What matters, context or sentiment?:
Analysing the influence of news in U.S. elections using Natural Language Processing

Federico Albanese¹,†, Sebastián Pinto²,³, Viktoria Semeshenko⁴, and Pablo Balenzuela²,³

¹Instituto en Ciencias de la Computación, CONICET- Universidad de Buenos Aires, Argentina
²Instituto de Física de Buenos Aires (IFIBA), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
³Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
⁴Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. CONICET-Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Av. Córdoba 2122, C1120 AAQ, Buenos Aires, Argentina
†falbanese@dc.uba.ar

September 19, 2019

keywords: text analysis, Mass Media, sentiment analysis, topic detection, time series, news articles

Abstract

A key question in the analysis of collective social behaviour is related to know if and how mass media can influence public opinion. In this paper, we explore quantitatively the relation between a specific manifestation of public opinion and the intention to vote a given candidate, with the information related to the candidates in the Mass Media. We analyse the political news articles related to the US presidential campaign during the year 2016, using techniques of natural language processing. We applied recursive deep models for semantic composition over sentiment treebank to be able to detect the sentiment
of sentences, and topic detection methods in order to characterise how media outlets get involved in the election coverage. The results of the analysis were compared to the outcomes of political polls in order to know which of these two aspects of the information have more influence and if there exists any causal relationship between them. Our results suggest that the sentiment content of the news by itself is not enough to understand the differences in the polls between both candidates (validating the popular quote “There is no such thing as bad publicity”) but the topic coverage distribution does and the sentiment content discriminated by topic is consistent with these results. In particular, we found that one topic (news related to Clinton’s email scandal) is the more relevant in this analysis because it correlates negatively and shows a significant causal relation with the difference between Clinton and Trump in the polls and also explain the negative values of sentiment analysis in Fox News coverage.

1 Introduction

The more popular recent view is that mass media influence is significant. The mass media provide facts and events about what is happening in society. Media interpret events and issues and meaning so that individuals understand their role in society. Therefore, media do not only have the capacity to reflect reality through their own perspective, but also can determine the importance they attribute to the different topics [1]. The term that best fits this description is agenda setting, which means mass media do not tell you what to think, but what to think about. In other words, mass media tell you what is and what is not important and to what extent. The way to control attitudes is to provide a partial selection of information to process, because the way to influence what individuals think is precisely shape what they think about. These agenda setting strategies can influence the political opinion of citizens in electoral contexts.

Previous research shows how the public’s view of a political event can be modified by mass media [2]. This influence is manifested on the basis of topics emphasised by media, and those that are omitted [3]. In addition, several studies demonstrated that reading different media in a sustained manner leads individuals to modify their political ideology and makes them vote in a different way than they did in the past, aligning their votes in lines with the editorial viewpoint of the newspaper [4]. For instance, Gerber and Dean [5] studied explicitly what is the influence that reading newspapers may have on the political opinion of its readers. The authors observed that during the governor elections of the year 2005 in the state of Virginia, reading the
newspaper did not generate an increase in the absolute number of voters, however it did influence the votes qualitatively. The authors also detected that the news produce changes in the perception that audience has about politicians. King et al. [6] analysed how the number of tweets increased on a specific topic after individuals have been exposed to pertinent news. The authors concluded that the conversation in Twitter social network was 20% higher one day after the news was published, and found that the effects were similar regardless of gender, religion or ideology.

On the other hand, Yasseri and Bright [7] proposed the number of mentions of a given candidate, based on visits to Wikipedia site, as the best predictor of the results of elections. However, the authors did not take into account whether mentions of political parties and candidates are positive or negative. Sentiment of news (positive or negative connotation) has been explored within the framework of how the tone is related to the variation of economic indicators [8], how it affects public expectation about economy [9, 11], or how it shapes the public opinion about a given prominent issue [11]. It has also been addressed how the sentiment of bots’ or influential users’ tweets induces the tone in which their followers express in Twitter [12, 13].

In this paper, we study the relationship between the mass media coverage and the political opinions during the US presidential campaign in 2016, using methods for text mining and for analysing time series data. We consider here the number of mentions of each candidate, the sentiment content of the news articles where they are mentioned and the evolution of the relative coverage of a set of topics of the same articles (political media agenda). In particular, we would like to know the following:

1. There are exist correlations between the intention to vote a given candidate and the information related to them in mass media?

2. Is the sentiment content of the news articles or a given set of topic of the media agenda what better explains the fluctuations in the polls?

3. Is there any significant causal relation between topics and polls?

In section 2 we describe the data used in the study. In section 3 we describe the text mining tools applied to the article news. In section 4 we analyse the time series of the polls with the number of mentions of the candidates in the news articles, the sentiment analysis of the news content, the topic evolution of the political media agenda, the sentiment analysis discriminated by topic and the calculation of Granger causality. Finally, we discuss the results in section 5.
2 Data Collection

In this section we describe the data sources employed in our analysis. We retrieve data from two different mass media outlets, The New York Times and Fox News, and opinion polls. We provide an overview of these sources and describe how the data was collected.

2.1 Mass Media Data

We analysed news articles from the online editions of The New York Times and Fox News during the electoral period of 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. 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 corpus analysed is made up by 9664 news articles: 3915 from The New York Times and 5749 from Fox News. These mass media have been selected due to their national widespread coverage and their different political orientations. The New York Times was the most searched online newspaper in all the states during the 2016 election campaign [14]. This media was classified as a democrat media, based on the frequency of conservative or liberal arguments used in its articles, as well as the Fox News portal was classified as a republican political orientation media [15].

2.2 Polls

In order to quantify the change of public opinion during the election period, we based our analysis on a total of 263 national surveys conducted by different agencies (an average of 2.7 surveys per day), in which the forecaster of vote 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 and the date that it was done, was downloaded from Real Clear Politics website [16] using a python web scraper. All national surveys that are presented in this work used a demographic balanced sample.

Figure 1 displays the time series of potential percentage of votes for each candidate and the difference between these time series (the percentage of Clinton minus the percentage of Trump). Each point in the time series corresponds to a particular date, but its value is the average of the entire week before. In other words, a 7-days sliding window average was used and
in consequence each point in the time series takes into account an average of 19 polls.

Figure 1 also shows that Clinton kept up an advantage over Trump during the time period. However, this advantage was affected by drops in the middle and next to the end of the period. If we inspect the individual series, we can see that the initial decreasing of the advantage is due to the gradual ascending of the Trump’s image until October. After that, his image went down sharply until it was recovered near the election date, which explains the last decrease in the Clinton’s advantage, which slightly increased in the last days. The aim of this paper is to explore how mass media contributed to the formation of these fluctuations and try to infer if and how the mass media can influence in the vote of the citizens.

3 Text Mining Methods

In this section we describe the text mining techniques which were used to extract useful and relevant information from the news articles, and characterise the behaviour of media during the electoral period. We assess the choices of sentiment and frames adopted by the articles with respect to both candidates by determining the connotation with which each candidate was mentioned (sentiment analysis), and the main topics in which candidates were involved (topic analysis). The relationship between media and polls will be explored through Spearman’s correlations \cite{17} and Granger-causality \cite{18} tests.

3.1 Sentiment Analysis

In order to understand if the frequency of positive and negative mentions of a candidate leaves impact on his performance in the surveys, a sentiment analysis algorithm was adapted and implemented.

The sentiment analysis was performed utilising deep recursive models for the semantic composition applied to sentiment trees \cite{19}, in particular by the Stanford CoreNLP implementation of natural language processing \cite{20}. Sentiment analysis with CoreNLP is very straightforward. The 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. Using deep machine learning techniques, 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 root node which also corresponds to the total phrase.

Even though there exist other alternative algorithms to perform sentiment analysis, such as those based on the extraction of characteristics of the sentences [21] or lexicon-based approaches to opinion mining [22, 23], the Stanford CoreNLP is suitable for analysing our corpus of news given that each sentence is well-formulated with correct grammar and spelling and it uses this exact information to determine the sentiment of a phrase.

3.2 Topic Detection

As previously mentioned, mass media set the agenda by either mentioning or omitting certain news, or by covering certain topics or not [3]. When the audience identifies what is relevant in the topics, or mass media tell its audience what is important and what is not, these can be thought as agenda-setting [1, 24]. In order to understand how the coverage of the main topics affects the candidates’ performance during elections, it is important to identify those topics and make the follow up over time. Therefore, we classify documents as thematic topics using unsupervised learning techniques following the procedure sketched in [25].

The articles are described as numerical vectors through the term frequency - inverse document frequency (tf-idf) representation [26]. The dimension of the vectors \((t = 49172)\) is given by the total amount of words of the corpus after removing non-informative ones such as prepositions and conjunctions.

Once the document vectors are constructed, we put them together in a document-term matrix \((M)\), which has dimensions of number of documents in the corpus \((d)\) by number of terms \((t)\). The next step is to find how the news articles can be grouped in clusters (called topics), according to some similarity criteria. A topic is defined as a group of similar articles which roughly talks about the same subject.

In order to detect the main topics in the corpus, we perform non-negative matrix factorization (NMF) [26, 27] on the document-term matrix \((M)\). Similar results were obtained by applying Latent Dirichlet Allocation (LDA) method [28] on the same corpus.

Once the articles were grouped in topics, we can estimate their coverage by calculating the topic’s weights. The weight of topic \(i\) \((W_i)\) is defined as the combination of the amount of news articles of the topic (weighted by their degree of membership) times the length of the article. It can be defined on a daily basis (time-dependent distribution) or for the whole period (average distribution). For a single day \(d\), it is sketched in the Eq. 1, where \(l(j)\) is the number of words of 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 \(\delta\)
is the Kronecker delta. Providing by the fact that each document vector can have all non-zero components, it is allowed that a document contributes to more than one topic weights. In order to reduce noise, we apply a linear filter with a three day wide sliding window, and finally we normalize the temporal profiles.

\[
W_i(d) = \sum_j l(j) \cdot h_{ji} \cdot \delta_{d_i,d} 
\]  

(1)

4 Results

Applying the procedures described above, we analyse the data in order to infer if and how mass media influence the citizens’ political opinions. Does the sentiment of a statement matter? Is the topic of the statement the crucial part of the interaction? In this section we present the results of the performed analysis.

4.1 Total number of mentions

As a first approach, we simply seek to compare the time series of the surveys against the total number of mentions of each candidate in both media. These curves can be seen in Figure 2. In this first analysis we completely ignore the context in which the phrases appear and also if the mention is in favour or against the candidates.

We calculate the Spearman’s correlation coefficient [17] between the total number of mentions of each candidate (Fig. 2) with the spread between polls data (Fig. 1). This coefficient is a measure of the rank relationship between two variables and it is independent of the escalation of the variables. In particular, we used the python-scipy implementation [29]. From all time series, we previously remove the linear trend if there was any. At the time of calculating the correlations one must take into account that changes in media coverage may not be instantaneously reflected in the polls, either because it is not clear the time scale of media influence or because the publication date of the polls if posterior to the data collection. Therefore, we seek for a range of lags when most of the correlations are significant.

We found that the number of mentions of both candidates in New York Times positively correlates with the spread between Clinton and Trump in the polls (on average over a range of lags, the correlation coefficient is 0.646). This means that when mentions increase, the difference Clinton minus Trump increases. Meanwhile these correlations are negative for similar time series
in Fox News (average correlation coefficient $-0.476$), i.e., when mentions increase, the difference Clinton minus Trump decreases. These results are statistically significant ($p < 0.001$) for a lag between 6 and 18 days for both newspapers, and similar conclusions can be achieved when studying the mentions of the candidates separately.

In conclusion, the sign of the correlation depends on the Mass Media outlet but not on the candidate. These results are in line with previews works \cite{15} where New York Times seems to show a clear preference in favour of the democratic candidate whereas Fox News shows a bias to the republican candidate. However, these biases could have different causes. For example, it might be the case that New York times portrays Hillary Clinton in a positive way and the Republican candidate in a negative way. Instead, for instance, it could be due to the topic the articles are covering. In order to understand the cause of these biases, we apply sentiment analysis and topic modelling.

4.2 Sentiment analysis

In order to study the positive and negative connotation of each candidate we applied the sentiment classifier algorithm to news articles selected previously. The procedure was the following:

1. 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.

2. 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 us to understand when the text refers to a given candidate in a positive or negative way.

In this way we can register, for every day in the studied period, the number of phrases related with each candidate as well as their sentiment score. Based on this, we define a Sentiment Bias statistic $SB$ in equation $2$ 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.

$$SB = (\#C_+ - \#C_-) - (\#T_+ - \#T_-)$$ (2)
The Sentiment Bias statistic $SB$ is a measure of the bias towards one of the candidates. If $SB > 0$, the bias is positive towards Clinton compared with Trump. If $SB < 0$ is the same towards Trump.

We calculated the accumulated value of $SB$ for both media (by taking into account all the articles). The results of this analysis is $SB_{NYT} = 0.162\pm0.008$ for New York Times and $SB_{FN} = 0.046\pm0.008$ for Fox News. While in both cases $SB$ is a positive value (which will be discussed below), this statistic bias is larger in the case of New York Times, indicating that is significantly more positive to Hillary Clinton than Fox News. In all cases, the statistical significance and the error bars were obtained by the bootstrapping method [30, 31], consisted on sampling with replacement from the original distribution of $\#C^+, \#C^-, \#T^+$ and $\#T^+$ of each news outlet and recalculating the statistic $SB$.

The fact that the accumulated value of the Sentiment Bias statistic is positive for both media can be interpreted as the sentiment by itself is not relevant to understand the results obtained in the previews section, where the sign of the correlation of both candidates mentions with the spread between Clinton and Trump in the polls depend on the news outlet (and were positive for New York Times and negative for Fox News).

However, if we analyse the temporal dependence of the Sentiment Bias statistic ($SB$), we will see that meanwhile the time series for the New York Times is consistently positive (which means that the sentiment is positive towards Clinton), the corresponding to Fox News show periods of time where it takes negative values (which means positive bias towards Trump).

These results suggest that something is missing in the relation between news and polls, and this is the context in which both candidates are mentioned. The Agenda Setting of Mass Media is vitally important during political elections in order to understand the behaviour of the society in the polls previous to election as we will see in next subsections.

### 4.3 Topic analysis

Following the procedure described in section 3.2, we decompose the news corpus depicted in section 2 in six topics or related groups of news. The reason behind factorising the corpus in six topics was based on having a low dimensional representation of the corpus and a clear interpretation of the topics due to our prior knowledge of the political background. We found that this factorisation allowed us to draw useful conclusions. However, more sophisticated methodologies to estimate the number of topics in a corpus can be taken into account in future researches.

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 analyse political news during the campaign period, but we choose to discard it given its lack of specificity.

The other topics are very informative and reveals the subjects discussed during the electoral campaign, which we label and briefly describe as:

- “Clinton email controversy”: Covers the famous controversy Hillary Clinton faced during the elections she used her private email server for official communication.
- “Clinton Foundation Scandal”: This is about the allegations of possible conflicts of interest due to the fact that Clinton was secretary of state and the foundations accepted foreign donations.
- “Economy”: It discuss particularly on taxes, income, jobs and business.
- “Immigration”: It is about the fortified wall between Mexico and USA proposed by Donald Trump to stop Latin illegal immigration.
- “Foreign affairs”: It deals with the case if the Russian government interfered in the election in order to increase political instability and damage Hillary Clinton’s campaign [32, 33].

The keywords which define these five topics are represented in the word clouds of the top panels in Figure 3.

In Figure 4, the radar plot shows the accumulated coverage and figure 3 shows the evolution of the topics. We can observe how The New York Times emphasises the topics Economy and Foreign Affairs, while Fox News gives more coverage to Clinton’s scandals and Immigration. Figure 4 shows that agendas are consistent with the expected style of each media outlet during the electoral period.

Similarly to the total number of mention’s correlation study depicted in subsection 4.1, we calculate the Spearman’s correlation coefficient and its p-value for a range of lags between the time series of the topics and the spread between Clinton and Trump of the polls. We found that most of the correlation coefficients are significant for a lag range of about 8 to 20 days (except topic Immigration for NYT), on which we calculated the average correlation coefficient. The results can be seen in Table 1.

The results shown in Table 1 indicate that the context in which the candidates are mentioned in news articles plays a key role in order to understand the behaviour of the polls. For instance, the time series corresponding to topics Email controversy, Foundation Scandal and Immigration negatively
Table 1: **Linear correlation.** Average Spearman’s coefficient between topic coverage and polls difference over a statistically range of lags (SRL), with $p < 0.001$. Non-significant values are not reported.

| Topic                           | NYT (SRL) | Fox (SRL) |
|---------------------------------|-----------|-----------|
| Clinton email controversy       | -0.47 (8-20) | -0.46 (11-20) |
| Economy                         | 0.59 (5-20)  | 0.60 (8-20)  |
| Clinton Foundation Scandal      | -0.51 (3-20) | -0.46 (14-20) |
| Immigration                     | -0.45 (0-13) | -0.43 (7-20)  |
| Foreign affairs                 | 0.49 (17-20) | -           |

correlates with the polls difference (i.e. the coverage of these three topics affects Clinton’s image), conversely to what happens with the topics *Economy* or *Foreign Affairs*. This means that each topic has a positive or negative connotation, regardless the media it belongs to. But the role that the different media outlets play is setting the agenda they introduce to their audience. As was shown in the radar plot in Figure 4, the accumulative coverage of each topic during the electoral period differs for each media outlet. Topics like the Clinton’s foundation scandal, email leaking scandal or immigration were clearly more emphasised by Fox News. Consistently, these topics are the three that favour Donald Trump, as can be seen by the negatives correlations in table 1. On the other hand, topics such as the relationship with Russia, taxes, and general themes about economy, have a great coverage in The New York Times, which favours Hillary Clinton and emphasised the democrat orientation of this medium.

### 4.4 Combining Sentiment and topic analysis

Even though the average Sentiment Bias statistic (SB) is always positive regardless the media outlet, the fact that this quantity is larger for the New York Times than Fox News ($SB_{NYT} > SB_{FN}$) and the existence of a period of time where $SB_{FN} < 0$ suggest that a sentiment analysis based on topic discrimination could help to understand these behaviours.

The Sentiment Bias statistic (SB) for each topic, discriminated for the New York Times and Fox News, are reported in Table 2. If we compare with Table 1, we can observe that in four of the five topics (excluded Immigration) the sign of the correlation match with the sign of the Sentiment Bias statistic (SB) (in all cases we reject the hypothesis that $SB$ has an opposite sign with $p < 0.001$). It means that those topics that favours Clinton have positive sentiment towards Clinton, meanwhile those in favour of Trump have negative sentiment towards Clinton (or positive towards Trump). The only exception
Table 2: The sentiment bias per topic. The 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 $p < 0.001$ is the Immigration topic, for which the results are not consistent with the others.

The power of this combined analysis can be seen in Figure 5, where it can be observed in top panel, the evolution of $SB$ for both newspapers (New York Times in blue and Fox News in red). Here, $SB_{NYT}(t)$ is always positive, while $SB_{FN}(t)$ shows negative regions. This result can be interpreted as New York Times positively supporting the democratic candidate and being negative on the republican candidate. In contrast, Fox News sentiment bias takes negative values (supporting Donald Trump and criticising Hillary Clinton), specially the last week before the elections. In low panel of same figure, we can see the radar plots with the agenda of each newspaper for two specifics dates corresponding to regions where $SB_{FN}(t) < 0$: In both cases the difference between both agenda correspond to the topic related to the Clinton emails controversy which is emphasised by Fox News more than the New York Times. Summarising, when the Sentiment Bias statistic of the Fox News ($SB_{FN}$) became negative (positive sentiment towards Trump and/or negative to Clinton) is because the agenda is largely dominated by the Clinton emails controversy.

4.5 Causality

The core of this section is to go further than simply detecting correlations between time series and try to find if there exist any causal relation between some of the topics coverage time series and the polls performance of the involved candidates. Within the limitation of the data, we will say that there exist a causal relation we mean that the information about the coverage of the topics is necessary to be able to explain the variations observed in the polls. This is the traditional framework where a causal relation is settled.
down, known as Granger causality [18].

We need to model the spread of polls between Clinton and Trump as a function of time in order to perform a causal test. Due to the fact that these are not stationary series but the first difference is (Augmented Dickey-Fuller test [34], $p < 0.05$), we studied the full and the partial auto-correlation of the first difference, and noticed that the spread is essentially a random-walk described by Eq. (3), where \( w_t \) is a standard normally distributed random value. In order to proceed with the causal analysis, we add to Eq. (3) a factor which includes the information about topics coverage within a certain lag. The topic information is also introduced as a first difference. This new model is referred in Eq. (4).

\[
\Delta \text{gap}(t) = \text{gap}(t) - \text{gap}(t - 1) = w_t
\]

\[
\Delta \text{gap}(t + \text{lag}) = \beta \cdot \Delta \text{topic}(t) + w_{t+\text{lag}}
\]

We say that the coverage of a given topic effectively affects the spread when the parameter \( \beta \) in Eq. (4) differs from zero significantly. In other words, including information of a topic improves the model of the spread series. Notice that when \( \beta \) is zero, the model of Eq. (3) is recovered. Due to the fact that the first differences are already involved in the models, the proper interpretation is that a non-zero value of \( \beta \) implies that the growth or decrease of a given topic “causes” a variation in the spread after a certain lag.

The values of \( \beta \) significantly different from zero ($p < 0.05$) found in our analysis belong to the topics: Clinton’s email controversy, with a negative value of \( \beta \) for Fox News and a lag of 19; Economy, with a positive value of \( \beta \) for Fox News and a lag of 16; and Clinton Foundation Scandal, with a negative value of \( \beta \) for New York Times and a lag of 11. Both the sign and the corresponding lags are consistent with the results of linear correlations calculated in previous sections. However, considering the fact that a larger \( p \) has to be used, this results are not as robust as the other analyses in this paper.

In particular, Clinton’s email controversy become one of the most relevant topics at the end of the studied period and therefore closer to the election day. This relevance is shown in the time series of Figure 3 and it is specially observed in the Fox News coverage. It is also clear in the radar plots of Figure 5 where it is showed that the topic has a clearly bigger coverage in the last week. Our analysis suggest that when this topic has a large fluctuation in the coverage level, an important reduction occurs in the difference of the surveys and this is the reason why our model reports a causal relationship between
that this topic and the difference of the surveys. Curiously, that happens at the end of the period with which we could conclude that it was an important factor in the determination of the electoral result.

5 Discussion

The interplay between mass media and opinion formation is a rather complex one. However, we demonstrated the existence of an explicit relationship between different aspects of the mass media coverage and the political opinion in society during the 2016 US presidential elections. Applying natural language processing algorithms to news articles, we got some insights about the nature of this relationship and can conclude the following:

• 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 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 behaviour.

• The topic analysis, which allow 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 results (except for Immigration), given that the topics in favour of Clinton shows positive sentiment towards Clinton and those in favour of Trump are negative towards Clinton (or positive towards Trump).

• There is a causal relation ($p < 0.05$) between three topics and the difference between Clinton and Trump 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.
3. It explains the negative values of the Sentiment Bias statistic of Fox News ($SB_{FN}(t)$)

Even though it is not possible discard other influence in the behaviour of the polls, we have demonstrated how a combined analysis of sentiment content and agenda of the media allow us to observed the influence of Mass Media in public opinion.

References

[1] McCombs, Maxwell E and Shaw, Donald L. The agenda-setting function of mass media. Public opinion quarterly, vol. 36, no. 2, pp. 176187, 1972.

[2] Besley, Timothy and Burgess, Robin The political economy of government responsiveness: Theory and evidence from India. The Quarterly Journal of Economics, vol. 117, no. 4, pp. 14151451, 2002.

[3] Brians, Craig Leonard and Wattenberg, Martin P. Campaign issue knowledge and salience: Comparing reception from TV commercials, TV news and newspapers. American Journal of Political Science, pp. 172-1993, 1996.

[4] Oberholzer-Gee, Felix and Waldfogel, Joel Media markets and localism: Does local news en Espanol boost Hispanic voter turnout?. The American economic review, vol. 99, no. 5, pp. 21202128, 2009.

[5] Gerber, Alan S and Karlan, Dean and Bergan, Daniel. Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. American Economic Journal: Applied Economics, vol. 1, no. 2, pp. 3552, 2009.

[6] King, Gary and Schneer, Benjamin and White, Ariel How the news media activate public expression and influence national agendas. Science, vol. 358, no. 6364, pp. 776780, 2017.

[7] Yasseri, Taha and Bright, Jonathan. Wikipedia traffic data and electoral prediction: towards theoretically informed models. EPJ Data Science, vol. 5, no. 1, pp. 115, 2016.

[8] Soroka, Stuart N and Stecula, Dominik A and Wlezien, Christopher. It’s (change in) the (future) economy, stupid: economic indicators, the media, and public opinion. American Journal of Political Science, vol. 59, no. 2, pp. 457474, 2015.
[9] Lischka, Juliane A. *What follows what? Relations between economic indicators, economic expectations of the public, and news on the general economy and unemployment in Germany, 2002-2011*. Journalism & Mass Communication Quarterly, vol. 92, no. 2, pp. 374-398, 2015.

[10] Hopkins, Daniel J and Kim, Eunji and Kim, Soojong. *Does newspaper coverage influence or reflect public perceptions of the economy?*. Research & Politics, vol. 4, no. 4, 2017.

[11] De Vreese, Claes H and Boomgaarden, Hajo G. *Media effects on public opinion about the enlargement of the European Union*. CMS: Journal of Common Market Studies, vol. 44, no. 2, pp. 419–436, 2006.

[12] Bae, Younggue and Lee, Hongchul. *Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers*. Journal of the American Society for Information Science and Technology, vol. 63, no. 12, pp. 2521-2535, 2012.

[13] Gorodnichenko, Yuriy and Pham, Tho and Talavera, Oleksandr. *Social media, sentiment and public opinions: Evidence from # Brexit and # USElection*. JNational Bureau of Economic Research, 2018

[14] Google Trends of U.S. newspapers and Mass Media.

[15] Datascience Berkeley Staff. *Exploring Political Bias with the Bitly Media Map*. 2013.

[16] Real Clear Politics. https://realclearpolitics.com/.

[17] Lehman, Ann. *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Institute, 2005.

[18] Granger, Clive WJ. *Investigating causal relations by econometric models and cross-spectral methods*. Econometrica: Journal of the Econometric Society, pp. 424-438, 1969.

[19] Socher, Richard and Perelygin, Alex and Wu, Jean and Chuang, Jason and Manning, Christopher D and Ng, Andrew and Potts, Christopher. *Recursive deep models for semantic compositionality over a sentiment treebank*. JProceedings of the 2013 conference on empirical methods in natural language processing, pp. 1631-1642, 2013.
[20] BManning, Christopher D and Surdeanu, Mihai and Bauer, John and Finkel, Jenny Rose and Bethard, Steven and McClosky, Davi. The stanford corenlp natural language processing toolkit. ACL (System Demonstrations), pp. 55-60, 2014.

[21] Doddi, Kiran S and Haribhakta, Mrs YV and Kulkarni, Parag. Sentiment Classification of News Article. Diss. College of Engineering Pune, 2014.

[22] Taboada, Maite and Brooke, Julian and Tofiloski, Milan and Voll, Kimberly and Stede, Manfred. Lexicon-based methods for sentiment analysis. JComputational linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[23] Muhammad, Aminu and Wiratunga, Nirmalie and Lothian, Robert. Contextual sentiment analysis for social media genres. Knowledge-based system, vol. 108, pp. 92-101, 2016.

[24] McCombs, Maxwel. A look at agenda-setting: Past, present and futures. Knowledge-based system, vol. 6, no. 4, pp. 543-557, 2005.

[25] Sebastin Pinto and Federico Albanese and Claudio O. Dorso and Pablo Balenzuela. Quantifying time-dependent Media Agenda and public opinion by topic modeling. Physica A: Statistical Mechanics and its Applications, vol. 524, pp. 614-624, 2019.

[26] Xu, Wei and Liu, Xin and Gong, Yihong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 267-273, 2003.

[27] Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 788, 1999.

[28] Blei, David M and Ng, Andrew Y and Jordan, Michael I. Latent dirichlet allocation. Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[29] Jones, Eric and Oliphant, Travis and Peterson, Pearu. {SciPy}: open source scientific tools for {Python}. 2014.

[30] Efron, Bradley and Tibshirani, Robert J. An introduction to the bootstrap. CRC press 1994.

[31] Efron, Bradley and others. Second thoughts on the bootstrap. Statistical Science, vol. 18, no.2, pp. 135–140, 2003.
[32] Office of the Director of National Intelligence *Assessing Russian Activities and Intentions in Recent US Elections*. National Intelligence Council Washington, DC, 2017.

[33] Torres Soriano, Manuel R. *HACKEANDO LA DEMOCRACIA: OPERACIONES DE INFLUENCIA EN EL CIBERESPACIO*. ININVESTAM 2017.

[34] Seabold, Skipper and Perktold, Josef. *Statsmodels: Econometric and statistical modeling with python*. 9th Python in Science Conference 2010.
6 Figures

Figure 1: Survey curves of 2016 US presidential elections. The figure on the top shows the curves of Hillary Clinton (Democrat candidate) and Donald Trump (Republican candidate) for the 2016 presidential election. The figure on the bottom shows the spread between these curves.
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 (New York Times on the top and Fox News on the bottom).
Figure 3: **Topic coverage for both The New York Times and Fox News.** We show the time dependency 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 Scandal; Topic 3 (yellow): Economy; Topic 4 (blue): Immigration; Topic 5 (magenta): Foreign Affairs.
Figure 4: **Topic coverage for both The New York Times and Fox News.** The radar plot shows the accumulative coverage within the whole period, i.e. the agenda of each media outlet over the five topics.
Figure 5: **Time series of sentiment bias and box plots** The curves on top show the statistic $SB$ for each newspaper, New York Times in blue and Fox News in red. 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.