RESEARCH

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

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 the different topics of the agenda? 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's dynamics and public attention. 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 measure Agenda's diversity as a function of time using Shannon's entropy and difference between Agendas using Jensen-Shannon's distance. We found that the Public Agenda is less diverse than the Media Agenda, and we are also capable to detect periods of time where coverage of certain issues are biased (coverage bias).

Keywords: agenda-setting; agenda diversity; coverage bias; topic modelling

Introduction

Agenda Setting Theory

To understand the ecosystem of information flow and opinion formation is one of the goals in these days. In this framework, many ingredients can be part of different mechanism of social influence. A major role in this ecosystem is played by the Mass Media outlets, which used to be the source of information of many of us. Informed people tends to interact with each other, either via personal interactions or through social networks. In this scenario, becomes essential to understand the role of Mass Media Influence in a given social group.

This question has been raised years ago. 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".

In its basic stage, the theory 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. This stage is sometimes 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. 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 are not necessarily true: the public spreads its attention in an homogeneous way across the multiple sources, and that the agendas of that sources are different [3].

From the Chapel Hill research, several directions of agenda-setting were established [3]. 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. This theory is linked with *framing* [4, 5]. 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 [6], which is in the opposite direction of one of the assumptions that predict the end of agenda-setting, as was mentioned before.

Since the irruption of internet in our lives and the access to big data, there were many attempts to analyze from a quantitative point of view the theory of Agenda Setting based on the original work of McCombs [1]. Many of them follows what we called a "static approach", i.e. by summarizing certain issues present in news media and looking for effects in the audience [7, 8, 9]. For instance, in Argentina, [10] examines the difference between public and journalists preferences. Several works by E. Zunino, like [11] and [12], studied the coverage given by the main newspapers of particular events which have happened during the former administration of Argentinian president *Cristina Fernández de Kirchner* (from 2007 to 2015), where the government started to confront with news organizations that were critical to its management as they were an opposition party [13].

Other line of research have been focused on detecting bias in the media, either by taking into account the number of mentions related to a preferred political party

[14, 15] or by identifying the ideology through the position of the media regards to certain issues or actors [16, 17]. New trends in agenda-setting theory propose to represent ideological aspects or issues emphasized as networks like mind maps, and the basic idea is the comparison between media's knowledge representative network with its counterpart in the audience [18, 19].

Most of the works cited above follow typical methods of social science research. However, a useful tool in the analysis of large corpus of documents, which is not widely used in the agenda setting theory, is unsupervised topic modeling. These methods are an alternative to the dictionary-based analysis, which is the most popular automated analysis approach [20], and allows to work with a corpus without a prior knowledge, letting the topics emerge from the data. Despite the popularity of these methods in other fields, they have been under-employed in the agenda setting framework. 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 [21, 22, 23]. The temporal profile of topics is usually embedded in the context of topic tracking [24, 25], or in the recognition of emerging topics in real-time [26] mostly applied to social media.

A typical dynamical analysis of Agenda Setting is focused on a single issue. For instance, in [27] it is shown that the newspapers and Twitter have an opposite reaction to the changes of the unemployment rates. In [4] the competition of frames about gun control is explored, and in [28] the authors show how fluctuations of twitter activity in different regions depend on the location of terrorist attacks. A remarkable exception is found in [29], 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. However, a drawback of working with predefined issues is that they usually reflects general themes, such as unemployment or crime, and the issues selection must be made by the researcher.

In this work we propose a novel method in order to study the dynamics of Mass Media Agenda as a set of topic evolving in time and their correspondence with the Public and Social Network Agendas. Our work intends to contribute another quantitative approach which complements the agenda-setting theory describe above. It mainly stands within the framework of first-level agenda setting, but issues about the comparison between media agendas and framing are also discussed. 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 [30].

Materials and Methods

The Media Agenda

In this work we analyze a three month period of the Argentinian Media's Agenda composed 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* [31]. On the other hand, *Infobae* has the website with more visitors, above the websites of *Clarín* and *La Nación* [32]. 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* (Politics) of the respective news portal, while the articles which belong to *Página 12* were taken from the section *El país* (The country).

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 [33]. 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 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 put them together in a document-term matrix (M), which has dimensions of number of documents in the corpus (d) per number of terms selected (t).

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

We perform *non-negative matrix factorization (NMF)* [33, 34] on the document-term matrix (M) in order to detect the main topics in the corpus. A topic is a group of similar articles which roughly talks about the same subject. *NMF* is an algorithm which factorize the 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 [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 per 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.

The matrix H is the representation of the documents in the topic-space. Each row is normalized to unit l_1-norm so that their components can be viewed as a degree of membership of a given document in the set of topics. In particular, the index of the

[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)

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 group of words which represent more accurately each topic (called keywords) and therefore an insight of what the topic is talking about. 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 $\tilde{M} = H \cdot W$. This was performed *NMF* through the python module *scikit-learn* [35].

A time-dependent representation is obtained by defining the temporal profile of the topic i, $W_i(day)$, by 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 and δ is the Kronecker delta. Providing by the fact that each document's vector can have all non-zero components, it is allowed that a document contributes to more than one topics' weight. In order to reduce noise, we apply a linear filter with a sliding window of size 3, 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. This last normalization prevent us against the differences in the number of articles that each newspaper publishes.

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

The Google and Twitter Agendas

In order to have a proxy about how the Media Agenda is reaching the public, we look for the same topics in *Google* and *Twitter* in the same period of time. We take advantage of topics' keywords on the one hand by making queries to the *Google Trends* tool and getting the relative weight of Google searches that people made about the identified topics, and in the other hand, by making queries to the advance search tool in the social media Twitter in order to get the relative amount of related tweets.

Outliers identification

We identify outliers values in a data set of N observations by following the box plot construction. The quantities (called fences) in 4 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 $IQ = Q3 - Q1$.

$$lowerinnerfence(LIF) = Q1 - 1.5IQ \tag{4}$$
$$upperinnerfence(UIF) = Q3 + 1.5IQ \tag{5}$$
$$lowerouterfence(LOF) = Q1 - 3IQ \tag{6}$$
$$upperouterfence(UOF) = Q3 + 3IQ \tag{7}$$

A point above the upper inner fence considered a mild outlier and a point above an outer fence is considered an extreme outlier. The same holds for the lower fences [36].

Jensen-Shannon distance

In probability theory and statistics, the Jensen-Shannon divergence (JS_{Div}) is a method to measure the similarity between two probability distributions. It is based on the Kullback-Leibler divergence (D_{KL}), but have also useful properties such as it is symmetric and it is always a finite value. The square root of the Jensen-Shannon divergence (see eq.8, where $M = \frac{P+Q}{2}$) is a metric [37] often referred as Jensen-Shannon distance (JSD).

$$D_{KL}(P||Q) = -\sum P(i) log(\frac{Q(i)}{P(i)}) \tag{8}$$

$$JSD(P,Q) = \sqrt{JS_{Div}(P,Q)} = \sqrt{\frac{1}{2}[D_{KL}(P||M) + D_{KL}(Q||M)]} \tag{9}$$

Normalized Shannon's H Information Entropy

The normalized Shannon's entropy (eq.10) is a way to measure how spread is a distribution, taking the maximum value where all outcomes are equally probable in the case of a discrete distribution.

$$H[p] = \frac{-\sum_{i=1}^{N} p(x_i) * ln(p(x_i))}{ln(N)} \tag{10}$$

Results

The starting point of our analysis is the decomposition of the corpus in ten topics. This is an arbitrary decision, validated in some manner by our knowledge of the corpus, which could be revised in future works. From the initial ten topics, three of them could be interpreted as part of the same macro-topic called *Elections*. The same happen with two others which we classified as part of the macro-topic called *Missing person*, leaving the final description of the agendas as made with seven topics. The meaning of the topics or macro-topics is contextualized in section of Supplementary Material.

After the procedure described in the previous section, we construct the **Media Agenda (MA)** and the **Public Agenda (PA)**, and all their derivations, as time-dependent distributions.

The Public and Media agendas

In figure 1 we show a ten 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). 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 reflects its importance in the topic's 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 table 1. In table 2 we show the linear correlation between the topics' temporal profiles from the Public Agenda and their counterparts in the Media Agenda.

In figure 2 we show the time evolution of the topics with a bump chart of the Agendas. The bump chart provides a very useful visualization tool for displaying

the relative weight of the topics at the same time that their ranking. In figure 2 we also highlight some important events related to the dynamics of the topics. [2]

The figures introduced above show not only the topics and their differences between the Agendas, but also the dynamics of the topics. For instance, we can see in the radar plot of figure 1, a greater interest of the audience in the topic *Missing person* than the Media, or inversely in the topic *Prosecutor's death*, at least in the analyzed time-lapse. We can also observe a great similarity between **GT** and **Tw** agendas which both are different faces of the Public Agenda.

On the other hand, figure 2 allow us to appreciate how the main topic change in time and have a glance of the qualitative differences between agendas. The linear correlations shown in table 2 are positive and statistically significant in all the cases, This can be interpreted as a validation of the topics found in the corpus and the keywords that describe it. Even though it is expected that the Media's and public's interest should generally follow a similar a pattern due to the external events, we are interested in those periods where they significantly differ. A non positive (or a non significant) correlation may imply, besides the obvious conclusion of agenda's disengagement, that we could eventually fail in properly detect 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.

A quantitative description of the Agendas

Agenda diversity

In order to quantify the diversity of the Agendas, we start by representing them as distributions in the topic's space. Following [38], we calculate the normalized Shannon's entropy (H, see eq.10) 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, where it can be focused in those periods of time where 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 where can be detected four local minima of Shannon's entropies. Three of them are outliers as defined in section , two of them from **GT** and one from **Tw**.

A lower value in the agenda's diversity is due to the fact that the most important topic attracts practically all the attention of the public and the media. In the radar plots included in figure 3 we can see that two of these outliers (**a** and **d**) belong to the topic *Elections* and are related with the primary and general legislative elections

[2] Due to different characteristics in the search tool of Twitter, we adapted the queries employed here but preserving 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 + 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

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. Why is that? The radar plot of the Twitter agenda for this day displays an association between the topic *Elections* and the *Current President* which could be decreasing the important of this topic. Discussions in Twitter about elections appear also in point (c), when the other agendas seems to be more diverse. On the other hand, we focused in point (b), despite not being detected as a outlier, because it belong to the topic *Missing person* and this date correspond to the rally that took place one month after the disappearance of *Santiago Maldonado* (see section). It is important to notice that the Shannon's Entropy of GT Agenda displays a minimum (collapsing agenda) not shown neither in the Media nor in the Twitter Agendas. We emphasize the discussion about this topic because we can appreciate 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 dropouts 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.

Distance between Media and Public Agenda's

Given our interpretation 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 Public and Media Agenda, i.e., specific events where the public interest do not match with media offer. In figure 4 we show the Jensen-Shannon distance between Media and Publics Agendas as a function of time. We focus in three points that seems to be relevant enough. In all cases, the topic distributions at that particular dates displayed by 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 highlight this topic, but the Media do not disregard other topics, so the corresponding distance between them can be interpreted of another way of seeing the diversity effect 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)** 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 becomes the most important in public's interest (both **GT** and **Tw**) days before that it happen in the Media Agenda. It can be associates 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. It was very important and was initially underestimated by the main Media outlets in Argentina (see section).

It is important to notice that it is our interpretation of the data based on the knowledge of the context, and we are not measuring causality between Agendas (we will say a few words about it in section), 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.

Agenda bias in different Media outlets

In this section we leave aside the Public Agenda as a whole and we study how the Media agenda of each media outlet. In figure 5 we show the bump charts corresponding to each newspaper analogously to figure 2. The topics are the same introduced in the wordclouds of figure 1, but when computing the topics' weights, the articles are discriminated 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 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 outliers behavior of a given newspaper respect to 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 as 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 corresponds to differences between *Página 12* and the other newspapers and correspond to differences in the coverage of the topic *Missing person*.

The point **(a)** corresponds to the first news of the Santiago Maldonado's disappearance reported by *Página 12* before the primary elections, and in point **(c)** to the two months rally after the disappearance took place (see section). Another difference in point **(c)** corresponds to a greater coverage of *Página 12* in the topic *Social leader* where the others 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* [39]. 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* respect the other newspapers at the beginning of the period. We can, for instance, quantify this difference calculating the median of the signals. If we focus in the period between July 31th and August 27th, the median of the topic's 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 word-clouds of topic's keywords but separating those who are more frequently mentioned by each newspaper, and filtering the words that are common to all and basically define the factual details of the topic. 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 section). We think that a more deeply study of topics' keywords could be a first approximation in the study of framing, which will constitute the core of futures works.

After all: Who sets the Agenda?

The behavior of the Media Agenda and the Public one, either by looking at Google Trends or Twitter, shows periods when there is a strong similarity among them (mostly in the presence of an unexpected event) and periods when they 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 causality: Is the audience a passive actor who follows what the Media says, 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 too naive.

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 have something to say in this work about causality. We think that the *Missing person* topic is the most adequate topic to be discussed because:

- It caused a great impact in either the Media and the audience
- its coverage fully deploys along the time lapse analyzed in this manuscript (see section).

In figure 8 we show the topic's relative weight from the Public and Media Agendas. After the initial coverage, the agendas seem to differentiate around August 15th, when topic started to became more important in the Public Agenda than in the Media one. Around August 24th, the topic abruptly increase in public's interest 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, 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 the cumulative coverage 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 responsible to accumulate 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; it is amplified by public in the social networks and also expressed by reiterative Google searches. Then the rest of the Media pay attention to this subject and it becomes an important topic in the Media Agenda. This interpretation try 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 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 (see section).

We took the question about causality 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 behavior of the Media and the Public but not to establish direction of causation, which we hope this to be the core of future ones.

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 quantitative analysis about causality.

Supplementary Material

Context

We provide here a more detailed explanation of the topics discussed among this work. The ideology of the media in Argentina expresses the highly polarized politic climate observe in Argentinian society. During the administration of *Cristina Fernández de Kirchner* (2007-2015), the government maintained a conflict with several news organizations. It led media such as *Clarín*, *La Nación* and the portal news *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 a opposite ideological inclination, supporting the former administration of *Cristina Kirchner* and therefore being very critical with the current *Maurico 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 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 *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 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.

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* that 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 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.

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.

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* 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.

Comparison between NMF and LDA

In this section we apply other topic model, Latent Dirichlet Allocation [40] (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 [41], 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 11, 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.

$$TC = \sum_{i<j}^{N} \frac{2d_{ij}}{d_i + d_j} \tag{11}$$

We perform a decomposition into 10 topics using LDA with the python module *gensim* [42], 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 3, 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., Valenzuela, S.: Agenda-setting theory: The frontier research questions. Estados Unidos: Oxford handbooks online (2014)
3. McCombs, M.: A look at agenda-setting: Past, present and future. Journalism studies **6**(4), 543–557 (2005)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. Zunino, E., Aruguete, N.: La cobertura mediática del conflicto campo-gobierno. un estudio de caso. Global Media Journal **7**(14) (2010)
12. 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)

13. 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)
14. Lazaridou, K., Krestel, R.: Identifying political bias in news articles. Bulletin of the IEEE TCDL **12** (2016)
15. 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)
16. Elejalde, E., Ferres, L., Herder, E.: On the nature of real and perceived bias in the mainstream media. PloS one **13**(3), 0193765 (2018)
17. 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)
18. Guo, L.: The application of social network analysis in agenda setting research: A methodological exploration. Journal of Broadcasting & Electronic Media **56**(4), 616–631 (2012)
19. Vu, H.T., Guo, L., McCombs, M.E.: Exploring "the world outside and the pictures in our heads" a network agenda-setting study. Journalism & Mass Communication Quarterly **91**(4), 669–686 (2014)
20. 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)
21. 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
22. 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
23. 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
24. 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
25. 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)
26. 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
27. 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)
28. 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)
29. 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)
30. 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)
31. Instituto Verificador de Circulaciones. http://www.ivc.org.ar
32. Alexa. https://www.alexa.com/topsites/countries/AR
33. 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
34. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
35. 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)
36. Natrella, M.: Nist/sematech e-handbook of statistical methods (2010)
37. 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
38. 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)
39. 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
40. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
41. 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)
42. 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
Tables
Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.

Figure 1 Radar plots of the Media and Public Agendas represented by the ten topic distribution and their corresponding 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 the keywords used to define the topics in the Google Trends and Twitter queries (see table 1) and therefore in our construction of the Public Agenda.

Figure 2 Bump graph of the Media Agenda (MA) and Public Agenda extracted from Google Trends (GT) and Twitter (Tw). The curves width and their ordering are related to the topics relative weight. Some important events related to the topics are pointed out: **A**: First news about Santiago Maldonado's disappearance; **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. A more detailed explanation is given in section .

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. **E**: Elections; **FPm**: Former Planning minister; **FVP**: Former Vice-President; **Sl**: Social leader; **Pd**: Prosecutor's death; **Mp**: Missing person; **CP**: Current President.

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. **E**: Elections; **FPm**: Former Planning minister; **FVP**: Former Vice-President; **Sl**: Social leader; **Pd**: Prosecutor's death; **Mp**: Missing person; **CP**: Current President.

Figure 5 Bump charts of newspapers' Agenda and radar plot of the average distributions. The figure shows, in a qualitative way, the bias in the different newspaper's 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.

Figure 6 Jensen-Shannon distance between the newspapers agenda and the Media Agenda as a function of time. *Página 12* 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* at pint (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 word-clouds of frequent newspapers' 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 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 and Media agenda seem to differentiate around August 15th (vertical grey line (1)) and the Public increase 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 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.

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

Topic's 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

Table 2 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 $p < 0.05$.

Topic's name	Correlation MA and GT	MA and Twitter	GT and Twitter
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**

Table 3 Correlation between the temporal profiles of the topics found in NMF and associated topics in LDA.

Topic's 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**