-8. Rapport d'activité du Comité d'experts sur les concentrations des media et le pluralism. MCM (94)5. Strasbourg: Council of Europe, 1994, p. 8.

Downs, A. *An economic theory of democracy*. New 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". In: Berkowitz L. (ed.). *Advances in Experimental Social Psychology*. Vol. 2. New York: Academic Press, 1965, 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*. New 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 [United States]: 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*. New 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, July, 2015.

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

Mikolov, T.; Quoc, V. Le.; 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) 1031–1053 (2005).

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, 13 November 2013, Madison, Wisconsin.

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>