Agenda diversity and coverage bias: A quantitative approach to the agenda-seeting theory

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

[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, Argentina.
[2]Instituto de Física de Buenos Aires (IFIBA), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
[3]Instituto de Cálculo, CONICET, Intendente Güiraldes 2160, Pabellón 2, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
[*]Corresponding author: spinto@df.uba.ar

Abstract

The agenda-setting theory is a practical framework in order to understand the role of mass media on a society. The theory treats mass media as a very important actor that is able to make people think about, and in many cases how to think about certain topics. When the media successes in this task, we say that the media *set the agenda.* In this work we study the agenda of Argentinian newspapers in comparison with public's interests through a quantitative approach by performing topic detection over the news, identifying the main topics covered and their evolution over time.We look for characterizing the differences and similarities over time between what what we call the Media Agenda and the Public Agenda. On the other hand, we aim to detect coverage bias among the newspapers involved in the analysis in the emerging topics.

1 Introduction

1.1 Agenda Setting Theory

In the famous study performed in Chapel Hill during the US presidential elections in 1968 [1], Maxwell McCombs and Donald Shaw found that those aspects of public affairs that are prominent in the news become prominent among the public. This study is considered the founding of the agenda-setting theory, which focus in the influence of mass media in public opinion. From [2], *"The media agenda is the pattern of news coverage over a period of days, weeks (...) for a set of issues or other topic. In other words, the media agenda is a systematic compilation of the issues or topics presented to the public that identifies the degree of emphasis on these topics."*

Since the Chapel Hill research, several directions of agenda-setting were established [3]. In its basic stage, the theory is focus on the comparison between the topics coverage by the media and the public agenda, i.e. the topics that the public consider as priority. It looks for answering the question if the media is able to set the public agenda, which would transform the media as an important actor in the formation of public opinion. The interaction between media and public is rather complex, for instance, [4] shows that not necessarily the journalists and public preferences coincide.

With the emergency of the Internet, the end of agenda-setting were predicted due to the audience fragmentation onto multiple sources, which would virtually lead to a highly individualized agenda. However, it is based on two assumptions that not necessarily are true: that the public spreads its attention in an homogeneous way across the multiple sources, and that the agendas of that sources are different [3].

The basic agenda-setting sometimes is called the "first level agenda-setting". The very often quoted phrase of Bernard C. Cohen *"The press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about."* illustrates its object of study. On the other hand, the "second level agenda-setting", sometimes called *attribute agenda-setting*, studies the *objects* (in a social psychology way, where an *attitude object* designate a thing that an individual has an attitude or opinion about) present in the media agenda. When the media talks about an object some attributes are emphasized, and others not.

The "second level agenda-setting" is linked with *framing* [5] [6]. To frame is to *select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation and/or treatment recommendation* (Robert Entman) [3]. This stage of agenda-setting theory can

be summary in the phrase *"the media not only can be successful in telling us what to think about, they also can be successful in telling us how to think about it."*

Other interesting stage of agenda-setting concerns with the sources of media agenda, i.e. if the media set the public agenda, *who sets the agenda media?* Within this framework, *intermedia agenda-setting* observes the competition between different media and how they influence each other. The competition between mass media for the same audience can lead to a homogenization of the agendas [7], which is in the opposite direction of one of the assumptions that predict the end of agenda-setting, as was mentioned before.

1.2 Related works

Typical works in agenda-setting theory in any of its stages are based on the comparison between the media agenda and the public agenda. HABLAR SOBRE METODOS TRADICIONALES ....

In this work we propose to employ unsupervised topic modeling to detect the most important topics and follow their dynamics over time. Unsupervised topic modeling is a widely used technique in the analysis of a large corpus of documents. It is an alternative to the dictionary-based analysis, which is the most popular automated analysis approach in social science research [8], and allows to work with a corpus without a prior knowledge, letting the topics emerge from the data. Despite the popularity of this methods, we believe that there is still a lack in the employment of these ones through the lens of the agenda setting framework.

As mentioned above, many works based on news corpus emphasize the performance of the topic model over a labeled corpus, focusing on the proper detection of the topics [9][10][11]. The temporal profile of topics is usually embedded in the context of topic tracking [12][13], or in the recognition of emerging topics in real-time [14] mostly applied to social media. Typically a dynamical description is not carried out in a set of topics but rather focused on a single issue. For instance, in [15] it is shown that the newspapers and Twitter have an opposite reaction to the changes of the unemployment rates, in [5] the competition of frames in this case about gun control is explored, and in [16] it is shown how twitter activity varies in different regions depending on the location of terrorist attacks. A remarkable exception is [17] where they work with a set of predefined issues. In this work the question of *causality* is also faced up. Their study shows that sometimes the traditional media set the agenda and sometimes, the social media does. They show that social media is always more interested in social issues than the traditional one, and

despite the existence of correlation, the social media agenda can not be seen as a *slave* of the traditional media.

In this work we propose a simple method to the study of Mass Media and audience response through topic detection algorithms. Our work intends to contribute another quantitative approach which complements the agenda-setting theory describe above. Rather than focus on a single issue or a set of independent topics, we work with the agendas (the media and the public) as an object in their own, studying their evolution over time. On the other hand, we aim to take an insight about Media dynamics and Public response in order to create useful tools at the time of constructing mathematical models about the interaction between Media and Public, investigation that we started in [18].

2 Description of the work

We analyzed a corpus of news' articles that were published between July 31th and November 5th in the section *Politics* of the electronic 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* [1]. On the other hand, *Infobae* has the website with more visitors, above the websites of *Clarín* and *La Nación* [2].

The corpus analyzed is constituted 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* of the respective news portal, while the articles which belong to *Página 12* were taken from the section *El país*.

The analysis made basically consist of topic detection over the corpus' articles in order to describe it as a set of topics which evolve over time. Topic detection is a powerful computational technique that allows us to analyze a big amount of texts that can be impossible otherwise [19]. For a careful description of the methodology implemented please see section A. This methodology not only gives us the evolution over time of the topics, but also a set of keywords that allow us to interpret and understand what the topics are talking about.

We take advantage of topics' keywords on the one hand by making queries to the *Google Trends* tool and getting the relative size of *Google* searches that people made about the identified topics, and in the other hand by making

[1] www.ivc.org.ar

[2] https://www.alexa.com/topsites/countries/AR

queries to the advance search tool in the social media *Twitter* in order to get the relative amount of tweets related. We take these two tools as a way to measure audience interests in the space of topics defined by the Media.

After all this proceedings, which we are going to give more details during the description and discussion of the work, we obtain two objects of study which we call the **Media Agenda (MA)**, and the **Public Agenda (PA)**, which at the same time has two faces, one giving by **Google Trends (Gt)** and the other by **Twitter (Tw)**. In part of the analysis, the Media Agenda will even be described by the agendas of each of the newspapers (or portal news) taken into consideration. After normalization, all agendas are described as a time dependent distribution over the topic's space, where the time scale is day by day. Therefore the agendas give us the relative importance of a given topic respect the other (i.e. it does not give us absolute values such as the number of titles or tweets associated).

3 Discussion

As was mentioned above, the starting point of our analysis is the topic decomposition of the corpus. We chose to decompose the corpus in 10 topics, which three of them we interpret to be talking about the same macro-topic which we called *Elections*, while other two were intereted as talking about other macro-topic called *Missing person*, as can be seen below. The meaning of the topics or macro-topics are more explained in section B. So the final description of the agendas is made with only 7 topics. We opted for choosing this arbitrary number of topics in order to describe the corpus with as least information as possible, but it is not more than an arbitrary decision, validated in some manner by our knowledge of the corpus. After the proceedings describe in section A, we construct the **Media Agenda (MA)** and the **Public Agenda (PA)**, and all their derivations, as time-dependent distributions.

3.1 Media and Public agenda: a qualitative approach

In figure 1 we show the radar plots of the average distributions of the **MA** and the **PA** discriminated by **GT** and **Tw** together with the introduction of the topics' names. A radar plot is an alternative of histograms that allows the visual comparison of distributions easier. In this figure we also show the wordclouds of the keywords that define each topic, where the size of the word belongs to the importance in the topic's definition given by the topic detection algorithm. 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 table 1 together with the linear correlation between the topics' temporal profiles that form the Public Agenda and their counterparts in the Media Agenda.

On the other hand, in figure 2 we show a bump chart of the Agendas, which is the dynamical representation of the Agendas of figure 1. A bump chart provides us a very useful visualization tool for displaying the relative weight of the topics and at the same time their ranking, putting on the top the most important topic at a particular date. In figure 2 we also point out some important events related to the topics.

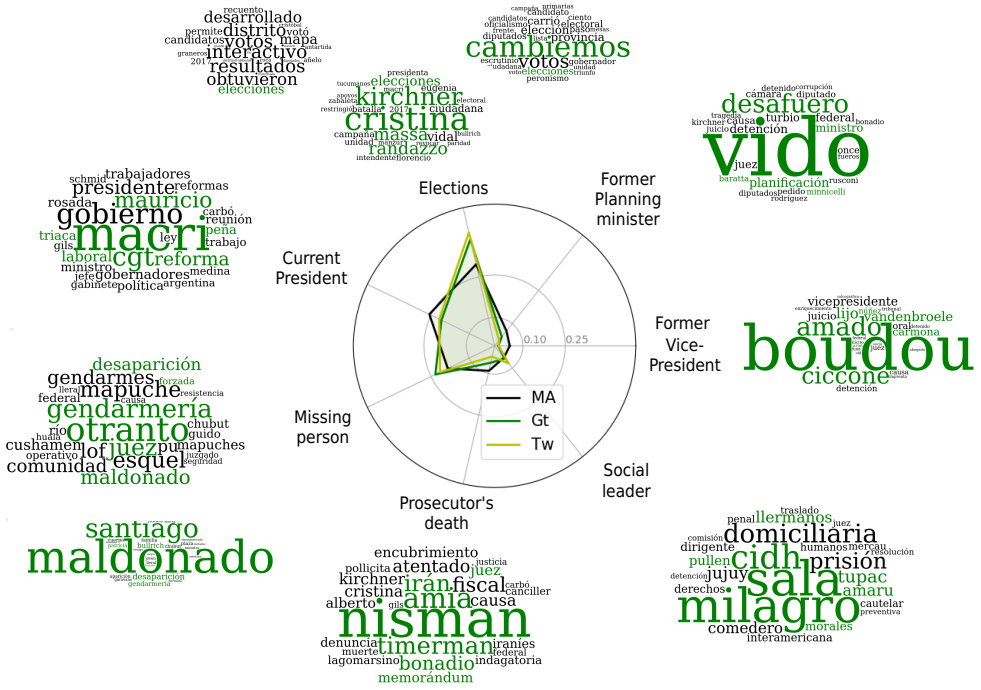

Figure 1: **Radar plots of the Media and Public Agenda average distributions and topics' wordclouds.** The Public Agenda is represented either by Google Trends (Gt) and Twitter (Tw). Topics' names are introduced together with the wordclouds containing the most important keywords involved in the definition of each topic. In green color we show those included in the Google Trends and Twitter queries (see table 1) and therefore in our construction of the Public Agenda.

3

[3]Due to different characteristics in the search tool of Twitter, we adapted the queries

	Google Trends query	Correlation MA and Gt	MA and Twitter	Gt and Twitter
Elections	elecciones + cambiemos + cristina kirchner + massa + randazzo	**0.81**	**0.59**	**0.75**
Missing person	santiago maldonado + juez otranto + patricia bullrich + gendarmería + desaparición forzada	**0.68**	**0.76**	**0.89**
Former Planning minister	de vido + desafuero + ministro de planificación + minnicelli + baratta	**0.92**	**0.82**	**0.87**
Current President	mauricio macri + cgt + reforma laboral + peña + triaca	**0.77**	**0.75**	**0.63**
Social leader	milagro sala + cidh + tupac amaru + pullen llermanos + morales	**0.49**	**0.25(*)**	**0.57**
Prosecutor's death	nisman + amia + memorándum con irán + timerman + juez bonadio	**0.56**	**0.59**	**0.75**
Former Vice-President	amado boudou + ciccone + ariel lijo + vandenbroele + núñez carmona	**0.90**	**0.92**	**0.97**

Table 1: Queries performed in Google Trends in order to made up the Public Agenda. We also shown the correlation between the topics' temporal profiles of the Public Agenda and their counterpart in Media Agenda. All correlation values are statistical significant ($p < 10^{-9}$), except (*) which is significant with $p < 0.05$.

The figures introduced above show in a qualitative way the differences between the agendas, and the dynamics of the topics, i.e. the range of dates in which a given topic was an important one, for which topic it was replaced, and so on. We can see, for instance in the radar plot of figure 1, a more interest of the audience in the topic *Missing person* than the Media, or inversely in the topic *Prosecutor's death*, when we see all the period analyzed as a whole. Also we can observe a great similarity between **Gt** and **Tw** agendas which form the Public Agenda.

On the other hand, the linear correlations of table 1 are in all cases positive and statistically significant, which we interpret as a form of validation of the topics found in the corpus and the keywords that describe it. We expected that the Media's and public's interest should generally follow a similar a pattern due to the external events, although the periods where those differ are of particular interest for us. A non positive (or a non significant) correlation may imply that we are not properly detecting the keywords or features that describe a particular topic, so the Google Trends' or Twitter's pattern would not be able to reflect a similar behavior that its counterpart in the Media.

---

employed here but preserving the at least at we can the most important keywords. The queries employed by topic were:
Elections: elecciones + cambiemos + kirchner + massa + randazzo
Missing person: maldonado + otranto + gendarmeria + desaparicion
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

3.2 A quantitative approach

3.2.1 Agenda diversity

In order to quantify the similarities and differences between the Media Agenda and the Public Agenda, we start by asking how is the distribution of each agenda among the topic's space. In particular, we measure how diverse is each agenda. Following [20], we calculate the normalized Shannon's entropy (H, see eq.7) in order to measure the diversity of the **MA** and **PA**.

In figure 3 we can see the value of H as a function of time. We can see that there are periods where the diversity is lower than the usual, more notorious in the Public Agenda giving by **Gt**. We are going to pay attention to four dates in the Public Agenda, which three detected as outliers of the typical behavior, two from **Gt** and one from **Tw**.

A small value in the diversity is due to the fact that the most important topic attracts practically all the attention of the public or the media. In the radar plots included also in figure 3 we can see that two of the points (**a** and **d**) belong to the topic *Elections* and coincide with the primary and general legislative elections that took place in August 13th and October 22th. In all the agendas these points where detected as outliers except point (d) in Twitter Agenda: By inspecting the radar plot the diversity in this agenda can be the result of the association between the topic *Elections* and the *Current President*. Discussions in Twitter about elections appear also in point (c), when the other agendas seems to be more diverse. On the other hand, we inspect the point (b) despite not being detected as a outlier, which belong to the topic *Missing person* and related to a month after the disappearance of Santiago Maldonado (see section B). We emphasize the discussion about this topic because we see interesting facts that appear along the analysis.

From the measure of H we can also see that the median of the Public Agenda diversity is statistical significant lower than the Media Agenda's one. Specifically $H_{Gt} = 0.73$ and $H_{Tw} = 0.74$ are statistical significant 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 falls in the diversity in response to specific events. We conclude that its an important fact about audience behavior: given a finite set of topics, **the Public Agenda is less diverse than the Media Agenda**, because the public seems to focus more in the most important topic than the Media can do, maybe due to editorial decisions.

3.2.2 Public Agenda's distance

The measurement of the Shannon's entropy made above is an independent property of each distribution. Here we directly compare the Agendas by computing the Jensen-Shannon distance. We again identify outliers and aim to interpret them. In figure 4 we show the Jensen-Shannon distance as a function of time. We inspect three points that seems to be of particular interest. In all cases, the radar plots shows that a greater distance is associated with a more interest of public in the topic *Missing person*.

Points **(c)** and **(d)** shows that both the public and the Media are interest in that topic, but the Media have to cover other topics, so the distance value can be seen as a derivation of the diversity effect discussed in the last section. However points **(a)** (we take this point due to be a extreme of the distance in the middle of the period despite not being an outlier) and **(b)** seem to show an 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 reached the first place in public's interest in both **Gt** and **Tw** before that in the Media. We associate this fact with a campaign made in social media like Facebook and Twitter in August 26th, that paid for the appearance of Santiago Maldonado and had a great repercussion, maybe at first underestimated by the Media (see section B).

It is important to recall that it is our interpretation based on the knowledge of the context, and that we are not studying causality (we will say a few words about it in section 3.4), i.e. we can't say, for instance, that in this case the Public Agenda set the Media Agenda. However, the Jensen-Shannon distance, in conjunction with the measurement of the agenda diversity given by the Shannon entropy, give an insight of independent behavior of the Public and the Media, and its identification can be a starting point to study the Media reaction to a change in audience's interests.

3.3 Media Agenda: differences between newspapers

In this section we leave aside the Public Agenda and we study how the Media agenda varies when one consider the newspapers separately. In figure 5 we show the bump charts which belongs to each newspaper analogously to figure 2. The topics are the same introduced in the wordclouds of figure 1, but at computing the topics' weights the articles are separated by newspaper. We also show the radar plots showing the average distribution, as made in figure 1.

In figure 5 we can see in a qualitative way the slightly differences between the newspapers' agendas. For instance, we can see how *Página 12* gave more

importance to the topics *Missing person* and *Social leader*, while it did not pay too much attention to the *Former Planning minister* as the others did.

3.3.1 Independent behavior

In other to detect an independent behavior of a newspaper respect the others we again calculate the Jensen-Shannon distance between the 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 is a function of time. We detect three points as outliers, although we discard the point **(b)** due to the low information of *Infobae* in that period. The other two points belongs to a difference between *Página 12* and the other newspapers. We interpret that the coverage of the topic *Missing person* is the principal cause of the outliers. *Página 12* paid more attention to it than the Media Agenda at point **(a)** when the first notices of the Santiago Maldonado's disappearance before the primary elections, and in point **(c)** when a march two months after the disappearance took place, and had not the same coverage as the other march a month before (see section B). Also, in point **(c)**, it can be seen a greater coverage of *Página 12* in the topic *Social leader* while the others seemed to be more interested in the topic *Former Vice-President*.

3.3.2 Coverage bias

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*. 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 a more coverage of *Página 12* respect the other newspapers at the first of the period. We can for instance quantify this difference by the median of the signals. If we focus in the period between July 31th and August 27th, the median of the topic relative weight for *Página 12* is roughly 0.14 and this is statistically significant 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 wordclouds of topic's keywords but separate those that are most frequent mentioned by each newspaper, filtering the words that are common to all and basically define the factual details of the topic. Although most words are less informative, some are interesting to inspect, for instance, the word *represión* (repression) when *Pagina 12* talks about the topic *Missing Person* and the word *Cristina* (Fernández de Kirchner) which is employed by all newspapers except *Pagina 12* when talk about the topic *Former Planning minister* (see section B). We don't go farther but we think that a more deeply study of topics' keywords is a first approximation in the study of framing, which will constitute the core of futures works.

3.4 After all: Who sets the Agenda?

The behaviour of the Media Agenda and the Public one, either by looking at Google Trends or Twitter, shows periods when there is a strong similarity, mostly in the presence of an unexpected event, and periods when seem to be unrelated. These are mere observation that talk about the correlation between the agendas. However the agenda-setting theory is in its nature a theory about causations: Is the audience a passive actor who follows what Media say, or there are periods where the Media must paid attention at public's interests? Paraphrasing the question, who sets the agenda, the Media or the audience? In a world where social media exists and the feedback between a Media and its audience is common currency, nowadays the idea that the Media sets the agenda and the audience blindly follows it, as it's seemed to be suggested in the original work of McCombs, is a very weak one.

In spite of the fact that we think that to establish a causation, i.e. the direction towards information flows, is a task that must be made with extreme care, we can not totally skip the question about causation. Therefore we will discuss at least in a qualitative way the *Missing person* topic. This is the more adequate topic to be discussed because it caused a great impact in either the Media and the audience, and on the other hand, its more important related events are fully contained in the period under studying (see section B). Therefore we can study the way the topic acquired the importance showed along this work.

In figure 8 we show the topic relative weight from the Public Agenda and the Media Agenda. After the initial news, the agendas seems to differentiate around August 16th, when the topic started to acquire more importance in the Public Agenda than in the Media, but around August 24th the topic abruptly increase in Public's interest while in the Media the reaction is slower, showing a significant peak in the plot of the discrete differences. 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, i.e, when the audience set the Media Agenda? After all the audience get involved about this topic by the Media, so how was the coverage before those events? By calculating the cumulative sum we can see how the coverage in a cumulative way was during the first events related to the topic. This measure can be seen as the numerical integrate of the temporal profiles of figure 7 between the initial date and the current date. It is interesting to note that the newspaper which accumulate coverage during the initial stage was the minority one: *Página 12*. Our suggestion of how the agenda was set in this particular topic is the following: one of the newspapers gave great coverage to this one, the audience magnified this in social media and also expressed its interest by reiterative Google searches, and then the rest of the Media oughted to pay attention to this behavior. This suggestion is of course a very reductive one but try to catch in a qualitative way how the information flow was. 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 sensible theme for the Argentinian society, and second, there are political reasons in why *Página 12* was particularly interested in cover this topic while the other Media did not follow this interest.

We took the question about causation only in a qualitative way. There are different metrics that would help in this question, such as Granger causality or correlation by sliding windows. We think that this analysis must be performed with extreme care since the acquirement of the data. With this work we aimed to provide a quantitative characterization of the behaviour of the Media and the Public but not to establish direction of causation, which we hope this to be the core of future ones.

4 Conclusions

The study of Mass Media, and in particular the agenda setting theory, can be empowered by the used of data mining or machine learning algorithms. In this work, through the implementation of a topic detection algorithm we could describe the Agenda of the Media as a distribution which evolves over time and which is defined in a topic's space which emerges from the analysis of the corpus. This gave us an insight of how we can construct and follow the Public's interests, the Public Agenda, in order to compare with the Media Agenda, i.e. 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 focus on this one, when the Media has to cover the other too. On the other hand, the measurement of distances between Agendas can be employed to rapidly detect periods when the Public may have an independent behavior respect to the Media. The methodology implemented here also allow 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 Mass Media and Public interaction. 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 quantitave analysis about causation.

A Methodology

A.1 Numerical representation of the corpus

In order to perform the analysis of the articles in the corpus, we describe them as numerical vectors through the *term frequency - inverse document frequency (tf-idf)* representation. *Tf-idf* gives greater values to terms that appear in less documents of the corpus (i.e more specific terms) and/or to those which appear more frequent in a document.

Given the set of terms made up by all the corpus' words after removing the 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 equation 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 document's vector is normalized to unit Euclidean length. Once the documents' vector are constructed, we join them in a document-term matrix (M), which has dimensions of number of documents in the corpus (d) per number of terms selected (t).

$$\begin{aligned} \text{idf}_j &= 1 + \log(\frac{1+d}{1+n_j}) \\ x_{ij} &= \text{tf}_{ij} \cdot \text{idf}_j \end{aligned} \tag{1}$$

A.2 Topic detection

We perform *non-negative matrix factorization (NMF)* on the document-term matrix (M) in order to detect the main topics in the corpus. *NMF* is an algorithm which factorize a matrix M into two matrices W and H (eq.2), with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect, and very suitable for topic detection.

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

As can be seen in eq.2, the resulting matrix H has dimensions of number of documents per k, and matrix W has dimensions of k per number of terms. The number k is interpreted as the number of topics in the documents and it is a parameter that must be set before the factorization. The matrix H is interpreted as the representation of the documents in the topic-space, and the matrix W as the topics represented in the original term-space.

This factorization usually can not be made exactly so it is approximated by minimizing the reconstruction error, i.e. the distance between matrix M and its approximated form $\tilde{M} = H \cdot W$. We performed *NMF* through the python module *scikit-learn* [21].

A.3 Topic interpretation and temporal profiles

The matrix H obtained by *NMF* gives the representation of the documents in the topic-space. In order to improve its interpretation we normalize each document's vector described in that space to unit l_1-norm. Therefore the components of these vectors can be viewed as a degree of membership of a given document in the set of topics. The index of the largest component tells us which is the most representative topic of the document.

On the other hand, each row of matrix W represent the topic over the term-space. Therefore, the terms associated with the largest components of the i-row are the most representative ones and give an insight of what the topic is talking about.

We define the temporal profile of the topic i, $W_i(day)$ by the eq. 3 where $l(j)$ is the number of words of the document j, and h_{ji} is the degree of membership of document j on topic i. This definition allows all documents to contribute to any topic weight, providing by the fact that each document's vector can have non-zero components in more than one topic.

As last steps of the data's preprocessing, we filter the topics' weight in order to reduce the noise but keeping the most details as possible, by redefining $W_i(day)$ as the mean value of a 3 days width window, centered on the day, as described in equation 4.

Finally we normalize again all the temporal profiles in order to describe each newspaper as dsitribution over the topics' space which evolves over time. This last normalization prevent us against the differences in the number of articles that each newspaper publishes.

$$W_i(day) = \sum_{j}^{d} l(j) \cdot h_{ji} \cdot \delta_{d,day} \tag{3}$$

$$\tilde{W}_i(day) = \frac{1}{3}(W_i(day) + W_i(day-1) + W_i(day+1)) \tag{4}$$

A.4 Outliers identification

We identify outliers values in a data set of N observations by following the box plot construction. The quantities (called fences) in 5 are used, where $Q1$

is the lower quartile (range of the distribution where lies the 25th percent of the data), $Q3$ is the upper quartile (where lies the 75th percent of the data) and $IQ = Q3 - Q1$.

$$\begin{aligned} \text{lower inner fence (LIF)} &= Q1 - 1.5\text{IQ} \\ \text{upper inner fence (UIF)} &= Q3 + 1.5\text{IQ} \\ \text{lower outer fence (LOF)} &= Q1 - 3\text{IQ} \\ \text{upper outer fence (UOF)} &= Q3 + 3\text{IQ} \end{aligned} \tag{5}$$

A point above the upper inner fence considered a mild outlier and a point aove an outer fence is considered an extreme outlier. The same holds for the lower fences. [4]

A.5 Jensen shannon divergence

In probability theory and statistics, the Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. It is based on the Kullback–Leibler divergence (D_{KL})6, but have useful properties such as it is symmetric and it is always a finite value. The square root of the Jensen–Shannon divergence 6 is a metric often referred to as Jensen-Shannon distance. [22]

$$\begin{aligned} D_{KL}(P||Q) &= -\sum P(i) log(\frac{Q(i)}{P(i)}) \\ \text{JS Divergence}(P||Q) &= \frac{1}{2}[D_{KL}(P||M) + D_{KL}(Q||M)] \\ \text{JS Distance (JSD)} &= \sqrt{(}\text{JS Divergence}) \end{aligned} \tag{6}$$

A.6 Normalized Shannon's H Information Entropy

The normalized Shannon's entropy 7 as a way to measure how spread is a distribution, taking the maximun value where all outcomes are equally probable in the case of a discrete distribution.

$$H[p] = \frac{-\sum_{i=1}^{N} p(x_i) * ln(p(x_i))}{ln(N)} \tag{7}$$

[4] http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

B Context

B.1 *Elections*

Two legislative elections were celebrated during the period in great part of the Argentina: Primary elections on August 13th and the general elections on October 22th. 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 to the alliance *Cambiemos*, which is the alliance of the current President *Mauricio Macri* and the current governor of Buenos Aires province, *Maria Eugenia Vidal*.

B.2 Current President

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

B.3 *Missing Person*

Santiago Maldonado vanished on August 1st after a minor clash between the Gendarmerie (Border Guards) and a group of Mapuches, which recognize as themselves 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 of a **forced disappearance**.

A very massive campaign in social media took place on August 26th, under the motto "Where is Santiago Maldonado?", followed by two massive protest marches to the *Plaza de Mayo* took place on September 1st and October 1st, 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 on 17th October 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 responsability of the current administration is still being discussed.

B.4 *Former Planning minister* and *Former Vice-President*

Julio de Vido was the Planning minister during the administration of *Nestor Kirchner* and *Cristina Fernandez 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 inmmediatly jailed on October 27th.

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

B.5 *Social leader* and *Prosecutor's death*

Milagro Sala is an indigenous leader's. 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* 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 on 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.

A Comparison with other topic model

In this section we apply other topic model, Latent Dirichlet Allocation [23] (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 studying should be in some manner robust to the election of the topic model. On the other hand, as was discussed in [24], 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-occurrence 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* [25], which allow 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 the figure and the table 2, 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.

	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**

Table 2: Correlation between the temporal profiles of the topics found in NMF and associated topics in LDA.

References

[1] M. E. McCombs and D. L. Shaw, "The agenda-setting function of mass media," *Public opinion quarterly*, vol. 36, no. 2, pp. 176–187, 1972.

[2] M. McCombs and S. Valenzuela, "Agenda-setting theory: The frontier research questions," *Estados Unidos: Oxford handbooks online*, 2014.

[3] M. McCombs, "A look at agenda-setting: Past, present and future," *Journalism studies*, vol. 6, no. 4, pp. 543–557, 2005.

[4] E. Mitchelstein, P. J. Boczkowski, C. Wagner, and S. Leiva, "La brecha de las noticias en argentina: factores contextuales y preferencias de periodistas y público," *Palabra Clave*, vol. 19, no. 4, 2016.

[5] L. Guggenheim, S. M. Jang, S. Y. Bae, and W. R. Neuman, "The dynamics of issue frame competition in traditional and social media," *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 207–224, 2015.

[6] O. Tsur, D. Calacci, and D. Lazer, "A frame of mind: Using statistical models for detection of framing and agenda setting campaigns.," in *ACL (1)*, pp. 1629–1638, 2015.

[7] C. J. Vargo and L. Guo, "Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online us news," *Journalism & Mass Communication Quarterly*, p. 1077699016679976, 2017.

[8] L. Guo, C. J. Vargo, Z. Pan, W. Ding, and P. Ishwar, "Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling," *Journalism & Mass Communication Quarterly*, vol. 93, no. 2, pp. 332–359, 2016.

[9] X.-Y. Dai, Q.-C. Chen, X.-L. Wang, and J. Xu, "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, IEEE, 2010.

[10] L. Po, F. Rollo, and R. T. Lado, "Topic detection in multichannel italian newspapers," in *Semanitic Keyword-based Search on Structured Data Sources*, pp. 62–75, Springer, 2016.

[11] A. Brun, K. Smaïli, and J.-P. Haton, "Experiment analysis in newspaper topic detection," in *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, pp. 55–64, IEEE, 2000.

[12] X. Hu, "News hotspots detection and tracking based on lda topic model," in *Progress in Informatics and Computing (PIC), 2016 International Conference on*, pp. 248–252, IEEE, 2016.

[13] W. Li, J. Joo, H. Qi, and S.-C. Zhu, "Joint image-text news topic detection and tracking by multimodal topic and-or graph," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 367–381, 2017.

[14] M. Cataldi, L. Di Caro, and C. Schifanella, "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, ACM, 2010.

[15] S. Soroka, M. Daku, D. Hiaeshutter-Rice, L. Guggenheim, and J. Pasek, "Negativity and positivity biases in economic news coverage: Traditional versus social media," *Communication Research*, p. 0093650217725870, 2017.

[16] A. E. Ali, T. C. Stratmann, S. Park, J. Schöning, W. Heuten, and S. C. Boll, "Measuring, understanding, and classifying news media sympathy on twitter after crisis events," *arXiv preprint arXiv:1801.05802*, 2018.

[17] W. Russell Neuman, L. Guggenheim, S. Mo Jang, and S. Y. Bae, "The dynamics of public attention: Agenda-setting theory meets big data," *Journal of Communication*, vol. 64, no. 2, pp. 193–214, 2014.

[18] S. Pinto, P. Balenzuela, and C. O. Dorso, "Setting the agenda: Different strategies of a mass media in a model of cultural dissemination," *Physica A: Statistical Mechanics and its Applications*, vol. 458, pp. 378–390, 2016.

[19] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[20] A. E. Boydstun, S. Bevan, and H. F. Thomas, "The importance of attention diversity and how to measure it," *Policy Studies Journal*, vol. 42, no. 2, pp. 173–196, 2014.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, p. 31, IEEE, 2004.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[25] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010.

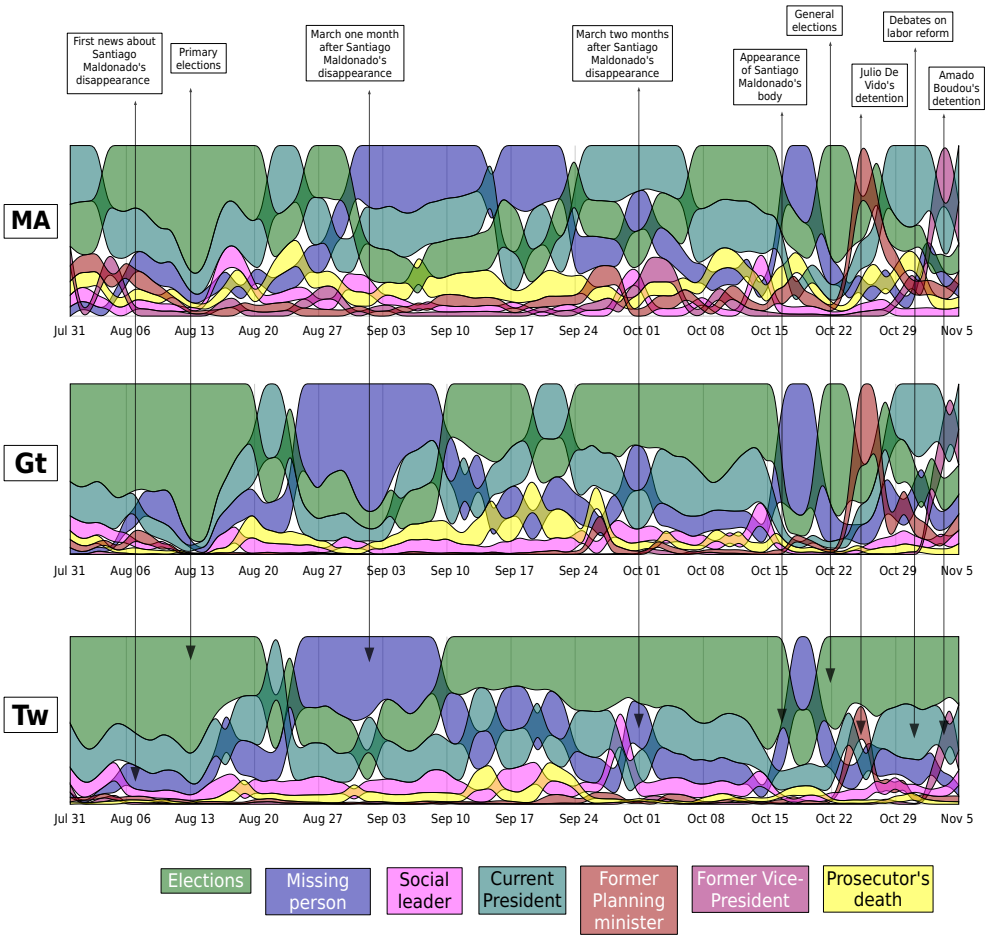

Figure 2: **Bump graph of the Media Agenda (MA) and Public Agenda extracted from Google Trends (Gt) and Twitter (Tw).** Some important events related to the topics are pointed out. The curves' width and their ordering are related to the topics' relative weight.

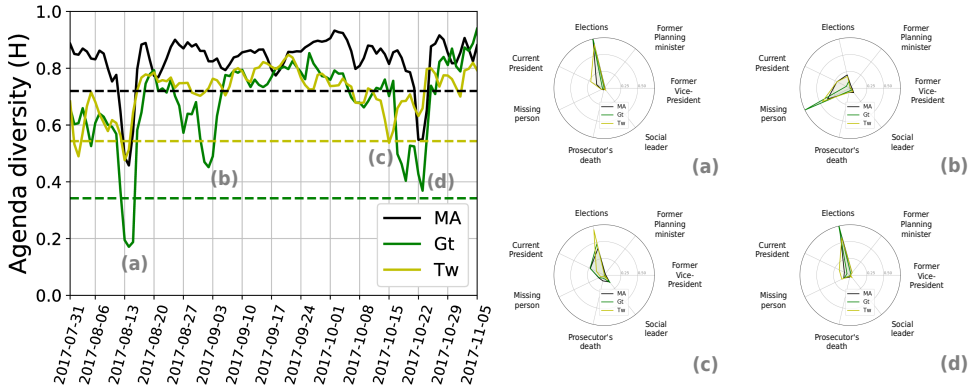

Figure 3: **Shannon entropy (H) as a measure of agenda diversity.** The Public Agenda show a less diverse behavior than the Media Agenda as can be seen in the left figure. The horizontal lines are the lower inner fences of each signal in order to identify outliers. The related radar plots shows that those dates when the agenda has a low diversity, the most important topic catches the most public's attention

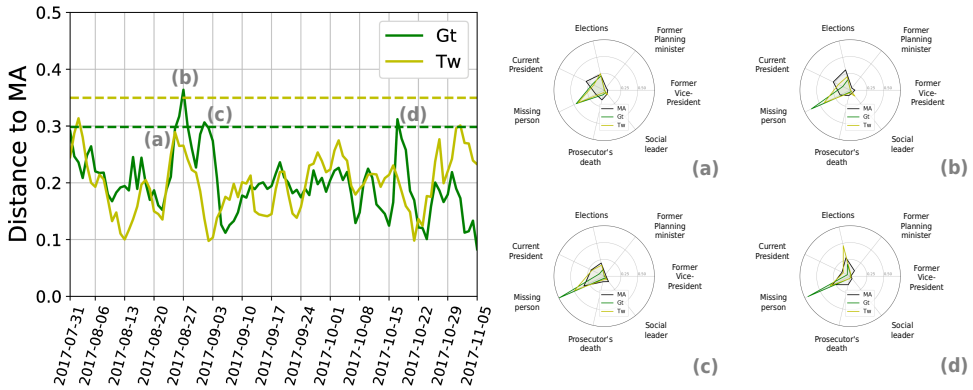

Figure 4: **Jensen Shannon distance between the Media Agenda and the Public Agenda as a function of time** (with upper inner fences pointed out). The larger distances are due to a greater interest of the audience in the topic *Missing person* which leads to lesser interest in the other topics, which the Media has to cover, maybe except in points (a) and (b) where the Media seems not to anticipate the public interest in the mentioned topic.

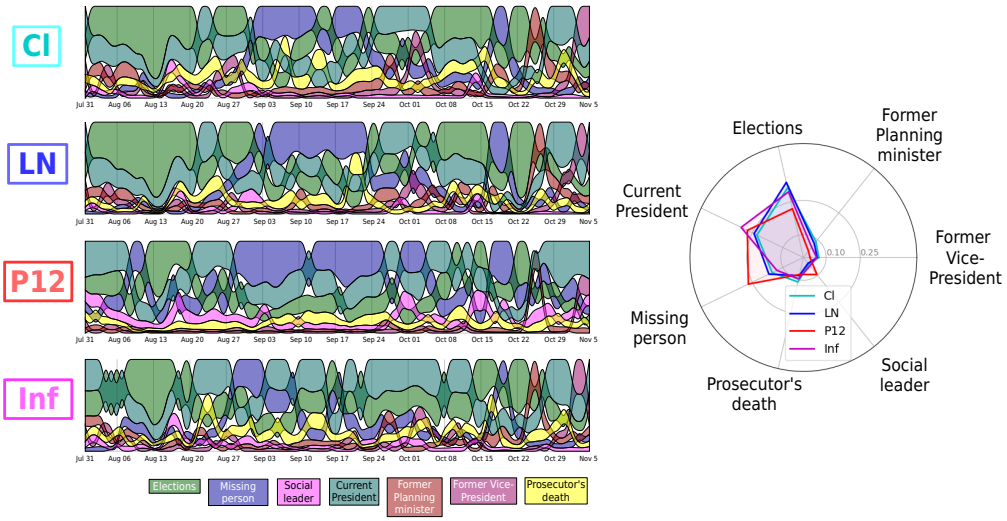

Figure 5: **Bump charts of newspapers' Agenda and radar plot of the average distributions.** The figure shows, in a qualitative way, the differences between the mentioned agendas, for instance, the greater interest of Página 12 (P12) in the *Missing person*'s topic and its slightly lesser interest in the *Former Planning minister* respect to the other newspapers.

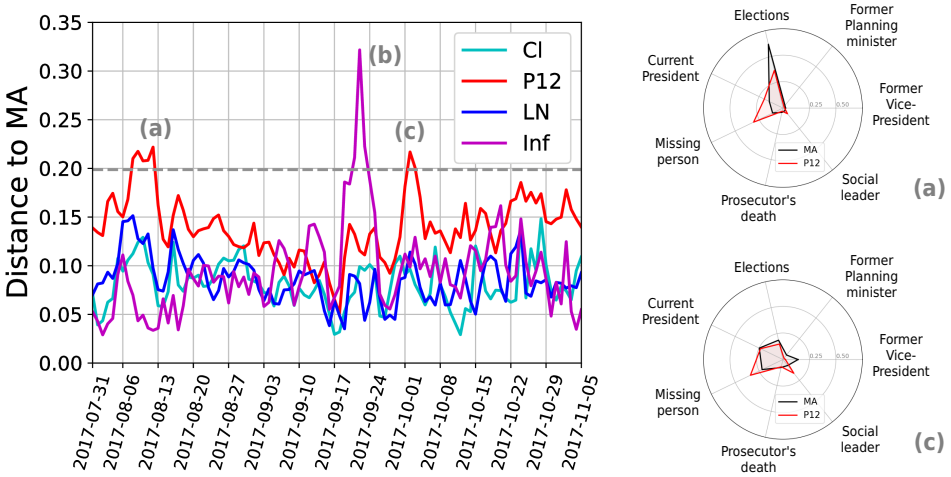

Figure 6: **Jensen Shannon distance between the newspapers' agenda and the Media Agenda as a function of time.** *Página 12* (P12) shows the more different behavior, motivated again by its interest in the *Missing person* and *Social leader* topics as can be seen in the radar plots which belongs to points (a) and (c). The anomalous behavior of *Infobae* (Inf) at point (b) is due to few articles around that date in our database, therefore we ignore its radar plot.

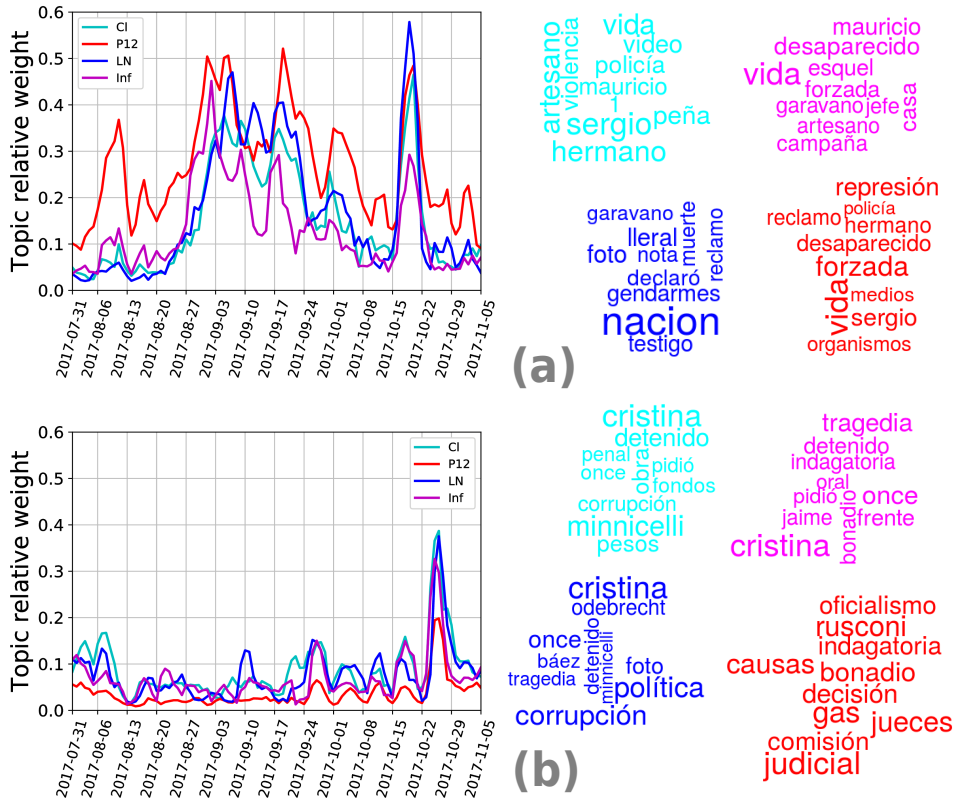

Figure 7: **Relative weight of the topics (a) *Missing person*, and (b) *Former Planning Minister*, and wordclouds of frequent newspapers' keywords**. We interpreted the differences shown in certain periods as an indicator of coverage bias. For instance, in figure (a) Página 12 pays a greater attention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

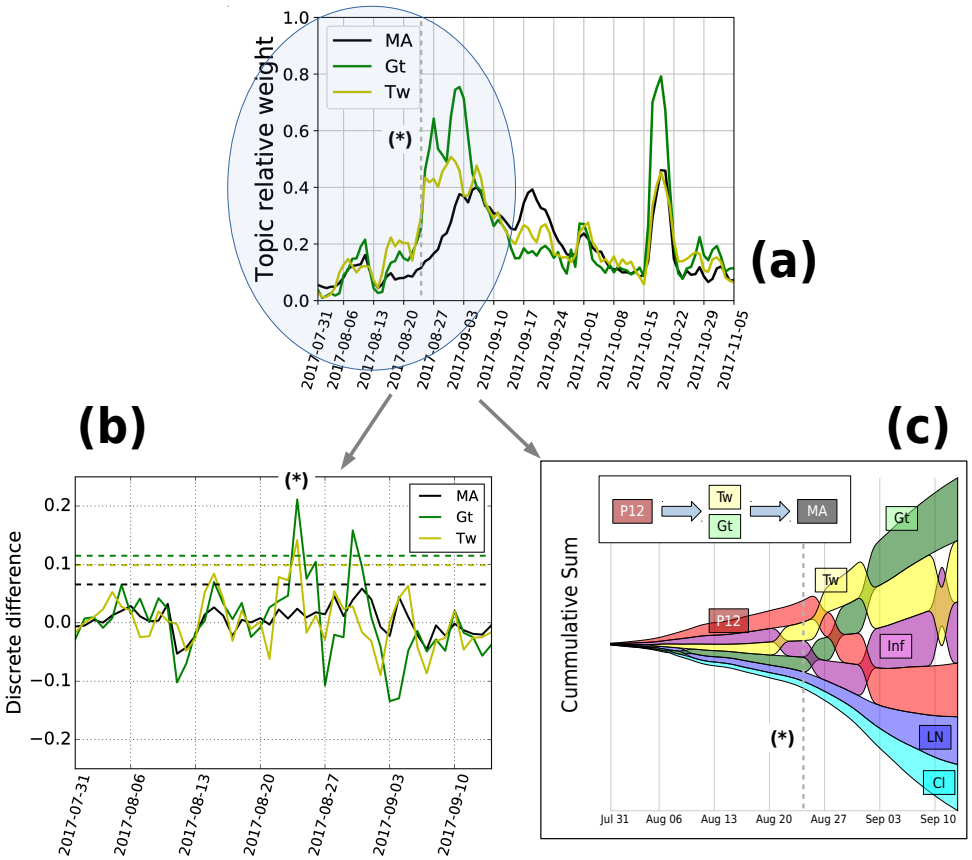

Figure 8: **Agenda setting direction in Missing person's topic?** The temporal profiles of figure (a) show that the Public increased abruptly its interest in the topic around August 24th, which belongs to the peak **(*)** shown in figure (b), where the discrete differences were computed. This date is also pointed out in the rest of the figures with a vertical grey line together with the asterisk. With the computing of the cumulative sum 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 Public's interest cause the coverage of the other Media.

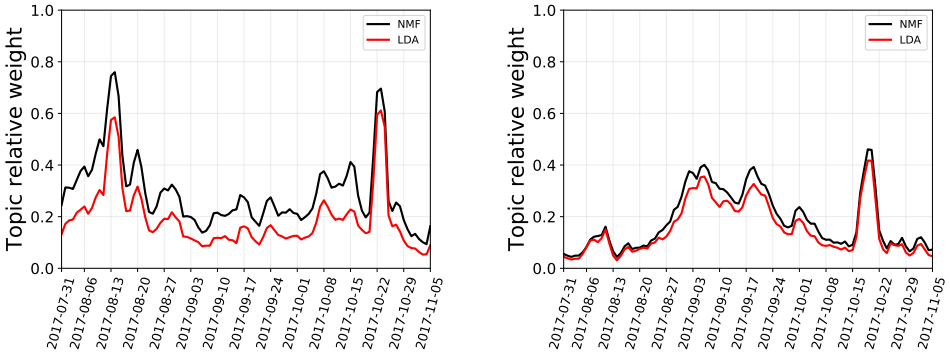

Figure 9: Temporal profiles of topics *Elections* (left) and *Missing Person* (right) for both LDA and NMF. All the topics found by applying NMF have a highly correlated counterpart in LDA.