RESEARCH

A quantitative analysis of Media Agenda and Public Opinion using time-evolving topic distribution

Sebastián Pinto[1,2*], Federico Albanese[3], Claudio O Dorso[1,2] and Pablo Balenzuela[1,2]

*Correspondence: spinto@df.uba.ar
[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, AR
[2]Instituto de Física de Buenos Aires (IFIBA), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, AR
Full list of author information is available at the end of the article
†Equal contributor

Abstract

The mass media play a fundamental role in the formation of public opinion, either by defining the topics of discussion or by making a degree of emphasis on certain issues. Directly or indirectly, people get informed by consuming news from the media. But which is the dynamics of the agenda and how the people become interested in their different topics? The agenda-setting theory provides a conceptual framework in order to understand the role played by the mass media in public opinion formation, but the previous questions can not be answered without proper quantitative measures of agenda dynamics and public attention. In this work we study the agenda of Argentinian newspapers in comparison with public interests through a quantitative approach by performing topic detection over the news, identifying the main topics covered and their evolution over time. We measure agenda diversity as a function of time using the Shannon entropy and differences between agendas using the Jensen-Shannon distance. We found that the Public Agenda is less diverse than the Media Agenda, and we were also capable to detect periods of time where coverage of certain issues are biased (coverage bias).

Keywords: agenda-setting; opinion formation; topic detection; mass media influence

Introduction

One of the challenges in complex social system research is to understand the ecosystem of information flow and opinion formation. Within this framework, many ingredients can be parts of different mechanisms of social influence. A major role in this ecosystem is played by the mass media outlets, which is used to be the source of information for many people. Then, informed people tend to interact with each other via personal interactions or through social networks, giving rise to a complex dynamics where opinions are shaping and changing in time. In this scenario to understand the role of the mass media influence in a given social group becomes essential.

Naturally, how the media affects public opinion was first explored in the area of social sciences. In the seminal study performed in Chapel Hill during the US presidential elections in 1968 [1], Maxwell McCombs and Donald Shaw found that the aspects of public affairs that are prominent in the news become prominent in the public. This work is considered the founding of the agenda-setting theory. In its basic stage, known as first-level agenda-setting [2], the theory focuses on the comparison between the topics coverage by the media and the public agenda: The topics that the public consider as priority. For instance, within the agenda-setting framework, it was explored how media content correlate with audiences of different ages [3] and how people agendas differ based on the way they consume news [4]. On the other hand, the theory hypothesizes how the media affect the audience opinion, for instance, how political coverage and political advertisement shape candidate knowledge among the audience [5, 6], or how the coverage given by the media to a particular nation affects people perception about its importance to local political interests [7]. Related to our work, in Argentina, [8] examines the differences between public and journalists preferences, and other works like [9, 10], studied the coverage of the main newspapers on particular events related to a confrontation scenario between government and press [11].

The Chapel Hill investigation also induced other several research directions [2]. One of them focus on detecting bias in the media, either by taking into account the number of mentions related to a preferred political party [12, 13] or by identifying the ideology through the position of the media regards to certain issues or actors [14, 15]. This research line can be linked with the theory of *framing* [16, 17], which focus on the way the media emphasizes some attributes of an object, while understating others. Other investigations pay attention to the sources of the media agenda, theory known as *intermedia agenda-setting* [18, 19, 20], where the competition and the mutual influence between different media is observed.

From the irruption of internet, quantitative analysis based on the access to big data were possible and, for instance, take into account the temporal dependence of the media and public attention. In [21] it is shown that the newspapers and Twitter have an opposite reaction to the changes of the unemployment rates; in [16], the competition of frames about gun control is explored; in [22], the authors show how fluctuations of Twitter activity in different regions depend on the location of terrorist attacks; and in [23], the complex interplay between the social media and the traditional one is followed over time on a set of predefined, but general, issues. However the works cited above have performed a dynamical analysis of agenda-setting based on a single issue or on a set of predefined issues. These are usually selected by the researcher and use to reflect general subjects, such as “health” or “gun control”.

The question about what issues should be selected can be solved by an useful tool employed in the analysis of large documents corpus, but understated in the agenda-setting framework and its derivatives: Unsupervised topic modeling. It is an alternative to the dictionary-based analysis, which is the most popular automated analysis approach [24], and allows to work with a corpus without a prior knowledge, letting the topics emerge from the data. Although many works employ unsupervised topic modeling on news corpus, much of them emphasize the performance of the topic model over a labeled corpus, focusing on the proper detection of the topics [25, 26, 27]. In general, issues about the temporal profile of topics are embedded in the context of topic tracking [28, 29], or in the recognition of emerging topics in real-time [30], mostly applied to social media.

In this work we propose a novel method in order to study the dynamics of the mass media agenda, which consist of performing an unsupervised topic model on newspapers articles, and studying how the emerging topics evolve in time. We look also at their correspondence with the audience agenda, by looking at the Google searches and Twitter in the studied period. Rather than focus on a single issue or on a set of independent topics, this method allows us to define the agendas (both the media and the public) as an object which evolves in time. Our work focus on a quantitative approach which complements the agenda-setting theory describe above, which mainly stands within the framework of first-level agenda setting, but also allows us to face agenda bias and framing.

On the other hand, in recent years new approaches to study social dynamics through several tools taken from statistical physics have proposed mathematical models to explore the interplay between the mass media and society [31, 32, 33, 34, 35]. Much of them lack in being contrasted with real data. With our work we aim to gain a closer insight on the complex interaction between the media and the public, and provide a quantitative research that would be useful at the time of constructing better and more data-driven models.

Materials and Methods

The Media Agenda

In this work we analyze a three-month period of the Argentinian media agenda composed by a corpus of news articles that were published between July 31st, 2017 and November 5th, 2017. The articles come from the politics section of the online editions of the Argentinian newspapers *Clarín*, *La Nación*, *Página12*, and the news portal, *Infobae*. The first two lead the sale of printed editions in *Buenos Aires* city, but *Clarín* reaches roughly two times the readers of *La Nación*, and ten times the readers of *Página 12* [36]. On the other hand, *Infobae* has the most visited website, much more than *Clarín* and *La Nación* [37]. The corpus analyzed is made up by 2908 politics articles of *Clarín*, 3565 of *La Nación*, 3324 of *Página 12*, and 2018 of *Infobae*. Except *Página 12*, all articles were taken from the section *Política* (Politics) of the respective news portals, while the articles which belong to *Página 12* were taken from the section *El país* (The country).

The articles are described as numerical vectors through the *term frequency - inverse document frequency (tf-idf)* representation [38]. Given the set of terms contained in the corpus words, after removing non-informative ones such as prepositions and conjunctions, the *tf-idf* algorithm represents the i-document as a vector $v_i = [x_{i1}, x_{i2}, ..., x_{it}]$, where the component x_{ij} is computed by the eq.(1), where tf_{ij} is the number of times the j-term appears in the i-document; d is the number of documents in the corpus; and n_j is the number of documents where the j-term appears. Each vector is then normalized to unit Euclidean length. Once the document vectors are constructed, we put them together in a document-term matrix (M), which has dimensions of number of documents in the corpus (d) by number of terms (t).

$$x_{ij} = \mathrm{tf}_{ij} \cdot \mathrm{idf}_j = \mathrm{tf}_{ij} \cdot [1 + \log(\frac{1+d}{1+n_j})] \tag{1}$$

In order to detect the main topics in the corpus, we perform *non-negative matrix factorization (NMF)* [38, 39] on the document-term matrix (M). A topic is defined to be a group of similar articles which roughly talks about the same subject. *NMF* is an unsupervised topic model which factorizes the matrix M into two matrices W and H with the property that all three matrices have no negative elements (see eq.(2)). This non-negativity makes the resulting matrices easier to inspect, and very suitable for topic detection [1].

$$M^{(d\times t)} \sim H^{(d\times k)} \cdot W^{(k\times t)} \tag{2}$$

Such as the resulting matrix H has dimensions of number of documents by k and matrix W has dimensions of k per number of terms, the number k is therefore interpreted as the number of topics in the documents and it is a parameter that must be set before the factorization. In this work, we arbitrarily set $k = 10$. We found this value to be a suitable one based on our knowledge of the corpus. Since the factorization of eq.(2) usually can not be made exactly, it is approximated by minimizing the reconstruction error, i.e. the distance between matrix M and its approximated form $\tilde{M} = H \cdot W$. The *NMF* factorization was made through the python module *scikit-learn* [40].

The matrix H is the representation of the documents in the topic space. We normalized its rows to unit l_1-norm in order to view their components as a degree of membership of a given document in the set of topics. In particular, the index of the largest component tells us which is the most representative topic of the document. On the other hand, W gives the topics representation in the original term space. The largest components of a row give the most representative words of each topic, which we call *keywords*, and therefore an insight of what the topic is talking about.

After performing NMF, we represent the time-dependent Media Agenda as a time evolving distribution of topics. We define $W_i(day)$ to be the daily weight of the topic i, which is calculated following the eq.(3), where $l(j)$ is the number of words of the document j; h_{ji} is the degree of membership of document j on topic i; d_j is the date of document j; and δ is the Kronecker delta. Providing by the fact that each document vector can have all non-zero components, it is allowed that a document contributes to more than one topic weights. In order to reduce noise, we apply a linear filter with a three day wide sliding window, and finally we normalize the temporal profiles in order to describe each newspaper agenda as a distribution over the topics' space, which evolves over time.

$$W_i(day) = \sum_j l(j) \cdot h_{ji} \cdot \delta_{d_j,day} \tag{3}$$

The Google and Twitter Agendas

The other side of the Media Agenda is to have some measure of the public interests about its topics, by finally making up what we call the Public Agenda. To achieve

[1] Even though there are other techniques in topic detection, as for instance LDA (Latent Dirichlet Allocation), the NMF decomposition suites perfect for the kind of corpus we have analyzed in this work (A detailed comparison of both, NMF and LDA methods could be found in **Supplementary Material**).

this goal, we take Google and Twitter as proxies of the public interests by looking for the same topics in the same period of time. We take advantage of the topic keywords in order to make queries into the *Google Trends* tool and into the advance search tool of Twitter, and therefore to get the relative weight of searches and tweets in each respective platform. Therefore we describe the Public Agenda by describing, in an independent way, both the Google and Twitter Agendas. In section **Results** we will give a more detailed description of the keywords involved in their construction.

Normalized Shannon Entropy (H)

In fact, one of the key of our work is the representation of the Agendas as time evolving topic distributions, and the association between the different measures that we can derive from them with the social events observed during the period. In order to do that, the first measure we emphasize is the concept of *diversity of the attention*. *Diversity* is a very important variable that must be taken into account when dealing with multiple issues [41], due to the fact that it tells us how the attention is distributed across the different topics of discussion. As was proposed in [41], we take the normalized Shannon entropy as a very suitable way to quantify the diversity within our framework.

The normalized Shannon entropy $H[p]$ referred in eq.(4) gives us a measure of how spread is a -discrete- distribution, taking the maximum value of 1 when all outcomes are equally probable, and 0 when there is just one possible outcome. This last case, or an approximated one, will be of particular interest for us because it tells about a topic which absorbs all the attention either from the Media or the Public, while the former shows an absolutely unbiased interest towards any of the topics.

$$H[p] = \frac{-\sum_{i=1}^{N} p(x_i) * ln(p(x_i))}{ln(N)} \tag{4}$$

Jensen-Shannon distance

While the diversity is a property of each distribution, a natural question that arises from comparing different distributions is how similar they are. For instance, we will be particularly interested in measuring the similarity between the Media and Public Agendas, because those dates when the similarity is low tell us about distant interests of the media and its audience. We proposed here to measure the similarity between distributions by the use of the Jensen-Shannon distance (JSD). When the similarity of the distributions is low, the distance between them is high.

The Jensen-Shannon distance (JSD) is a metric between distributions based on the Jensen-Shannon divergence (JS_{Div}) [42], which is in turn a symmetric version of the well-known Kullback-Leibler divergence D_{KL} (eq.(5)). We recall that the JSD has the advantage of being symmetric and also a well-defined distance, which makes it conceptually easier to deal with. As can be seen in eq.(6), the JSD between the distributions P and Q is simply the square root of the Jensen-Shannon divergence, where $M = \frac{P+Q}{2}$.

$$D_{KL}(P||Q) = -\sum P(i) log(\frac{Q(i)}{P(i)}) \tag{5}$$

$$JSD(P,Q) = \sqrt{JS_{Div}(P,Q)} = \sqrt{\frac{1}{2}[D_{KL}(P||M) + D_{KL}(Q||M)]} \tag{6}$$

Outliers identification

As was mentioned above, once the concepts of diversity and distance between distributions were defined, we are particularly interested in those dates when these concepts take extreme values: When the diversity is low, we know that a topic is absorbing much of the attention of any of the Agendas; on the other hand, when the distance between two distributions is high, we can conclude that they have distant interests. However, we still lack the definition of what a low diversity or a high distance are. We face this problem by treating the measures of each observable as random samples from a population with unknown distribution, and identifying those extreme values as outliers of the distribution. In order to detect these outliers we follow the popular box-plot construction proposed by Tukey [43], which is a simple data-driven method and has the advantage of making no prior assumption about the distribution of the data. However, it is important to remark that the constants involved in the outliers definition (see eq.(7)) is taken from applying this method on a normal distribution.

In the box-plot construction a quartile division of the N observations is proposed. We name $Q1$ as the lower quartile, $Q2$ the median of the distribution, and $Q3$ the upper quartile. Recall that $Q1$ ($Q3$) is defined to be the division where the 25th (75th) percent of the observations lies below (by definition, the median $Q2$ separates the distribution in two equal parts). On the other hand, the inter-quartile range IQ is defined to be $IQ = Q3 - Q1$. This is the range where the bulk of the data lies inside. We are not interesting in the visualization of the box-plot in its own but instead in its procedure to identify outliers. Therefore, from the identification of the quartiles, new quantities called *fences* are defined in eq.(7): The *lower inner fence* (LIF), the *upper inner fence* (UIF), the *lower outer fence* (LOF), and the *upper outer fence* (UOF). The fences can be interpreted as the limits of the distribution.

$$\begin{aligned} LIF &= Q1 - 1.5IQ \\ UIF &= Q3 + 1.5IQ \\ LOF &= Q1 - 3IQ \\ UOF &= Q3 + 3IQ \end{aligned} \tag{7}$$

We then have all the ingredients to label a point as an outlier: A point which lies above the upper inner fence is considered a *mild outlier*, while a point that lies above the upper outer fence is considered an *extreme outlier*. The same holds for the lower fences, i.e. if a point lies below the lower inner (outer) fences is considered as a mild (extreme) outlier [44]. We will indicate the proper fences in each figure either when the diversity or the distance is being analyzed. We will pay attention not only to those values labeled as outliers, but also to those that are next to any of the fences despite not being strictly defined as that.

Results

We initially focus on the ten most important issues from the three-month period corpus of news reported above. These ten topics are represented in the wordclouds of figure 1. Given our interpretation of the keywords found in three of them, we joined these topics as being part of the same macro-topic which we called *Elections*. On the other hand, the same holds for other two topics which were classified as part of a macro-topic called *Missing person*. Therefore, the ten original topics were reduced to seven, which are pointed out in the radar plot of figure 1. The meaning of the topics or macro-topics is contextualized in the **Supplementary Material**.

Finally, by following the procedure described in the previous section, we construct the **Media Agenda (MA)** and the **Public Agenda (PA)**, in both its Google and Twitter derivations, as time evolving distributions in a space of seven topics.

The Media and Public Agendas

In figure 1 we show a seven topic decomposition of the whole corpus using radar plots for the Media **MA** and Public agendas **PA** discriminated by **GT** (Google Trends) and **Tw** (Twitter). In this figure we also show the wordclouds of the keywords that define each of the ten original topics, where the size of the word reflects its importance in the topic definition. In green color, we point out the words involved in the Google Trends and Twitter queries in order to construct the Public Agenda. The queries employed are also specified in tables 1 and 2. On the other hand, in table 3 we show the linear correlation between the topic temporal profiles from the Public Agenda and their counterparts in the Media Agenda.

We can observe that both GT and Tw look similar in this representation, but they show specific differences with the Media Agenda. For instance, a greater interest of the audience in the topic *Missing person* than the media is observed, or inversely, a lower interest in the topic *Prosecutor's death* takes place. However, this static representation is not able to show the complex dynamics of the agendas evolution and the importance of punctual and specific facts which can erase or amplify their differences.

This can be observed in figure 2 where the time evolution of the topics is shown in a bump chart of the Agendas. The bump chart provides a clear visualization of the relative weight of the topics at the same time with their ranking. In this figure we also highlight some important events related to the dynamics of the topics. It is possible to appreciate how the main topic changes in time and has a glance of the qualitative differences between the agendas. In particular, it can be observed some differences between the Public Agendas that were not observed in the previous figure, as for instance, the persistence of main topics is longer in Twitter than in Google Trends. This is more evident at the end of the analyzed period, where the topics discussed in Google Trends show more response to change in Media Agenda than in Twitter, maybe due to the existence of a different pattern of interaction in the social network, to which a deeper analysis could be devoted in future works.

The linear correlations between the same topics of MA and PA were also calculated. In all cases, we found that the correlations are positive and statistically significantly, as it is shown in table 3. We interpret this as a validation of the topics found in the corpus and the keywords that describe it. Even though we are particularly interested in those periods where the Agendas differ, it is expected that the media and public interests should generally follow a similar a pattern, mainly driven by external events A non positively (or a non significantly) correlation may imply, besides the obvious conclusion of agendas disengagement, that we can be eventually failing to properly detect the keywords or features that describe a particular topic, and therefore the comparison of the Google Trends or Twitter patterns with their counterpart in the Media Agenda would be wrong.

A quantitative description of the Agendas

Agenda diversity

How dominant is a main topic? Is the degree of dominance of a given topic in the Media Agenda reflected in the Public Agenda? In order to answer such kind of questions we quantify the diversity of the agendas through the normalized Shannonentropy H, which was introduced in section **Material and Methods**.

In figure 3 we can see the value of H as a function of time for the three agendas. It is important to pay attention to those periods of time when the diversity is lower than usual. This effect is notoriously more pronounced in the Public Agenda giving by GT, and in particular in four specific days when four local minimums of the Shannon entropy can be detected. Three of them are outliers as defined in section **Material and Methods**, two of them from GT and one from Tw. The other one has been not identified as an outlier but it is a pronounced minimum and therefore a point of interest in our description.

A lower value in the agenda diversity is due to the fact that the most important topic attracts practically all the attention of the public and the media, collapsing the agenda to one of the issues involved. In the radar plots included in figure 3 we can see how two of these outliers (**a** and **d**) belong to the topic *Elections*. They are related to the primary and general legislative elections that took place in August 13th and October 22nd respectively. In all the agendas these points were detected as outliers except point (d) in Twitter Agenda. Why is that? The radar plot of the Twitter agenda for this day displays an association between the topic *Elections* and the *Current President*, decreasing the importance of this topic. Discussions in Twitter about elections appear also in point (c), when the other agendas seem to be more diverse. On the other hand, and despite not being classified as outlier, we also focus in point (b) because the Shannon Entropy in the Google Agenda displays a minimum (collapsing agenda) which is not corresponded neither in the Media nor in the Twitter Agendas. Crawling in the context, we see that it belongs to the topic *Missing person* and this date corresponds to the rally that took place one month after the disappearance of *Santiago Maldonado* (see **Supplementary Material**). We would like to emphasize the discussion about this topic (*Missing person*) because its dynamics show interesting features, as we will show below.

From the measure of H we have also observed that the median of the Public Agenda diversity is statistical significant lower than the observed in the Media Agenda. Specifically $H_{GT} = 0.73$ and $H_{Tw} = 0.74$ are statistically significantly lower than $H_{MA} = 0.85$ with $p < 10^{-18}$, while there is no significant difference between the first two. However, from figure 3 we can see that GT shows more abrupt dropouts in the diversity in response to specific events. From all this analysis we can conclude that given a finite set of topics, **the Public Agenda is less diverse than the Media Agenda**, because the public seems to focus on the most important topics than the media can do, maybe due to editorial decisions.

Distance between Media and Public Agendas

Given our descriptions of the agendas as time-evolving distributions, we can compare them by computing the Jensen-Shannon distance. In this context, outliers in selected dates will correspond to divergences between the Media and Public Agenda: Specific events when the public interests do not match with media offer. In figure 4 we show the Jensen-Shannon distance between Media and Public Agendas as a function of time. We focus in three points that seem to be relevant enough. In all cases, the topic distributions at that particular dates displayed by the radar plots show that the increment in the distance between agendas is due to a greater interest of public opinion in the *Missing person* topic.

Points **(c)** and **(d)** show that both the public and the media highlight this topic, but the media do not disregard other topics, so the corresponding distance between them can be interpreted as lack of diversity in Public Agenda as discussed in the last section.

On the other hand, points **(a)** (we take this point due to be a local maximum despite not being an outlier) and **(b)** show a major interest of the public in the topic *Missing person* which it is not reflected in the Media. In figure 2 we can see that this topic becomes the most important in public interests (both in GT and Tw) days before that it happens in the Media Agenda. This fact can be associated with a social networks (like Facebook and Twitter) campaign in favor of the appearance of *Santiago Maldonado* ("The missing person") that took place on August 26th. This campaign was massive and initially underestimated by the main media outlets in Argentina (see **Supplementary Material**).

Finally, it is important to say that the Jensen-Shannon distance, in conjunction with the measurement of agenda diversity given by the Shannon entropy, give an insight of independent behavior, in certain particular dates, of the public and the media. Its identification can be a starting point to study the media reaction to a change in audience interests.

Agenda bias in different media outlets

In this section we leave aside the Public Agenda as an unified corpus and we study the composition and evolution of the Media Agenda of each media outlet. In figure 5 we show the bump charts corresponding to each of the analyzed newspapers analogously to figure 2. The topics are the same introduced in the wordclouds of figure 1, but when computing the topic weights, the articles are discriminated by newspaper. We also show the radar plots with the average distribution, as made in figure 1.

In figure 5 we can qualitative have a glance of the differences between the newspaper agendas. For instance, we can see how the newspaper called *Página 12* gave more importance to the topics *Missing person* and *Social leader*, while it reduces to minimum the coverage of the topic *Former Planning minister* as the others did.

In other to detect significant bias coverage of a given newspaper, we again calculate the Jensen-Shannon distance, but between the individual newspapers agenda and the Media Agenda. Note that this is the distance between the distributions of figure 5 and the top panel of figure 2. In figure 6 we show the Jensen-Shannon distance as a function of time. We detect three points as outliers, although we finally disregarded point **(b)** due to a lack of information of newspaper *Infobae* in that period. The other two points corresponds to differences between *Página 12* and the other newspapers and correspond to differences in the coverage of the topic *Missing person*.

Point **(a)** corresponds to the first news the disappearance of Santiago Maldonado, reported by *Página 12* before the primary elections, and point **(c)** corresponds to the two months' rally after the disappearance (see **Supplementary Material**). Another singularity of point **(c)** corresponds to a greater coverage of *Página 12* in the topic *Social leader* while the other media outlets seem to be more interested in the topic *Former Vice-President*.

The greater coverage in the topic *Missing person* by *Página 12* is even more clear if we inspect the temporal profile of the topic and compare the coverage given by each newspaper. A difference in the coverage is what it is called *coverage bias* [45]. In figure 7 we show the temporal profile of the topic *Missing person* (panel (a)) and the topic *Former Planning minister* in panel (b), as an example where the behavior is the opposite, as can be seen below.

From panel (a) of figure 7, we can see the larger coverage of *Página 12* in comparison to other newspapers at the beginning of the period. For example, we can quantify this difference calculating the median of the signals. If we focus in the period between July 31st and August 27th, the median of the topic relative weight in *Página 12* is roughly 0.14 and this is statistically significantly larger ($p < 10^{-7}$) than other medians, which are lower than 0.05. Analyzing the same period, but in panel (b), we again can show that the median in *Página 12*, which is roughly 0.01, is lower than the others, which oscillate around 0.05 ($p < 10^{-3}$). This quantification is proposed as method of studying coverage bias in the context of the methodology implemented in our work.

Finally, in figure 7 we also show the topic keywords wordclouds, highlighting the most frequently mentioned in each newspaper and filtering the common words to all newspapers. Although most of the words are not relevant enough, some of them are quite interesting, as for instance the word *represión* (repression) when *Página 12* talks about the topic *Missing Person* and the word *Cristina* (*Fernández de Kirchner*, former president) which is employed by all newspapers except *Página 12* when they talk about the topic *Former Planning minister* (see **Supplementary Material**). We think that a deeper study of the topic keywords could be a first approximation in the study of framing, which will constitute the core of futures works.

A brief discussion about agenda-setting

In a world where social media exist and the feedback between the media and audience is common currency, nowadays the idea that the Media set the agenda and the audience blindly follows it (as it's seemed to be suggested in the original work of McCombs) is too naive. Based on the data analyzed above, the behavior of the Media and Public Agendas, either by looking at Google Trends or Twitter, shows periods of strong similarity (specially in the presence of an unexpected event) and periods of disengagement. Therefore, it is not trivial to establish a causal relationship between agendas, specially when they are represented as evolving in time topic distribution as we did in this work. However, it is possible to discuss agenda setting if we focus in a single topic. We think that the *Missing person* topic is the most adequate topic to be discussed because:

- It caused a great impact in both the media and the audience
- its coverage fully deploys along the time lapse analyzed in this manuscript (see **Supplementary Material**).

In figure 8 we show the topic relative weight from the Public and Media Agendas. After the initial coverage, the agendas seem to differentiate around August 15th, when topic started to become more important in the Public Agenda than in the Media one. Around August 24th, the topic abruptly increases in the audience interests while the reaction in the media is slower, showing a significant peak in the plot of the discrete difference (panel (b)). This date is very close to August 26th, when a campaign in social media took place. After that, the Media increase its coverage about the topic.

Is this a case of reverse agenda-setting [23], i.e, when the audience set the Media Agenda? After all, the audience gets involved about this topic by the media, so how was the coverage before those events? We can answer these questions by calculating the cumulative coverage of this topic since the first events took place. This measure can be seen as the numerical integration of the temporal profiles of figure 7 between the initial date and the current date. It is interesting to note that the newspaper responsible for accumulating coverage during the initial stage was the minority one: *Página 12*. Our interpretation about the setting agenda dynamics of this particular topic is the following: A small newspaper (*Página 12*) gives great coverage to this topic; the topic is then amplified by public in the social networks and it is also expressed by reiterative Google searches. Then, the rest of the Media pays attention to this subject and it becomes an important topic in the Media Agenda. This interpretation tries to catch in a qualitative way how the information flow was in this specific topic. On the other hand, behind this interpretation there are two important facts that must be mentioned: First, the disappearance of a person is a very sensitive theme for the Argentinian society, and second, there are political reasons why *Página 12* was particularly interested in cover this topic while the other media did not follow this interest (see **Supplementary Material**).

Even tough we face the question about causality only in a qualitative way and just for a specific topic, it was possible to highlight the complex feedback dynamics that take place between public and media agendas.

Conclusions

The mass media play a fundamental role in opinion formation and therefore, it's of vital importance to have an accurate quantitative description of the Media and Public Agenda and their relationship in the framework of agenda-setting theory. In this work, through the implementation of a topic detection algorithm we describe the Media Agenda as a distribution which evolves in time and which is defined in a topic space which emerges from the analysis of the corpus. This gave us an insight of how we can construct and follow the audience interests, i.e the Public Agenda, in order to compare with the media interests.

Given the Agendas, we found that the Public one is usually less diverse than the Media, showing that when there is a very attractive topic, the audience focuses on this one, meanwhile the media keep certain degree of diversity. On the other hand, the measurement of distances between agendas can be employed to rapidly detect those periods when the public may have an independent behavior respect to the media. The methodology implemented here also allowed us to detect coverage bias in newspapers and gave us a first approximation in the theory of framing.

We hope that some of the elements studied here will give us insights at the time of proposing a mathematical model about the interaction between mass media and audience. Future works may include a more systematic study and its extension to international media, a deeper study of framing through topic detection and sentiment analysis, and a more quantitative analysis about causality.

Supplementary Material

Context

We provide here a more detailed explanation of the topics discussed within this work. The topics belong to the period between July 31th and November 5th, 2017. The ideology of the media in Argentina expresses the highly polarized political climate observed in Argentinian society. During the administration of *Cristina Fernández de Kirchner* (2007-2015), the government maintained a conflict with several news organizations. It led the media such as *Clarín*, *La Nación* and the news portal *Infobae* to be very critics of the *Fernández*'s administration, emphasizing the allegations of corruption related to it, as can be seen in the importance given to the topics *Former Planning minister* and *Former Vice-President*. On the other hand, *Página 12* has an opposite ideological inclination, supporting the former administration of *Cristina Kirchner* and therefore being very critical with the current *Mauricio Macri*'s administration, doing special emphasis on issues related to human rights, as can be again observed in the coverage given to the topics *Social leader* and *Missing person*.

Elections

Two legislative elections were celebrated during the period in great part of Argentina: Primary elections on August 13th and the general elections on October 22nd, 2017. A special focus was put on the elections in the Buenos Aires province, where the former President *Cristina Fernández de Kirchner* participated as a senator candidate representing the alliance *Unidad Ciudadana*, confronting *Cambiemos*, which is the alliance of the current President *Mauricio Macri* and the current governor of Buenos Aires province *Maria Eugenia Vidal*.

Current President

Mauricio Macri is the current Argentinian President since December 2015. Most of the articles in political sections are logically devoted to him under different contexts. However, it is important to point out that during the analyzed period, and specially after the general elections of October 22nd, 2017, a controversial labour reform promoted by the government was being discussed.

Missing Person

Santiago Maldonado vanished on August 1st, 2017 after a minor clash between the Gendarmerie (Border Guards) and a group of Mapuches (Patagonian native population), which recognize themselves as the original population of an area in the Patagonia. Since that event, the *Mauricio Macri*'s administration was accused by several people as the responsible for a **forced disappearance**.

A very massive campaign in social media took place on August 26th, 2017 under the motto “Where is Santiago Maldonado?”, followed by two massive protest marches to the *Plaza de Mayo* that took place on September 1st and October 1st, of which the first one had a great repercussion due to several incidents that took place during the march.

The body of *Santiago Maldonado* was found dead on October 17th, 2017 in the *Chubut* river, near the place where he was seen the last time, and the autopsy report told that *Santiago Maldonado* had died from “asphyxia after being submerged”, with no injuries on his body. However, the responsibility of the current administration is still being discussed.

Former Planning minister *and* Former Vice-President
Julio de Vido was the Planning minister during the administration of *Néstor Kirchner* and *Cristina Fernández de Kirchner* (2003-2015). In 2015, he was elected to integrate the Chamber of Deputies, which finally voted to strip *De Vido* of his congressional immunity over corruption allegations and was immediatly jailed on October 27th, 2017.

Amado Boudou was the Vice-President of the *Cristina Kirchner*'s administration. *Boudou* was arrested on November 3rd, 2017 on charges including money-laundering and hiding undeclared assets.

Social leader *and* Prosecutor's death
Milagro Sala is an indigenous leader. She has been incarcerated under pre-trial detention ever since she was first detained in January 2016. She faces allegations of embezzlement related to government funding for housing projects managed by Túpac Amaru, her social organization. Sala accused the government of "violating her human rights", and several people think that she is a political prisoner of the *Mauricio Macri*'s administration.

Alberto Nisman was a special prosecutor who were investigating the 1994 terror attack on the Argentine Israeli Mutual Association (AMIA), until his suspicious death in January 2015. During the period analyzed in this work, a team of experts led by the Gendarmerie (Border Guard) concluded that late prosecutor's death may have been a case of murder, not suicide.

Comparison between NMF and LDA

In this section we apply other topic model, Latent Dirichlet Allocation [46] (LDA), to our corpus and compare its results to the shown in this paper. Due to the increasingly use of LDA, we think that a few words about the performance of LDA in our work is necessary.

Naturally the topics found with LDA may not coincide with the NMF ones. However, one expects that the corpus under study displays some degree of robustness when considering different topic models. On the other hand, as was discussed in [47], NMF can be a more suitable topic modeling method in certain domains, in the way that it produces more coherent topics, while LDA tends to return higher levels of generality and redundancy. Topic coherence is defined as the semantic interpretability of the terms used to describe a particular topic, although the coherence of a topic may depend on the end user's expectations.

We define a simple coherence measure defined in equation 8, where d_{ij} is the number of documents where the term i and term j appear simultaneously, and d_x is the number of documents where appears the term x. The summation is over the N top terms of the topic. It's important to note that if two terms have no co-occurrences, the contribution to the summation is zero, and if these ones appear only together the contribution is one. A topic with higher coherence is a topic where the terms that define it co-occur frequently.

$$TC = \sum_{i<j}^{N} \frac{2d_{ij}}{d_i + d_j} \tag{8}$$

We perform a decomposition into 10 topics using LDA with the python module *gensim* [48], which allows us to modify the number of times the corpus is read, improving the coherence of the topics. Unlike to what we see with NMF, the LDA's performance depends strongly on the initial condition of the algorithm. After 10 iterations, we chose the one with highest mean topic coherence, and compared this with the NMF results.

In figure 9 we show the temporal profiles of topics *Elections* and *Missing Person* for both NMF and LDA. The association between topic models was simply made by looking at the topics which share common keywords. As can be seen from figure 9 and table 4, those LDA topics which can be linked to NMF ones or to a combination of these, show a temporal profile highly correlated.

Nevertheless, LDA returns other topics which can not be directly associated, some of them composed of very general words. By keeping only those topics which can be associated with NMF and re-defining the Media Agenda over this topic space with reduced dimension, we observed similar results by both methods. The same procedure is proposed in absence of an alternative topic model to which make the comparison: Keep only those topics easily interpretable and define the Agendas over this reduced space.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
Text for this section ...

Author details
[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, AR. [2]Instituto de Física de Buenos Aires (IFIBA), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, AR. [3]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, AR.

References

1. McCombs, M.E., Shaw, D.L.: The agenda-setting function of mass media. Public opinion quarterly **36**(2), 176–187 (1972)
2. McCombs, M.: A look at agenda-setting: Past, present and future. Journalism studies **6**(4), 543–557 (2005)
3. Coleman, R., McCombs, M.: The young and agenda-less? exploring age-related differences in agenda setting on the youngest generation, baby boomers, and the civic generation. Journalism & Mass Communication Quarterly **84**(3), 495–508 (2007)
4. Althaus, S.L., Tewksbury, D.: Agenda setting and the "new" news: Patterns of issue importance among readers of the paper and online versions of the new york times. Communication Research **29**(2), 180–207 (2002)
5. Brians, C.L., Wattenberg, M.P.: Campaign issue knowledge and salience: Comparing reception from tv commercials, tv news and newspapers. American Journal of Political Science, 172–193 (1996)
6. Gerber, A.S., Karlan, D., Bergan, D.: Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. American Economic Journal: Applied Economics **1**(2), 35–52 (2009)
7. Wanta, W., Golan, G., Lee, C.: Agenda setting and international news: Media influence on public perceptions of foreign nations. Journalism & Mass Communication Quarterly **81**(2), 364–377 (2004)
8. Mitchelstein, E., Boczkowski, P.J., Wagner, C., Leiva, S.: La brecha de las noticias en argentina: factores contextuales y preferencias de periodistas y público. Palabra Clave **19**(4) (2016)
9. Zunino, E., Aruguete, N.: La cobertura mediática del conflicto campo-gobierno. un estudio de caso. Global Media Journal **7**(14) (2010)
10. Koziner, N., Zunino, E.: La cobertura mediática de la estatización de ypf en la prensa argentina: un análisis comparativo entre los principales diarios del país. Global Media Journal **10**(19) (2013)
11. Mitchelstein, E., Boczkowski, P.J.: Information, interest, and ideology: Explaining the divergent effects of government-media relationships in argentina. International Journal of Communication **11**, 20 (2017)
12. Lazaridou, K., Krestel, R.: Identifying political bias in news articles. Bulletin of the IEEE TCDL **12** (2016)
13. Baumgartner, F.R., Chaqués Bonafont, L.: All news is bad news: Newspaper coverage of political parties in spain. Political Communication **32**(2), 268–291 (2015)

14. Elejalde, E., Ferres, L., Herder, E.: On the nature of real and perceived bias in the mainstream media. PloS one **13**(3), 0193765 (2018)
15. Sagarzazu, I., Mouron, F.: Hugo chavez's polarizing legacy: Chavismo, media, and public opinion in argentina's domestic politics. Revista de Ciencia Política **37**(1) (2017)
16. Guggenheim, L., Jang, S.M., Bae, S.Y., Neuman, W.R.: The dynamics of issue frame competition in traditional and social media. The ANNALS of the American Academy of Political and Social Science **659**(1), 207–224 (2015)
17. Tsur, O., Calacci, D., Lazer, D.: A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In: ACL (1), pp. 1629–1638 (2015)
18. Vargo, C.J., Guo, L.: Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online us news. Journalism & Mass Communication Quarterly, 1077699016679976 (2017)
19. Harder, R.A., Sevenans, J., Van Aelst, P.: Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times. The International Journal of Press/Politics, 1940161217704969 (2017)
20. Guo, L., Vargo, C.J.: Global intermedia agenda setting: A big data analysis of international news flow. Journal of Communication **67**(4), 499–520 (2017)
21. Soroka, S., Daku, M., Hiaeshutter-Rice, D., Guggenheim, L., Pasek, J.: Negativity and positivity biases in economic news coverage: Traditional versus social media. Communication Research, 0093650217725870 (2017)
22. Ali, A.E., Stratmann, T.C., Park, S., Schöning, J., Heuten, W., Boll, S.C.: Measuring, understanding, and classifying news media sympathy on twitter after crisis events. arXiv preprint arXiv:1801.05802 (2018)
23. Russell Neuman, W., Guggenheim, L., Mo Jang, S., Bae, S.Y.: The dynamics of public attention: Agenda-setting theory meets big data. Journal of Communication **64**(2), 193–214 (2014)
24. Guo, L., Vargo, C.J., Pan, Z., Ding, W., Ishwar, P.: Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. Journalism & Mass Communication Quarterly **93**(2), 332–359 (2016)
25. Dai, X.-Y., Chen, Q.-C., Wang, X.-L., Xu, J.: Online topic detection and tracking of financial news based on hierarchical clustering. In: Machine Learning and Cybernetics (ICMLC), 2010 International Conference On, vol. 6, pp. 3341–3346 (2010). IEEE
26. Po, L., Rollo, F., Lado, R.T.: Topic detection in multichannel italian newspapers. In: Semanitic Keyword-based Search on Structured Data Sources, pp. 62–75 (2016). Springer
27. Brun, A., Smaïli, K., Haton, J.-P.: Experiment analysis in newspaper topic detection. In: String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium On, pp. 55–64 (2000). IEEE
28. Hu, X.: News hotspots detection and tracking based on lda topic model. In: Progress in Informatics and Computing (PIC), 2016 International Conference On, pp. 248–252 (2016). IEEE
29. Li, W., Joo, J., Qi, H., Zhu, S.-C.: Joint image-text news topic detection and tracking by multimodal topic and-or graph. IEEE Transactions on Multimedia **19**(2), 367–381 (2017)
30. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, p. 4 (2010). ACM
31. Crokidakis, N.: Effects of mass media on opinion spreading in the sznajd sociophysics model. Physica A: Statistical Mechanics and its Applications **391**(4), 1729–1734 (2012)
32. González-Avella, J.C., Cosenza, M.G., San Miguel, M.: A model for cross-cultural reciprocal interactions through mass media. PloS one **7**(12), 51035 (2012)
33. Moussaïd, M.: Opinion formation and the collective dynamics of risk perception. PLoS One **8**(12), 84592 (2013)
34. Rodríguez, A.H., Moreno, Y.: Effects of mass media action on the axelrod model with social influence. Physical Review E **82**(1), 016111 (2010)
35. Pinto, S., Balenzuela, P., Dorso, C.O.: Setting the agenda: Different strategies of a mass media in a model of cultural dissemination. Physica A: Statistical Mechanics and its Applications **458**, 378–390 (2016)
36. Instituto Verificador de Circulaciones. http://www.ivc.org.ar
37. Alexa. https://www.alexa.com/topsites/countries/AR
38. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273 (2003). ACM
39. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
41. Boydstun, A.E., Bevan, S., Thomas, H.F.: The importance of attention diversity and how to measure it. Policy Studies Journal **42**(2), 173–196 (2014)
42. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: Information Theory, 2004. ISIT 2004. Proceedings. International Symposium On, p. 31 (2004). IEEE
43. Tukey, J.W.: Exploratory Data Analysis vol. 2. Reading, Mass., ??? (1977)
44. Natrella, M.: Nist/sematech e-handbook of statistical methods (2010)
45. Dallmann, A., Lemmerich, F., Zoller, D., Hotho, A.: Media bias in german online newspapers. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 133–137 (2015). ACM
46. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
47. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications **42**(13), 5645–5657 (2015)

48. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta (2010)

Figures

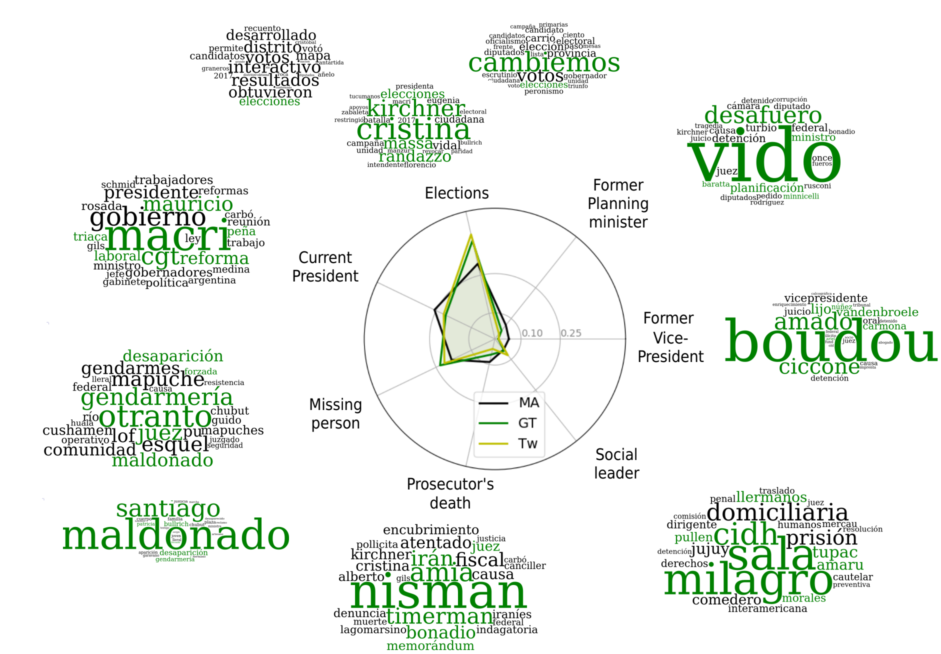

Figure 1 Radar plots of the Media and Public Agendas represented by the ten topic (then collapsed to seven) distribution and their corresponding wordclouds. The Public Agenda is represented both by Google Trends (GT) and Twitter (Tw). Topic names are introduced together with the wordclouds containing the most important keywords involved in the definition of each topic. In green color we show the keywords used to define the topics in the Google Trends and Twitter queries (see tables 1 and 2) and therefore in our construction of the Public Agenda.

Tables

Table 1 Queries used in Google Trends in order to build the Public Agenda.

Topic name	Google Trends query
Elections	elecciones + cambiemos + cristina kirchner + massa + randazzo
Missing person	santiago maldonado + juez otranto + patricia bullrich + gendarmería + desaparición forzada
Former Planning minister	de vido + desafuero + ministro de planificación + minnicelli + baratta
Current President	mauricio macri + cgt + reforma laboral + peña + triaca
Social leader	milagro sala + cidh + tupac amaru + pullen llermanos + morales
Prosecutor's death	nisman + amia + memorándum con irán + timerman + juez bonadio
Former Vice-President	amado boudou + ciccone + ariel lijo + vandenbroele + núñez carmona

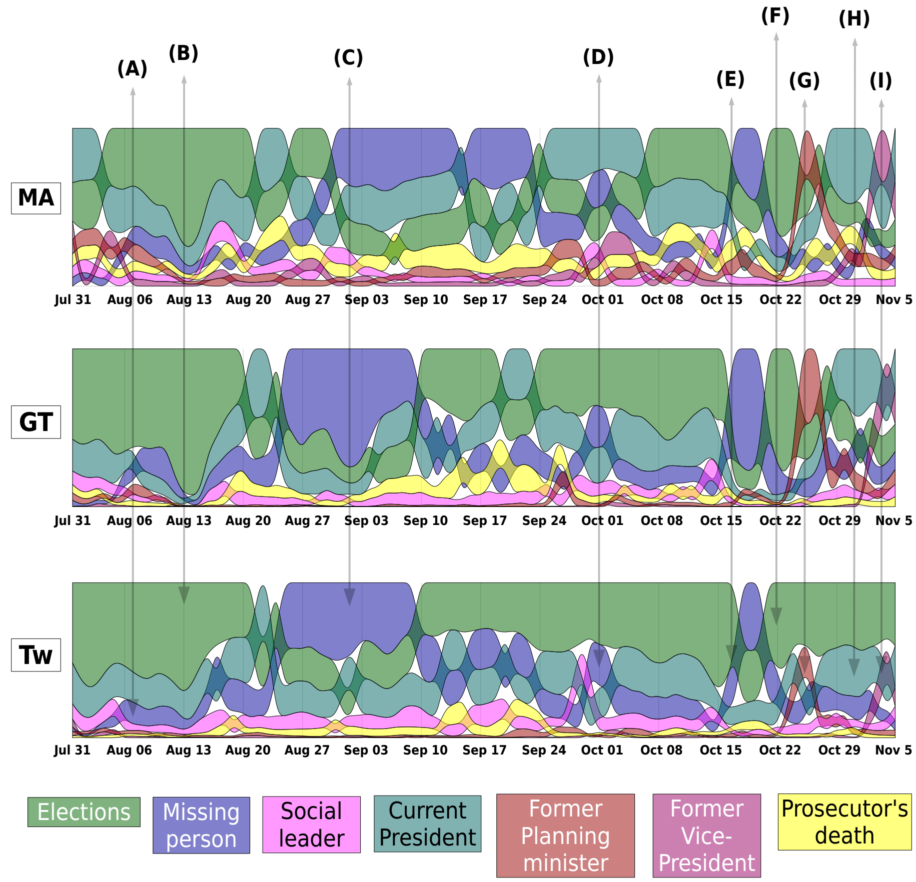

Figure 2 Bump graph of the time-dependent Media (MA) and Public Agendas extracted from Google Trends (GT) and Twitter (Tw). Widths and rankings of the curves encode topic's relative weight. Also, some important events related to the topics are pointed out: **A**: First news about Santiago Maldonado's disappearance (Missing person); **B**: Primary elections; **C**: March one month after Santiago Maldonado's disappearance; **D**: March two months after Santiago Maldonado's disappearance; **E**: Appearance of Santiago Maldonado's body; **F**: General elections; **G**: Julio De Vido's detention; **H**: Debates on labor reform; **I**: Amado Boudou's detention (Former vice-president). A more detailed explanation is given in section **Supplementary Material**.

Table 2 Queries used in Twitter: Due to different characteristics in the search tool of Twitter, we adapted the queries employed in Google Trends, but preserving, at least we can, the most important keywords.

Topic name	Twitter query
Elections	elecciones + cambiemos + kirchner + massa + randazzo
Missing person	maldonado + otranto + gendarmería + desaparición
Former Planning minister	vido + desafuero + minnicelli + baratta
Current President	macri + cgt + laboral + triaca
Social leader	sala + cidh + tupac + amaru + pullen + llermanos + morales
Prosecutor's death	nisman + amia + memorandum + timerman + bonadio
Former Vice-President	boudou + ciccone + lijo + vandenbroele + carmona

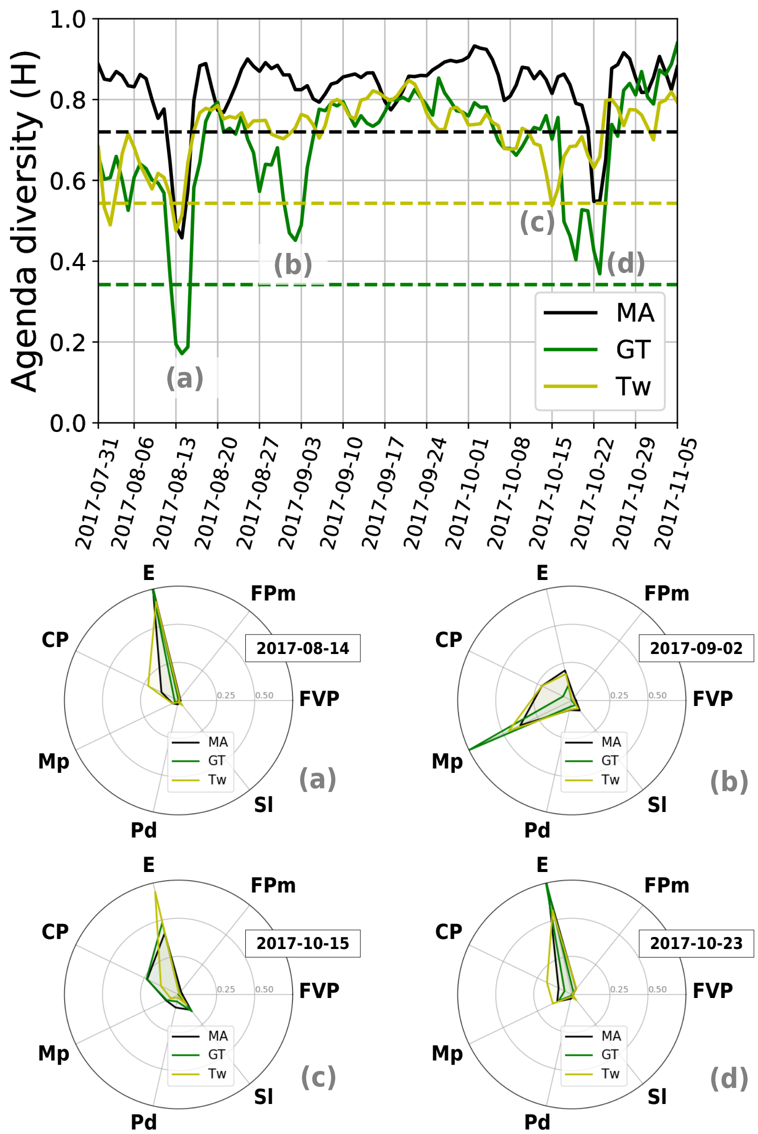

Figure 3 Shannon entropy (H) as a measure of agenda diversity. The Public Agenda shows a less diverse behavior than the Media Agenda as can be seen in the top figure. The horizontal lines correspond to the lower inner fences of each signal in order to identify outliers. The related radar plots show the agenda at the selected days where the time series exhibit dropouts (points a-d), indicating that the most important topic catches most of the public's attention. **E**: Elections; **FPm**: Former Planning minister; **FVP**: Former Vice-President; **SI**: Social leader; **Pd**: Prosecutor's death; **Mp**: Missing person; **CP**: Current President.

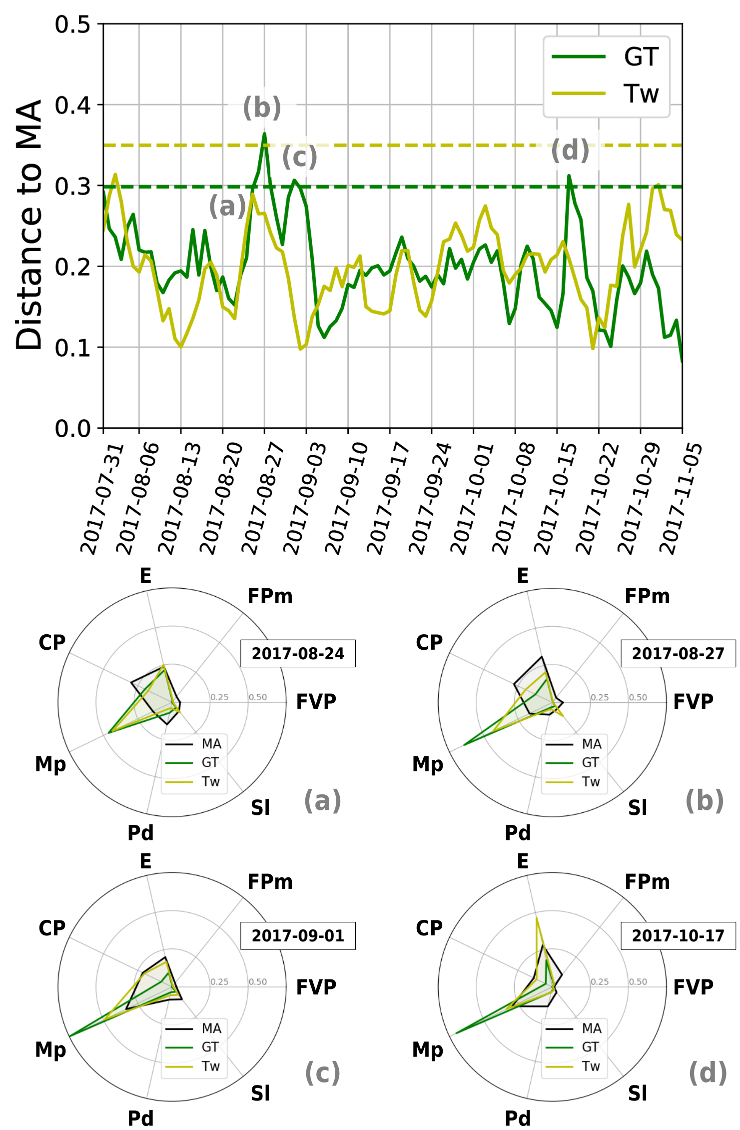

Figure 4 Jensen-Shannon distance between the Media and Public Agendas as a function of time (with upper inner fences pointed out). Larger distances are due to a greater interest of the audience in the topic *Missing person* which decreases the interest in other topics. On the other side, the Media Agenda still keeps certain degree of diversity. **E**: Elections; **FPm**: Former Planning minister; **FVP**: Former Vice-President; **SI**: Social leader; **Pd**: Prosecutor's death; **Mp**: Missing person; **CP**: Current President.

Table 3 Correlation between the topic temporal profiles of the Public Agenda and their counterpart in Media Agenda. All correlation values are statistically significant ($p < 10^{-9}$), except (*) which is significant with $p < 0.05$.

Topic name	Correlation MA and GT	MA and Tw	GT and Tw
Elections	**0.81**	**0.59**	**0.75**
Missing person	**0.68**	**0.76**	**0.89**
Former Planning minister	**0.92**	**0.82**	**0.87**
Current President	**0.77**	**0.75**	**0.63**
Social leader	**0.49**	**0.25(*)**	**0.57**
Prosecutor's death	**0.56**	**0.59**	**0.75**
Former Vice-President	**0.90**	**0.92**	**0.97**

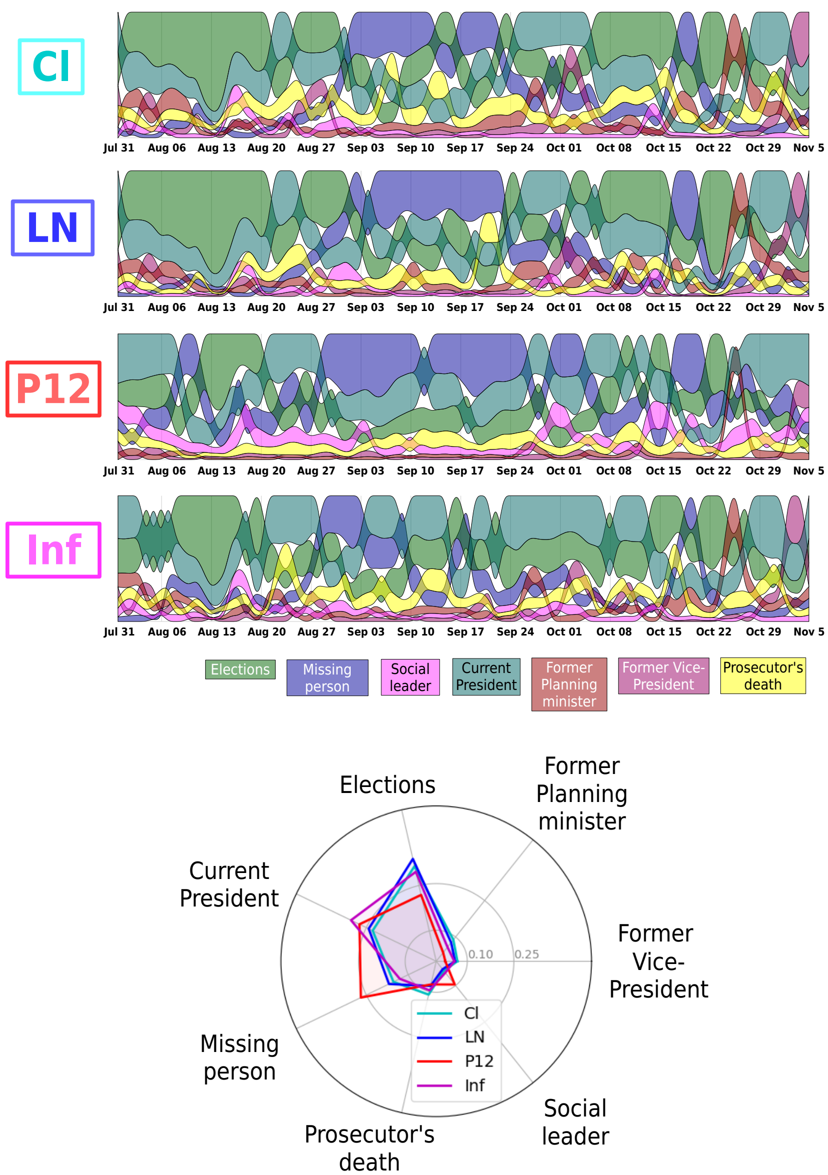

Figure 5 Bump charts of newspaper agendas and radar plot of the average distributions. The figure shows, in a qualitative way, the bias in the different newspaper agendas. For instance, the greater interest of *Página 12 (P12)* in the *Missing person* topic and its slightly lower coverage in the *Former Planning minister* respect to the other newspapers.

Table 4 Correlation between the temporal profiles of the topics found in NMF and associated topics in LDA.

Topic name	Correlation between NMF and LDA
Elections	**0.98**
Missing person	**0.99**
Former Planning minister + Former Vice-President	**0.89**
Current President	**0.94**
Social leader	**0.94**
Prosecutor's death	**0.83**

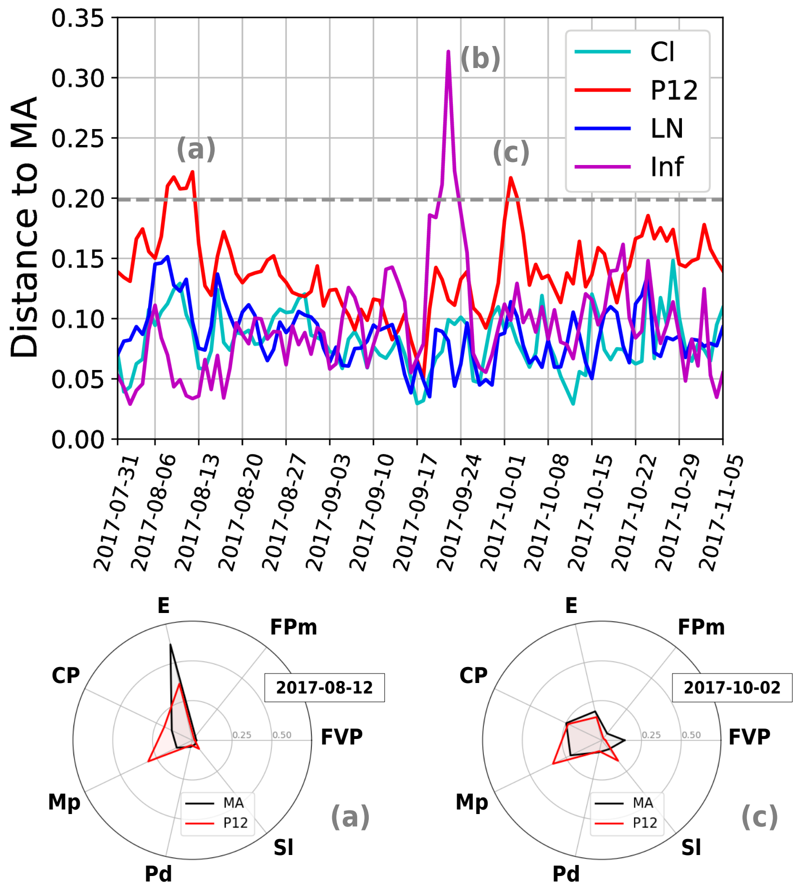

Figure 6 Jensen-Shannon distance between the newspaper agendas and the Media Agenda as a function of time. *Página 12* shows the most different behavior, motivated again by its interest in the *Missing person* and *Social leader* topics, as can be seen in the radar plots which belong to points (a) and (c). The anomalous behavior of *Infobae* at point (b) is due to few articles around that date in our database, therefore we ignore its radar plot. **E**: Elections; **FPm**: Former Planning minister; **FVP**: Former Vice-President; **Sl**: Social leader; **Pd**: Prosecutor's death; **Mp**: Missing person; **CP**: Current President.

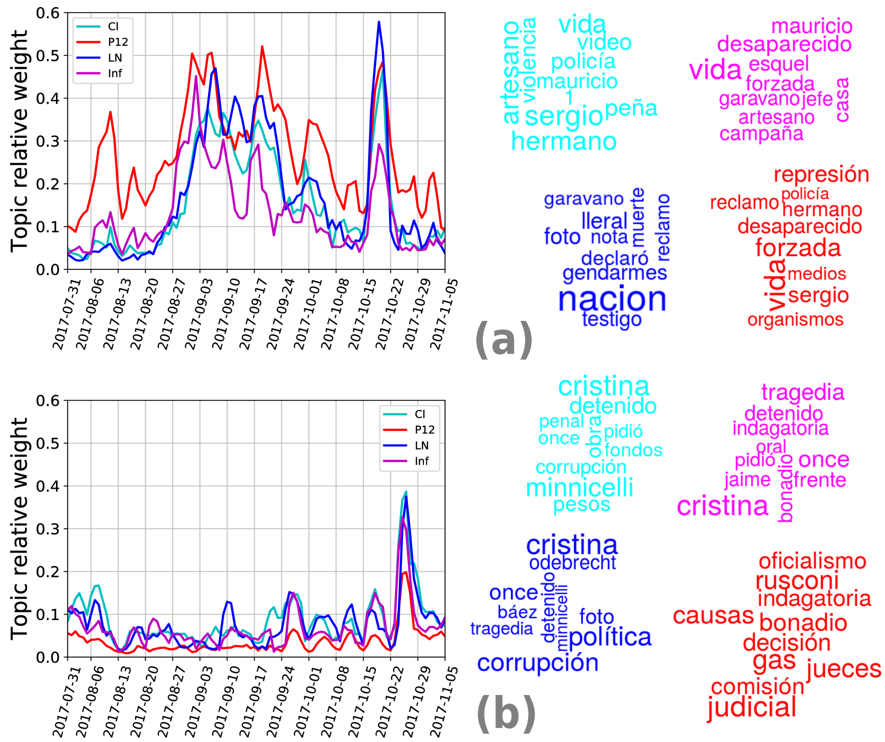

Figure 7 Relative weight of the topics (a) Missing person, and (b) Former Planning Minister, and their corresponding wordclouds of frequent newspaper keywords. We interpreted the differences shown in given periods as an indicator of coverage bias. For instance, in figure (a) *Página 12* pays a greater attention in the first days. In the word-clouds, we show which of the defining words are more frequently used by the corresponding newspaper. Most of them are less informative, but others seem to represent a first approximation in the study of framing.

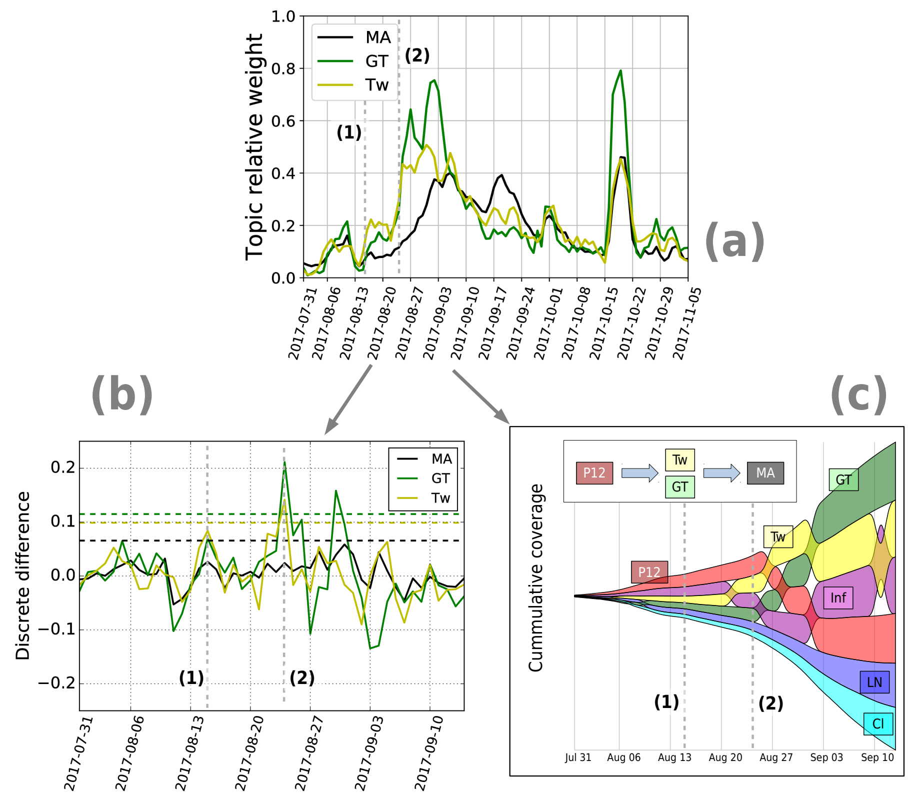

Figure 8 Can we guess the agenda-setting direction in the Missing person topic? The temporal profiles of figure (a) show that the Public and Media agenda seem to differentiate around August 15th (vertical grey line (1)) and the Public increases abruptly its interest in the topic around August 24th (grey line (2)). It can be seen also in figure (b), where the discrete differences were computed. With the computing of the cumulative coverage of figure 7 and figure (a), represented as a bump chart in figure (c), we suggest that the topic was first set by *Página 12* and then the audience's interest cause the coverage of the other media.

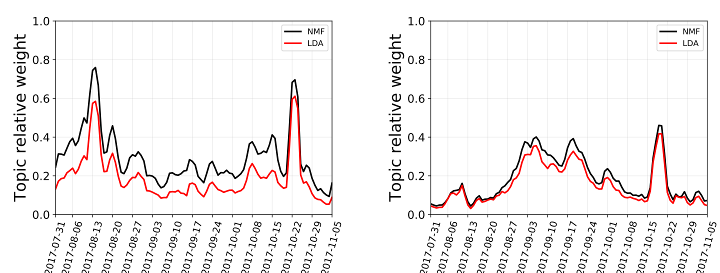

Figure 9 Temporal profiles of topics *Elections* (left) and *Missing Person* (right) for both LDA and NMF. All the topics found by applying NMF have a highly correlated counterpart in LDA.