Analyzing mass media influence using natural language processing and time series analysis

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
16 December 2019

REVISED
19 March 2020

ACCEPTED FOR PUBLICATION
7 April 2020

PUBLISHED
3 July 2020

PAPER

Analyzing mass media influence using natural language processing and time series analysis

Federico Albanese[1,6], **Sebastián Pinto**[2,3], **Viktoriya Semeshenko**[4,5] and **Pablo Balenzuela**[2,3]

[1] Instituto en Ciencias de la Computación, CONICET- Universidad de 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] 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
[4] Universidad de Buenos Aires, Facultad de Ciencias Económicas, Buenos Aires, Argentina
[5] CONICET-Universidad de Buenos Aires, Instituto Interdisciplinario de Economía Política de Buenos Aires, Av. Córdoba 2122, C1120 AAQ, Buenos Aires, Argentina
[6] Author to whom any correspondence should be addressed.

E-mail: falbanese@dc.uba.ar

Keywords: mass media influence, sentiment analysis, topic detection, time series analysis

Supplementary material for this article is available online

Abstract

A key question of collective social behavior is related to the influence of mass media on public opinion. Different approaches have been developed to address quantitatively this issue, ranging from field experiments to mathematical models. In this work we propose a combination of tools involving natural language processing and time series analysis. We compare selected features of mass media news articles with measurable manifestation of public opinion. We apply our analysis to news articles belonging to the 2016 US presidential campaign. We compare variations in polls (as a proxy of public opinion) with changes in the connotation of the news (sentiment) or in the agenda (topics) of a selected group of media outlets. Our results suggest that the sentiment content by itself is not enough to understand the differences in polls, but the combination of topics coverage and sentiment content provides an useful insight of the context in which public opinion varies. The methodology employed in this work is far general and can be easily extended to other topics of interest.

1. Introduction

Mass media play one of the important roles in the process of public opinion formation. Beyond informing about facts and events, mass media give an interpretation about such events, providing individuals a way to understand their relevance. Through its capacity to reflect reality from its own perspective, media determine the relative importance given to different topics, a process known as agenda-setting. The agenda-setting theory is usually summarized in the quote 'maybe media does not tell you what to think, but what to think about' [1, 2]. In other words, media can tell you what is and what is not important, and to what extent. Therefore, the agenda-setting power of mass media produce an important effect, that acquires relevance, for instance, in the political opinion formation during electoral contexts [3].

Previous research has shown how public perception of a political event is modified by mass media [4]. The influence is usually manifested on the basis of topics emphasized and omitted by the media [5]. In addition, other studies demonstrated that reading different media in a sustained manner leads individuals to modify their political ideology, aligning their votes with the editorial viewpoint of a given newspaper [6]. Gerber and Dean [7] studied the 2005 governor elections in the state of Virginia (US), and observed that the newspaper's reading affected the decisions of a finite number of voters, and produced induced changes in the perception of politicians.

In the last decades, given the availability of data and computational resources, the quantitative analysis of mass media influence has been addressed from different perspectives.

On one hand, the emergence of social media and availability of online data enriched the research about mass media impact. For instance, King *et al* [8] detected an increase in the number of tweets about a specific topic after being exposed to related news. Yasseri and Bright [9] showed that consulting for the number of mentions of a candidate in Wikipedia produces changes in vote shares for particular parties, regardless of whether those were positive or negative. The spreading and consumption of fake news in social media was analyzed in [10], where the authors show that people tend to share fake news that reinforce their ideological bias. The role of the news' sentiment (positive or negative connotation) has been explored within the framework of how the connotation is related to the variation of economic indicators [11], how it affects public expectation about economy [12, 13], or how it shapes the public opinion about a given prominent issue [14]. In the same line [15, 16], addressed how the sentiment of bots' or influential users' tweets induces the connotation of their followers expressions in Twitter. Other techniques such as topic modeling were also employed in order to describe topics dynamics in media [17–19], and how it is related to audience response.

On the other hand, computational models have also been implemented in order to evaluate different mechanisms of interactions between mass media and individuals, and how mass media influence the formation of collective public opinions [20–22]. In this kind of models, the behavior of mass media is generally not grounded on data, but the incorporation of relevant information contained in news articles, as is analyzed in this work, gives a powerful tool to explore different scenarios of interactions between mass media and public opinion [23].

In this paper, we study the relationship between mass media and public opinion using a combination of sentiment analysis and topic detection of news articles. As a proof of concept, we apply this methodology to a particular case study which is the 2016 US presidential campaign, analyzing news articles where the involved candidates are mentioned. We consider the number of mentions of each candidate, their sentiment content and the evolution of the relative coverage of a set of topics of the same articles (political media agenda). Beyond the particular case studied, the methodology introduced in this work is far general and can be extended easily to other case of interest.

This work is organized as follows: in section 3 we describe the natural language processing tools applied to the news articles. In section 2 we describe the data of the studied case. In section 4 we analyze the time series of the polls, the number of mentions of the candidates, the sentiment analysis of the news content, the topic evolution of the political media agenda, and the sentiment analysis discriminated by topic. We measure the Spearman correlation and the Granger causality to draw useful conclusions. Finally, we discuss these results in section 5.

2. Data: the 2016 US presidential elections

In 2016, a presidential election took place in the United States, facing up the Democrat candidate Hillary Clinton and the Republican candidate Donald Trump, who finally won the election. We centered our analysis in two types of data: polls, as a proxy of public opinion and news articles from four of the main US mass media. The whole analyzed period comprises from 28th of July (last party convention, where the candidates were formally defined) to the 8th of November of 2016 (election date).

2.1. Polls data

We analyzed a total of 263 national surveys conducted by different agencies (an average of 2.7 surveys per day), in which the forecast of the votes of each candidate was measured in a gap of a few days (around 3–5 days). This data, which shows the result of different pollsters, was downloaded from the real clear politics website [24]. All national surveys presented in this work belong to a demographic balanced sample. Polls data fulfill the terms and conditions specified in the site real clear politics and are available as supporting material (https://stacks.iop.org/JPCOMPLEX/1/025005/mmedia).

Figure 1 displays the time series of potential percentage of votes for each candidate (top panel), and the difference between these time series (the percentage of Clinton minus the percentage of Trump, bottom panel). Each point in the time series represent the average of the preceding week. In other words, a 7-days sliding window average was used, which means that each point in the time series takes into account an average of 19 polls.

Figure 1 shows that Clinton kept up an advantage over Trump during all the time period. However, this advantage was affected by drops in the middle and close to the end of the period. By looking at the individual time series, we can see that the initial decreasing of the advantage is due to the gradual ascending of the Trump's intention to vote until October. After that, his percentage went down sharply until it was recovered near the

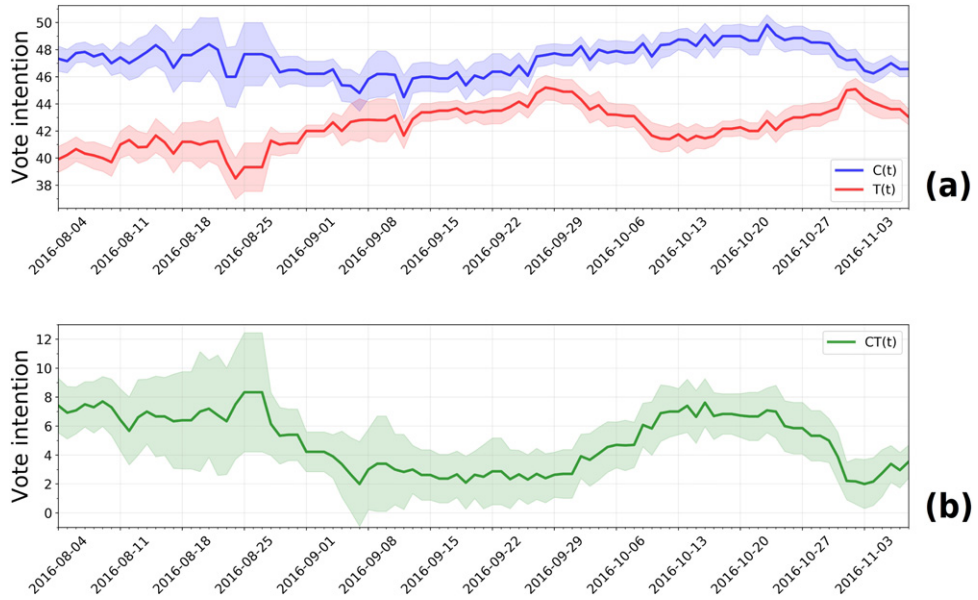

Figure 1. Survey curves of 2016 US presidential elections. Top: the curves of Hillary Clinton (Democrat candidate) and Donald Trump (Republican candidate) for the 2016 US presidential election. Bottom: the spread between these curves.

election date, which explains the last decrease in the Clinton's advantage, which slightly increased in the last days.

2.2. Mass media data

We selected the online editions of The New York Times, Fox News, CNN and USA Today to perform our analysis. The selection criteria is that they are the most popular mass media outlets in term of online searches in all US territory (see appendix A). For instance, the first one is a classical newspaper based in New York city with worldwide influence and readership and the second and third one are cable television news channels which broadcast to many countries all around the world. USA Today is an internationally distributed American daily, middle-market newspaper.

We selected the articles which contain at least the name of one of the two main candidates: Hillary Clinton (Democrat) and Donald Trump (Republican). The analyzed corpus is made up of a total of 15175 articles: 5672 from The New York Times, 5750 from Fox News, 2920 from CNN, and 833 from USA Today. We include all articles that mention one or both candidates irrespective of the section they belonged to. The full corpus is available as supporting material.

3. Methods

3.1. Text mining methods

We focused our analysis in text mining techniques, from which we extract useful information by applying both, sentiment analysis and topic detection, to news articles whose corpus will be detailed bellow.

3.1.1. Sentiment analysis

In order to measure the frequency of positive and negative mentions for a given candidate, we implemented a sentiment analysis algorithm. The sentiment analysis was performed through deep recursive models for the semantic composition applied to sentiment trees [27], in particular by the Stanford CoreNLP implementation of natural language processing [28]. This algorithm consists of assembling a tree from the grammatical structure and a syntactic analysis of each phrase. Then, each word (node) is assigned a sentiment value, taken from a database: very positive, positive, neutral, negative or very negative. In addition, this algorithm takes into account if the words are intensifiers, appeasers, deniers, etc. The algorithm assigns a sentiment value to each node starting from the inner nodes. After several iterations, it ends up assigning the corresponding sentiment value to the total phrase.

There exist several algorithms to perform sentiment analysis, such as those based on the extraction of characteristics of the sentences [29] or lexicon-based approaches to opinion mining [30, 31]. Given that our corpus of news is formed by grammatically correct sentences, the Stanford CoreNLP is the proper algorithm to perform sentiment analysis.

3.1.2. Topic detection

In addition to the sentiment analysis, we also performed a topic detection on the corpus of the news articles using unsupervised learning techniques as was implemented in [19].

We represent news articles as numerical vectors through the *term frequency–inverse document frequency (tf–idf)* representation. The value of each component is given by the frequency of each term in the text (tf) weighted by a measure of specificity (idf) [32]. Vectors are then compiled in a document-term matrix M, of dimension d (number of documents) per t (number of terms, $t = 56\,979$ in our work), which is given by the total amount of words in the corpus after stop-word removal [33]. On the other hand, we detect the main topics in the corpus by performing *non-negative matrix factorization (NMF)* [32, 34] on the document-term matrix (M). A topic is defined as a group of similar articles which roughly talk about the same subject. Let us mention, that analogous results were obtained by applying latent Dirichlet allocation (LDA) method [35] on the same corpus.

NMF decomposes matrix M as the product of two non-negative matrices, H and W (see equation equation (1)), where the first one is a document-topic matrix and gives us the representation of the documents in the space of topics, while the second one is a topic-term matrix and brings the topics described in the space of terms, from where we can extract the keywords which define each one. The inner dimension n in equation (1) is the number of expected topics, which is a parameter that must be set before the decomposition.

$$M^{(d\times t)} \sim H^{(d\times n)} \cdot W^{(n\times t)} \tag{1}$$

In order to calculate the coverage of mass media, we estimate the amount of articles and their relative importance in each topic. For this sake we define the weight of the topic i (T_i) as the product of the amount of news articles (weighted by their degree of membership) and the length of the article. The coverage can be defined on a daily basis (time-dependent distribution) or for the whole period (average distribution). equation (2) shows the coverage for a single day d,

$$T_i(d) = \sum_j l(j) \cdot h_{ji} \cdot \delta_{d_j,d}, \tag{2}$$

where $l(j)$ is the number of words in the document j; h_{ji} (element of matrix H) is the degree of membership of document j on topic i; d_j is the date of document j; and δ is the Kronecker delta. Given that each document vector can have all non-zero components, it is allowed that a document contributes to more than one topic weights. In order to reduce noise, we finally apply a linear filter with a seven day wide sliding window, and normalize the temporal profiles.

3.2. Correlation and causality measures

3.2.1. Spearman correlation

Spearman correlation coefficient is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. While Pearson correlation measures linear relation between the two variables, Spearman correlation assesses the monotonic relationships between them. Before computing the Spearman correlation, we removed the respective linear trend, if there was any, from all time series in order to avoid spurious correlations.

3.2.2. Granger causality

The Granger causality test [26] determines if one time series is able to forecast another one. Given a stationary time series, x_t, modeled by an auto-regressive-moving-average model equation (3), the Granger causality test basically determines if the model of equation (4) is better than the model of equation (3). In other words, this means that the additional information provided by a second time series, $y_{t-\tau}$, improves the forecasting of x_t.

The amount of terms that must be included in equations (3) and (4) are determined by studying both the autocorrelation and the partial correlation of x_t, where w_t's are white noise terms, and θ's, ϕ's, and β are just coefficients [39]. If β is significantly different from zero, we may say that y_t has a causal relation with x_t. Notice that when β is zero, the model of equation (3) is recovered.

$$x_t = \sum_i \phi_i x_{t-i} + \sum_j \theta_j w_{t-j} + w_t \tag{3}$$

$$x_t = \sum_i \phi_i x_{t-i} + \sum_j \theta_j w_{t-j} + w_t + \beta y_{t-\tau} \tag{4}$$

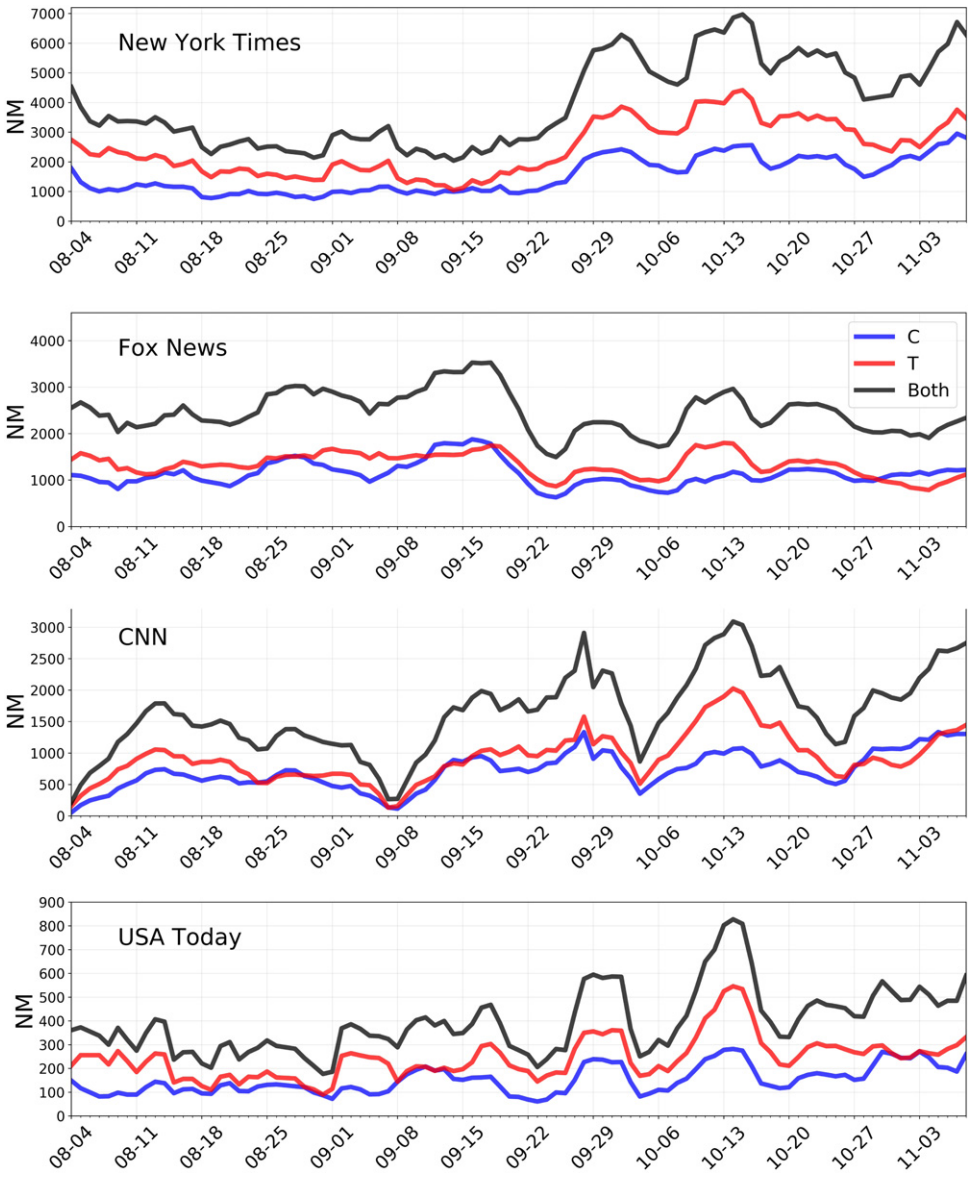

Figure 2. Time series of the total number of mentions. The curves shows the total number of mentions of Hillary Clinton (in blue), Donald Trump (in red) and the sum of both of them (in black) for each media outlet (from top to bottom, New York Times, Fox News, CNN, and USA Today).

4. Results

Here we analyze the polls data and the news articles (by extracting the sentiment content and the topic decomposition) in the electoral period from the 28th of July to the 8th of November of 2016. This period comprehends since the last 2016 party convention, where the candidates were formally defined, until the election date.

4.1. Total number of mentions

As a first approach, we compared the time series of the surveys with the total number of mentions of each candidate in both media [9]. These curves are shown in figure 2. This first analysis was performed regardless the context and sentimental connotation in which the phrases appeared.

We calculate the Spearman's correlation coefficient [25] between the total number of mentions of each candidate (figure 2) with the spread between polls data (figure 1(b)). We take into account that changes in media coverage may not be instantaneously reflected in the polls, either because the time scale of how media can exert influence is not clear, or the publication date of the polls is posterior to the data collection. Therefore, we calculate a lagged correlation between the time series for a range of lags.

We found that the number of mentions of both candidates in New York Times, CNN, and USA Today correlates positively with the spread between Clinton and Trump in the polls, with an average correlation coefficient of 0.663 (NYT), 0.426 (CNN), and 0.246 (USA) respectively. This means that when the number of

mentions increases in these outlets, the difference Clinton minus Trump increases, no matter which candidate is mentioned. On the other hand, these correlations are negative for similar time series in Fox News, with an average correlation coefficient of -0.476. In this case, when the number of mentions of any of the candidates increases in Fox News, the difference Clinton minus Trump decreases. These results are statistically significant ($p < 0.001$) for a lag between 7 and 15 days for The New York Times, Fox News and CNN, while for USA Today these are significant for a lag between 12 and 15 days ($p < 0.05$). Similar conclusions can be achieved when studying the mentions of the candidates separately.

The correlation signs depend on the media independently of candidates. In order to go deeply in the causes of this behavior, we apply sentiment analysis and topic modeling on the articles.

4.2. Sentiment analysis

To study the connotation with which each candidate is mentioned, we applied the sentiment classifier algorithm described above. The procedure is the following:

(a) In each text, we detected phrases mentioning terms 'Hillary', 'Clinton', 'Donald' or 'Trump'. In the case when more than one candidate is mentioned, we separated the sentences using syntactic analysis.

(b) We applied the sentiment analysis for these sentences and counted the amount of positive, negative, and neutral mentions for each of the candidates. In this step, the deep recursive models for the semantic composition play a central role, since the syntactic analysis of a sentence allows to understand when the text refers to a given candidate in a positive or negative way.

After this procedure, we registered for every day in the studied period, the number of phrases related to each candidate, as well as their sentiment score. Based on this classification, we define a sentiment bias statistic SB (equation (5)), where $\#C_+$ ($\#C_-$) stands for fraction of positive (negative) mentions of Hillary Clinton and $\#T_+$ ($\#T_+$) for positive (negative) mentions of Donald Trump in a given mass media outlet. SB is a measure of the bias towards one of the candidates: if $SB > 0$, the bias is positive towards Clinton compared with Trump, and on the other hand, if $SB < 0$, the bias is positive towards Trump.

$$SB = (\#C_+ - \#C_-) - (\#T_+ - \#T_-) \tag{5}$$

We calculated the value of SB for each media. The results of this analysis are $SB_{NYT} = 0.162 \pm 0.004$ for The New York Times, $SB_{USA} = 0.160 \pm 0.010$ for USA Today, $SB_{CNN} = 0.094 \pm 0.006$ for CNN, and $SB_{FN} = 0.046 \pm 0.005$ for Fox News. In all cases, we reject that SB is a negative value ($p < 0.001$, by bootstrapping [36, 37], see appendix B).

Although in all cases SB is a positive value, the sentiment bias statistic is significantly low for Fox News, while we did not find significant differences between The New York Times and USA Today. In summary, we found that $SB_{NYT}, SB_{USA} > SB_{CNN} > SB_{FN}$. For instance, this suggests that The New York Times, USA Today, and CNN seem to mention Hillary Clinton in a more positive way than Fox News.

The fact that SB is positive for all media suggest that this analysis alone is not enough to understand the behavior of the spread between Clinton and Trump in the polls. The temporal dependence of the sentiment bias statistic ($SB = SB(t)$) will provide a useful insight to understand differences in behavior between both media. In order to go further in this direction we need to analyze first the covered topics and their relative relevance.

4.3. Topic analysis

We classified the news corpus in six topics and called media agenda the relative importance that each media outlet gives to the set of topics, as calculated by equation (2) and defined in [19].

The first topic is about elections in general and it is represented by words like campaign, election, candidate, etc. This result is consistent with the fact that we analyze political news during the campaign period. We choose to discard it given its lack of specificity. The other five topics reveal the subjects discussed during the electoral campaign, which we label and describe briefly as:

- 'Clinton email controversy': covers the famous controversy which Hillary Clinton faced during the elections due to the use of her private email server for official communication.
- 'Clinton Foundation affair': is about the allegations of possible conflicts of interest due to the fact that Clinton was Secretary of State and her foundations accepted foreign donations.
- 'Economy': discusses particularly on taxes, income, jobs and business.
- 'Immigration': is about the discussion of immigration policies between Mexico and USA, raised in the Donald Trump's campaign.
- 'Foreign affairs': deals with the United States foreign policy. In particular, it centers on ISIS and the hypothetical interference of Russians in the electoral process.

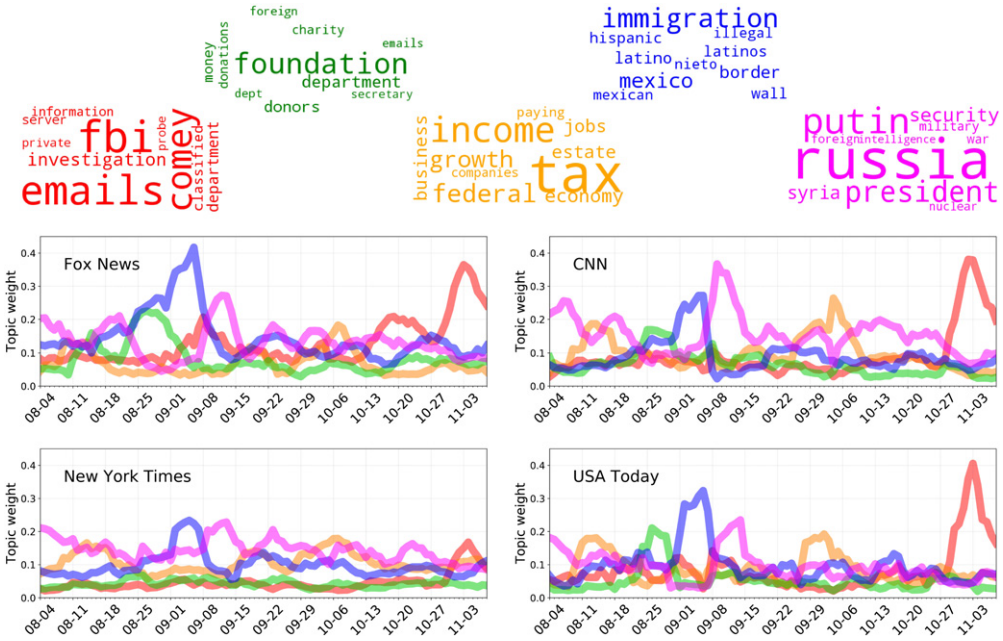

Figure 3. Topic coverage for each media outlet. We show the time evolution of the topics during the whole period, where the difference between these two media outlets can be also observed. Each topic is specified by the word-clouds on top of the figure (and also pointed out in the main text). Topic 1 (red): Clinton email controversy; topic 2 (green): Clinton foundation affair; topic 3 (yellow): economy; topic 4 (blue): immigration; topic 5 (magenta): foreign affairs.

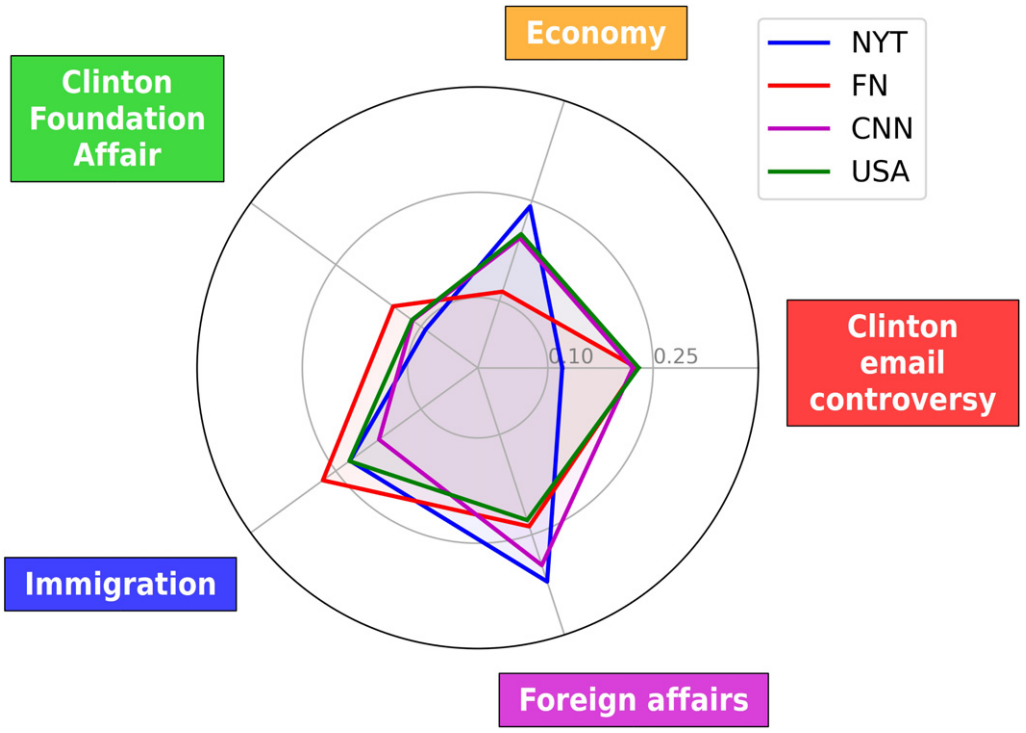

Figure 4. Topic coverage for each media outlet over the whole period. The radar plot shows the accumulative coverage, i.e. the agenda of each media outlet over the five topics.

The keywords which define these five topics are represented in the word clouds of the top panels in figure 3 and correspond to the most significant words which describe the similarity among the news articles grouped in a given topic. It is worth noting that this is an unsupervised method and therefore keywords emerge from the analyzed corpus of news and were not arbitrarily chosen. Figure 3 also shows the temporal evolution of these topics.

Table 1. Linear correlation.

Topic	NYT (SRL)	Fox (SRL)	CNN (SRL)	USA (SRL)
Clinton email controversy	−0.46 (10–20)	−0.42 (11–20)	−0.45 (13–20)	−0.54 (10–20)
Economy	0.56 (4–20)	0.59 (8–20)	0.48 (5–18)	0.40 (10–15)
Clinton foundation affair	−0.53 (3–20)	−0.43 (15–20)	−0.53 (1–20)	−0.40 (5–12)
Immigration	−0.42 (0–12)	−0.44 (5–20)	—	−0.47 (12–20)
Foreign affairs	0.44 (17–20)	—	—	—

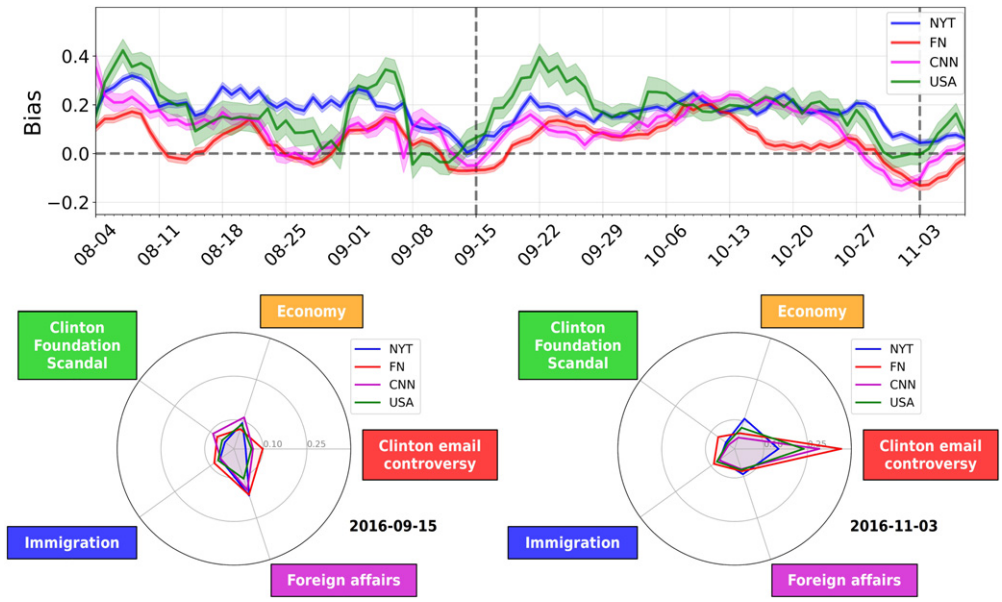

Figure 5. Time series of sentiment bias and radar plots. The curves on top shows time evolution of the sentiment bias statistic SB for each newspaper, New York Times in blue, Fox News in red, CNN in magenta, and USA Today in green. The figures on the bottom are the radar plots of the coverage over the five topics for two specific days: 2016-09-15 and 2016-11-03 (vertical dashed lines in top panel), which correspond to the largest negative values of SB for Fox News and CNN.

The comparative agenda of each media can be easily visualized in figure 4. We observe that the New York Times emphasizes the topics *economy* and *foreign affairs*, while Fox News gives more coverage to *Clinton's affairs* and *immigration*. On the other hand, we can see that both CNN and USA Today cover *economy*, sharing this interest with New York Times, but also pay attention to the topic *Clinton email controversy* as Fox news does.

As we did before with the number of mentions, we calculate the Spearman's correlation for a range of lags between the topic's coverage evolution and the spread between candidates. We found almost all correlation coefficients are significant for lags around 10 or 15 days, except foreign affairs which does not show significant correlation for three of the four media outlets considered (see table 1).

Average Spearman's correlation coefficient between topic coverage and spread between candidates on statistically significant range of lags (SRL) ($p < 0.001$). Non-significant values are not reported.

The results shown in table 1 indicate that the context in which the candidates are mentioned plays a key role in the correlation with the intention to vote. For instance, the time series corresponding to topics *email controversy*, *foundation affair* and *immigration* negatively correlates with the spread between candidates (i.e. the coverage of these three topics worsens Clinton's image), conversely to what happens with the topics *economy* and *foreign affairs* in New York Times. Notoriously, as we can see in figure 4, these two topics are more emphasized by New York Times and CNN, and the first one also by USA Today, while Fox News covers with more intensity the former three, which appears to affect Clinton's intention. These results are consistent with the calculation of SB in the previous section.

4.4. Combining sentiment and topic analysis

Although average sentiment bias statistic (SB) is positive regardless of the media outlet, there exists a period of time where for instance, $SB_{FN}(t) < 0$ or $SB_{CNN} < 0$ (see figure 5). Here we propose a combined analysis in order to better understand this behavior.

Table 2. The sentiment bias per topic.

Topic	SB_{NYT}	SB_{FN}	SB_{CNN}	SB_{USA}
Clinton email controversy	−0.475	−0.429	−0.302	−0.315
Economy	0.332	0.070	0.152	0.168
Clinton foundation scandal	−0.256	−0.257	−0.304	—
Immigration	0.501	0.347	0.306	0.382
Foreign affairs	0.146	0.053	0.115	0.166

In table 2, we calculated the sentiment bias statistic (SB) for each topic, discriminated for media outlet. We observe that in the first three topics the sign of SB matches the sign of the correlation displayed in table 1, while the sign of the last topic matches with the only significant correlation of that table. The only mismatching is in the topic *immigration*. These results reveal that sentiment analysis is much more informative when news articles are decomposed in topics than grouped all together.

The sentiment bias statistic SB calculated with the news of each topic and each newspaper together with the sign of the correlation of the same topic with the difference between Clinton and Trump in the polls. In all cases we reject the hypothesis that SB has an opposite sign with $p < 0.001$. Non-significant values are not reported.

The success of this combined analysis can be seen again in figure 5 where, in addition to the time evolution of SB for each outlet, we can see the radar plots of the agendas for two specifics dates, which belong to periods where $SB_{FN}(t), SB_{CNN} < 0$. In those dates, we can see that the difference between agendas can be partially explained by the topic Clinton emails controversy, which was more emphasized by Fox News and CNN than New York Times, more notorious at the end of the period.

4.5. Causality

In this section, we look for a causal relationship between the spread of polls ($CT(t)$) and the time series of the topics by applying the Granger causality framework described in section 3.2.2. Due to the fact the G is not a stationary series, we start by calculating its first difference ΔCT, which is stationary (augmented Dickey–Fuller test [38], $p < 0.05$) and therefore can be modeled by autoregressive models. By studying the full and partial auto-correlation of ΔCT, we noticed that it is essentially described by a random-walk according to equation (6), with w_t a standard normally distributed random value. We propose two models for the causal analysis: One describe by equation (6), and the other including the information about topics coverage within a certain lag τ (equation (7), where $T_i(t)$ is the weight of topic i at time t).

$$\Delta CT(t) = CT(t) - CT(t-1) = w_t \tag{6}$$

$$\Delta CT(t+\tau) = \beta \cdot \Delta T_i(t) + w_{t+\tau} \tag{7}$$

We say that the coverage of a given topic effectively affects the spread between candidates when the parameter β in equation (7) differs significantly from zero. Since both models involved only first differences, the proper interpretation is that a non-zero value of β implies that the growth or decrease of a given topic predicts a variation in the spread after a certain lag.

The topics with β significantly different from zero ($p < 0.01$), with a sign and lag consistent with the results of linear correlations calculated in previous sections, are: *Clinton's email controversy* ($\beta < 0$ and $\tau = 19$) for Fox News; *economy* ($\beta > 0$ and τ between 11 and 16) for Fox News, CNN, and USA Today; and *Clinton foundation affair* ($\beta < 0$ and $\tau = 19$) and *immigration* ($\beta < 0$ and $\tau = 10$), both for The New York Times.

Finally we would like to remark the role of the topic *Clinton's email controversy*. The analyzed data suggests that this topic plays a key role in the period of time close to the election day. This can be observed in the Fox News and CNN coverage of figure 3 and in the radar plots of figure 5, where this topic had a larger coverage during the last week. Our analysis suggests that, when this topic becomes the most important in news outlets, there is a notorious reduction in the difference between Clinton and Trump. This is the reason why our model reports a causal relationship between this topic and the difference of the surveys. Interestingly, that happens at the end of the period, which suggest that this was a key topic in the electoral result.

5. Conclusions

The influence of mass media on public opinion has been studied from different perspectives and methodologies, ranging from field experiments to data analysis and computational models. The vast availability of data coming from mass media communication and social media makes the analysis techniques based on data be important in the investigation of this kind of issues.

In this paper we suggest a set of tools grounded on natural language processing techniques to be applied in the analysis of the effects that mass media can produce on public opinion. In particular, we are interested in addressing which features of news articles are related to measurable changes in public opinion, using sentiment analysis and topic detection of news articles. Specifically, either the sentiment content of the news articles or the topic of the mentions is important.

Each method on its own, sentiment analysis and topic decomposition, has been widely applied to study related problems. However, to our knowledge, the use of the combination of the two methods to analyze the impact of mass media has not been done yet in systematic way.

By applying the developed methodology to the media coverage of the 2016 US presidential elections, we can understand key aspects in the relationship between mass media and public opinion. In this example, we analyzed news articles in which at least one of the two candidates involved were mentioned. We performed a sentiment analysis and topic detection on that corpus and compared them with measures of vote intention as a proxy of public opinion.

Our approach allows to extract useful insights to this example, as can be seen in the list below:

- The total number of mentions of both candidates in news articles correlates positively with the difference between Clinton and Trump in the polls in the New York Times, CNN, and USA Today, but negatively in Fox News, independently of the candidate.
- The average sentiment analysis of the news articles where both candidates were mentioned is not enough to explain previous behavior.
- The topic analysis, which allows us to appreciate the difference between the agenda of both media outlets, shows that the coverage of given topics are correlated with the difference between Clinton and Trump in the polls.
- The sentiment analysis discriminated by topic is consistent with these last results (except for immigration), given that the topics in favor of Clinton shows positive sentiment towards Clinton and those in favor of Trump are negative towards Clinton (or positive towards Trump).
- There is a causal relation ($p < 0.05$) in the Granger sense between four topics and the difference between Clinton and Trump in the polls, that means that these topics serve as good predictors for the variations in the polls.
- The topic related to the Clinton email controversy seems to be the most relevant because:

1. It negative correlates with the difference between Clinton and Trump in the polls.
2. It has a significant causal relation with the difference between Clinton and Trump in the polls, which, as we said before, determines this topic as a good predictor for the variation in the polls.
3. It explains the negative values of the sentiment bias statistic of Fox News ($SB_{FN}(t)$) and CNN ($SB_{CNN}(t)$).

It is important to notice that the choice of media outlets does not limit the analysis performed in this manuscript, given than we can add as many media as necessary. However, it could be interesting to analyze the role of social networks in the relation between news consumption and public opinion. Finally, it should be mentioned that the methodology implemented in this work is far general and can be easily extended to other topics of interest, not necessarily restricted to a political scenario.

6. Data availability

The data that support the findings of this study are available upon request from the authors.

Acknowledgments

We thank Marcos Trevisan for careful reading of the manuscript and helpful comments. This work was partially funded by the Agencia Nacional de Promoción Científica of Argentina via grant PICT 201-0215.

Appendix A

In this work we analyzed the role of the mass media and its influence in society during elections. Consequently, we selected mass media that are highly consumed by the American population. Therefore, Google trends, an official Google tool that compares the frequency with which different terms are searched, was used. In particular, we compared newspapers and news portals of the United States. The list includes: ABC, Boston Globe, CBS, Chicago Tribune, CNN, Fox News, Houston Chronicle, Los Angeles Times, New York Daily News,

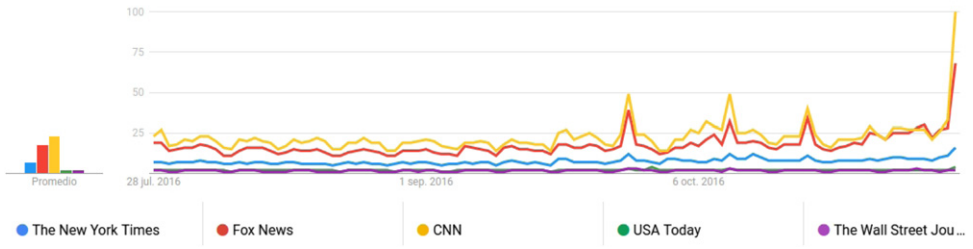

Figure A1. Google Trends of the Mass Medias. Left: the mean relative frequency of searches of five mass media between the 28th of July and the 8th of November; Right: the corresponding time series of the relative frequency of searches. The Google trends search is limited to the United States.

New York Post, Newsday, NPR, Tampa Bay Times, The Dallas Morning News, The Denver Post, The New York Times, The Wall Street Journal, The Seattle Times, USA Today and Washington Post. Among the two most important were Fox News and New York Times, as can be seen in figure A1. The search in Google trends was filtered by geographic location, limiting it only to the United States, and by period of time, from the 28th of July (2016 Democratic National Convention) until the 8th of November (the day of the election).

Appendix B. Bootstrapping

To test the significance of the sentiment bias (SB) calculated in section 4.2, as well its error bars, we employed the bootstrapping technique [36, 37]). It consists on approximating the unknown probability distribution of a given statistic by sampling with replacement from the data. By this way, one can construct confidence intervals in order to test the significance of a given result.

For instance, if we have the following data: 3 positive mentions and 1 negative mention for candidate A, while 2 positive mentions for candidate B and 3 negative mentions, this can be represented as (A+, A+, A+, A−, B+, B+, B−, B−, B−) (where A+ means a positive mention of candidate A). In this case, we would obtain SB $= 1/3$ (taking SB positive as favoring candidate A, see equation (5)). Then, we generate new data sets by sampling with replacement from the original one, as many elements as its length. For example, a generated data set could be (A+, A+, A+, A+, A+, B−, B−, B+, B+), where SB $= 5/9$. By successive repetitions of this process, we finally obtain an approximate distribution for SB.

ORCID iDs

Federico Albanese https://orcid.org/0000-0001-7140-2910
Pablo Balenzuela https://orcid.org/0000-0002-8581-4892

References

[1] McCombs M E and Shaw D L 1972 The agenda-setting function of mass media *Publ. Opin. Q.* **36** 176–87
[2] McCombs M 2005 A look at agenda-setting: past, present and futureless *Knowl. Base Syst.* **6** 543–57
[3] Fortunato J and Martin S 2016 The intersection of agenda-setting, the media environment, and election campaign laws *J. Inf. Pol.* **6** 129–53
[4] Besley T and Burgess R 2002 The political economy of government responsiveness: theory and evidence from India *Q. J. Econ.* **117** 1415–51
[5] Brians L C and Wattenberg M P 1996 Campaign issue knowledge and salience: comparing reception from TV commercials, TV news and newspapers *Am. J. Polit. Sci.* **40** 172–93
[6] Oberholzer-Gee F and Waldfogel J 2009 Media markets and localism: does local news en Espanol boost Hispanic voter turnout? *Am. Econ. Rev.* **99** 2120–8
[7] Gerber A S, Karlan D and Bergan D 2009 Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions *Am. Econ. J. Appl. Econ.* **1** 35–52
[8] King G, Schneer B and White A 2017 How the news media activate public expression and influence national agendas *Science* **358** 776–80
[9] Yasseri T and Bright J 2016 Wikipedia traffic data and electoral prediction: towards theoretically informed models *EPJ Data Sci.* **5** 22
[10] Allcott H and Gentzkw M 2017 Social media and fake news in the 2016 election *J. Econ. Perspect.* **31** 211–36
[11] Soroka S N, Stecula D A and Wlezien C 2015 It's (change in) the (future) economy, stupid: economic indicators, the media, and public opinion *Am. J. Polit. Sci.* **59** 457–74
[12] Lischka J A 2015 What follows what? Relations between economic indicators, economic expectations of the public, and news on the general economy and unemployment in Germany *Journal. Mass Commun. Q.* **92** 374–98
[13] Hopkins D J, Kim E and Kim S 2017 Does newspaper coverage influence or reflect public perceptions of the economy? *Res. Polit.* **4**

[14] De Vreese C H and Boomgaarden H G 2006 Media effects on public opinion about the enlargement of the European union *J. Common. Mark. Stud.* **44** 419–36
[15] Bae Y and Hongchul L 2012 Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular twitterers *J. Am. Soc. Inf. Sci. Technol.* **63** 2521–35
[16] Gorodnichenko Y, Pham T and Talavera O 2018 Social media, sentiment and public opinions: evidence from #Brexit and #USElection (National Bureau of Economic Research) NBER Working Paper No. 24631. Issued in May http://nber.org/papers/w24631
[17] Koltsova O and Koltcov S 2013 Mapping the public agenda with topic modeling. The case of the Russian live journal *Pol. Internet* **5** 1944–2866
[18] Korenvčić D, Ristov S and Šnajder J 2015 Getting the agenda right: measuring media agenda using topic models *Proc. of the 2015 Workshop on Topic Models: Post-Processing and Applications (TM -15)* (New York, NY, USA: Association for Computing Machinery) pp 61–6
[19] Pinto S, Albanese F, Dorso C O and Balenzuela P 2019 Quantifying time-dependent media agenda and public opinion by topic modeling *Phys. Stat. Mech. Appl.* **524** 614–24
[20] Shibanai Y, Yasuno S and Ishiguro I 2001 Effects of global information feedback on diversity: extensions to axelrod adaptive culture model *J. Conflict Resolut.* **45** 80–96
[21] Gonzalez-Avella J C, Eguiluz V M, Cosenza M G, Klemm K, Herrera J L and San Miguel M 2006 Local versus global interactions in non equilibrium transitions: a model of social dynamics *Phys. Rev.* E **73** 046119
[22] Pinto S, Balenzuela P and Dorso C O 2016 Setting the agenda: different strategies of a mass media in a model of cultural dissemination *Phys. Stat. Mech. Appl.* **458** 378–90
[23] Albanese F, Tessone C J, Semeshenko V and Balenzuela P 2019 A data-driven model for mass media influence in electoral context (arXiv:1909.10554 [physics.soc-ph])
[24] RealClearPolitics 2016 https://realclearpolitics.com/
[25] Lehman A, O'Rourke N, Hatcher L and Stepanski E 2005 *JMP for Basic Univariate and Multivariate Statistics: A Step-by-step Guide* (Cary, NC: SAS Institute)
[26] Granger C W J 1969 Investigating causal relations by econometric models and cross-spectral methods *Econometrica* **37** 424–38
[27] Socher R, Perelygin A, Wu J, Chuang J, Manning C D, Ng A and Potts C 2013 Recursive deep models for semantic compositionality over a sentiment Treeban *J Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* pp 1631–42 Anthology ID: D13–1170 https://aclweb.org/anthology/D13–11701
[28] Manning C D, Surdeanu M, Bauer J, Finkel J R, Bethard S and McClosky D 2014 The Stanford CoreNLP Natural Language Processing Toolkit *Proc. of the 52nd Ann. Meeting of the Association for Computational Linguistics, System Demonstrations* (Stroudsburg, PA: Association for Computational Linguistics) pp 55–60
[29] Haribhakta Y and Doddi K 2017 Shriniwas categorization of news articles using sentiment analysis *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2** 52–60
[30] Taboada M, Brooke J, Tofiloski M, Voll K and Stede M 2011 Lexicon-based methods for sentiment analysis *J. Comput. Ling.* **37** 267–307
[31] Muhammad A, Wiratunga N and Lothian R 2016 Contextual sentiment analysis for social media genres *Knowl. Base Syst.* **108** 92–101
[32] Xu W, Liu X and Gong Y 2003 Document clustering based on non-negative matrix factorization *Proc. of the 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval* pp 267–73
[33] Gerlach M, Shi H and Nunes Amaral L A 2019 A universal information theoretic approach to the identification of stop words *Nat. Mach. Intell.* **1** 606–12
[34] Lee D D and Seung H S 1999 Learning the parts of objects by non-negative matrix factorization *Nature* **401** 788
[35] Blei D M, Ng A Y and Jordan M I 2003 Latent Dirichlet allocation *J. Mach. Learn. Res.* **3** 993–1022 http://dl.acm.org/citation.cfm?id=944919.944937
[36] Efron B and Tibshirani R J 1994 An introduction to the bootstrap *Monographs on Statistics and Applied Probability* vol 57 (London: Chapman and Hall)
[37] Efron B others2003 Second thoughts on the bootstrap *Stat. Sci.* **18** 135–40
[38] Seabold S and Perktold J 2010 Stats models: econometric and statistical modeling with python *Proc. of the 9th Python in Science Conf.* vol 57 (SciPy) pp 92–6
[39] Shumway R H and Stoffer D S 2017 *Time Series Analysis and its Applications: With R Examples* (Berlin: Springer) (https://doi.org/10.1007/978-3-319-52452-8)