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

Pablo Balenzuela, Viktoriya Semeshenko, Sebastián Pinto & Federico Albanese

SoPhy Lab,

Physics Department,

FCEyN, University of Buenos Aires

Agenda-Setting Theory

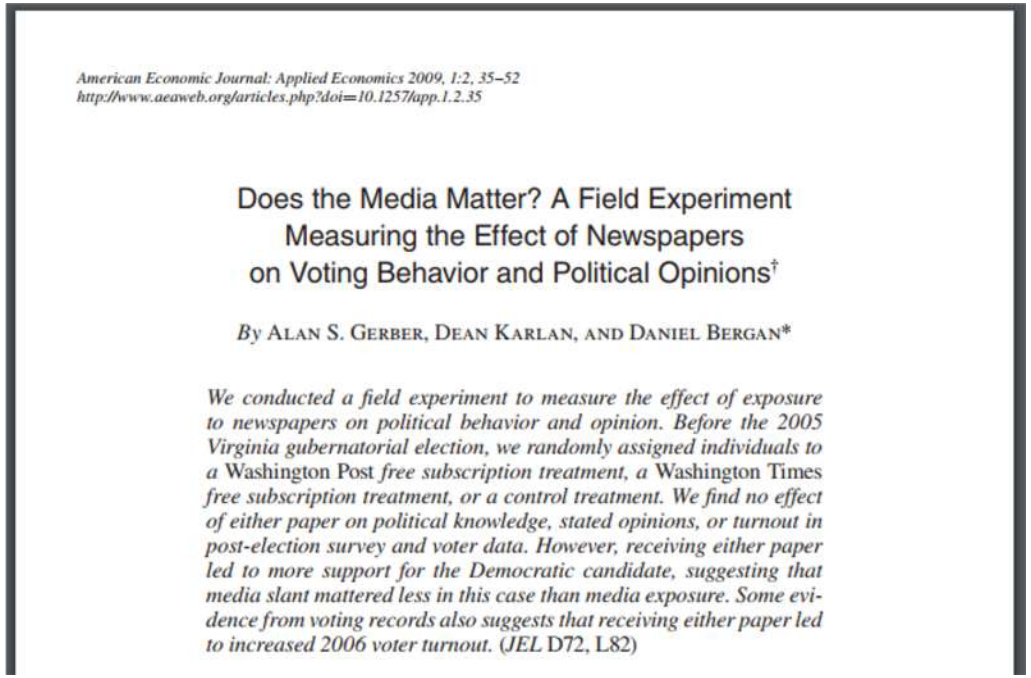
- How the media influence public opinion is a question usually faced and analyzed within the context of **Agenda Setting theory** (*Agenda-Setting, McCombs 1972*).

- Media outlets are represented by their agenda: compilation of topics presented with different emphasis.

- The theory, in its basic version, studies the correlation between the topics covered by the media and those that are prominent in the public.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, *36*(2), 176-187.

Introduction:
Field experiments



American Economic Journal: Applied Economics 2009, 1:2, 35–52
http://www.aeaweb.org/articles.php?doi=10.1257/app.1.2.35

Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions†

By Alan S. Gerber, Dean Karlan, and Daniel Bergan*

We conducted a field experiment to measure the effect of exposure to newspapers on political behavior and opinion. Before the 2005 Virginia gubernatorial election, we randomly assigned individuals to a Washington Post *free subscription treatment, a* Washington Times *free subscription treatment, or a control treatment. We find no effect of either paper on political knowledge, stated opinions, or turnout in post-election survey and voter data. However, receiving either paper led to more support for the Democratic candidate, suggesting that media slant mattered less in this case than media exposure. Some evidence from voting records also suggests that receiving either paper led to increased 2006 voter turnout. (JEL* D72, L82)

Field Experiment: (Governor elections in Virginia, 2005)

Three groups in Prince William County, Virginia:

- Group 1: free suscription (15/10 - 8/11) to Washington Post.
- Grupo 2: free suscription (15/10 - 8/11) to Washington Times.
- Group 3: Control

Subjects: registered voters. People already suscribed to a given newspaper were excluded.

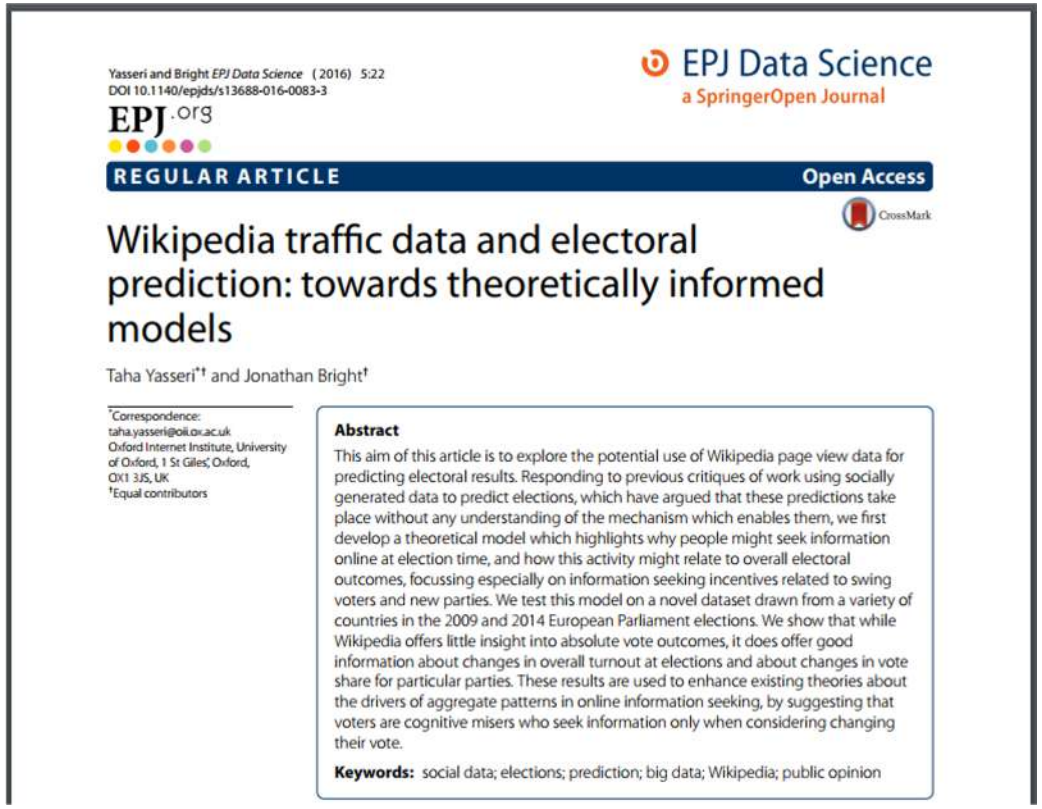
Post-elections survey: Current events, news, politics, political leanings, whether they voted, for whom.

Conclusions:

- While reading the newspaper did not increase the number of people who voted, it did influence who voted. They saw: 8% were more likely to vote for Democrats regardless of the newspaper they read, suggesting that media bias mattered less in this case than media exposure.

Gerber, A. S., Karlan, D., & Bergan, D. (2009). 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.

Introduction: Web data analysis



Yasseri and Bright *EPJ Data Science* (2016) 5:22
DOI 10.1140/epjds/s13688-016-0083-3
EPJ.org
EPJ Data Science
a SpringerOpen Journal
REGULAR ARTICLE
Open Access

Wikipedia traffic data and electoral prediction: towards theoretically informed models

Taha Yasseri*† and Jonathan Bright†

*Correspondence:
taha.yasseri@oii.ox.ac.uk
Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK
†Equal contributors

Abstract

This aim of this article is to explore the potential use of Wikipedia page view data for predicting electoral results. Responding to previous critiques of work using socially generated data to predict elections, which have argued that these predictions take place without any understanding of the mechanism which enables them, we first develop a theoretical model which highlights why people might seek information online at election time, and how this activity might relate to overall electoral outcomes, focussing especially on information seeking incentives related to swing voters and new parties. We test this model on a novel dataset drawn from a variety of countries in the 2009 and 2014 European Parliament elections. We show that while Wikipedia offers little insight into absolute vote outcomes, it does offer good information about changes in overall turnout at elections and about changes in vote share for particular parties. These results are used to enhance existing theories about the drivers of aggregate patterns in online information seeking, by suggesting that voters are cognitive misers who seek information only when considering changing their vote.

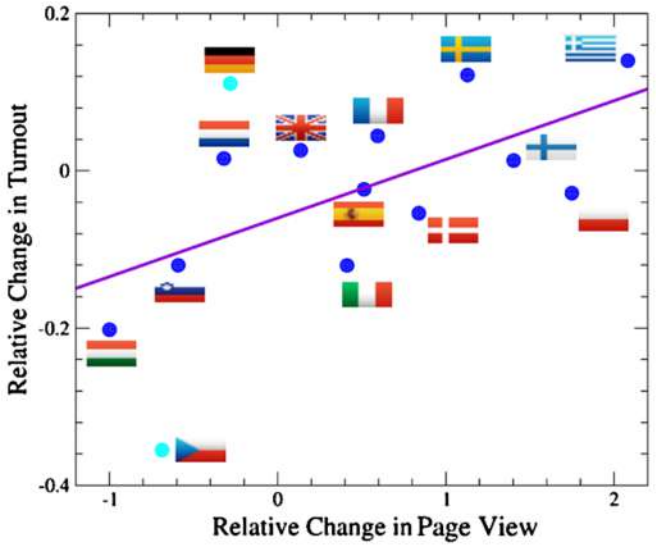
Keywords: social data; elections; prediction; big data; Wikipedia; public opinion

Results suggest that voters seek information only when they are considering changing their vote.

Analysis of information consumption on Wikipedia sites related to different European elections. Increases in the consumption of these pages correlates with:

- Increased participation in elections
- Increase of votes in given political parties.

2009 & 2014 European Elections



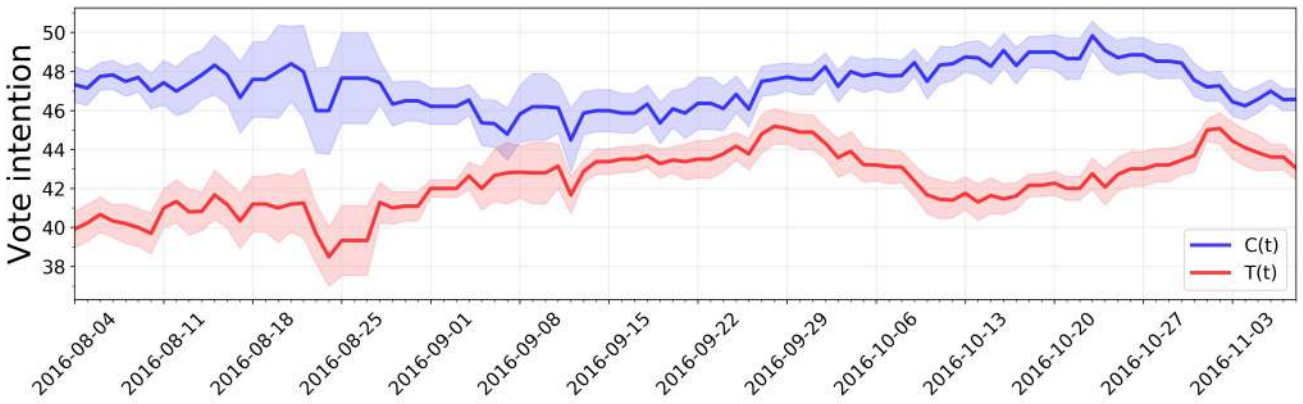
(R=0.72; p-value = 0.004)

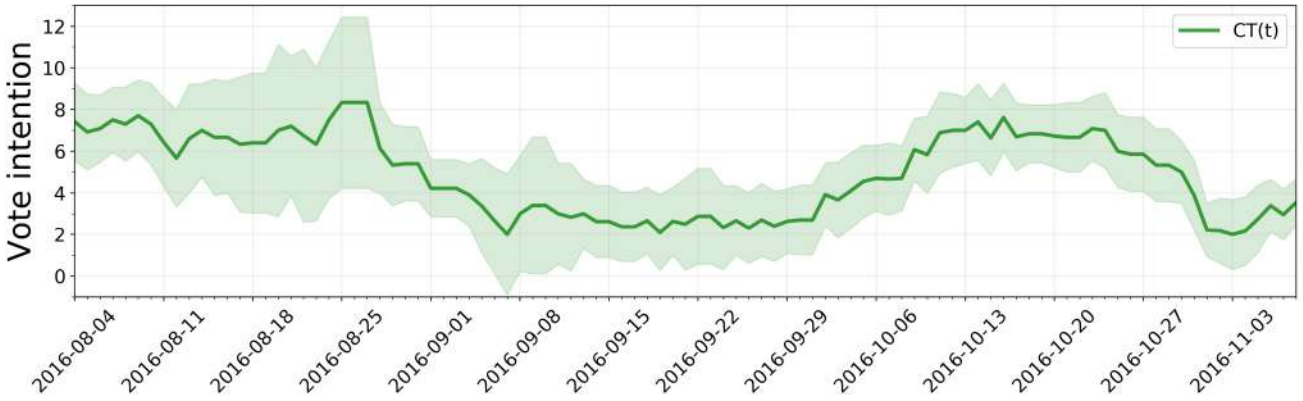
[1] Yasseri, T., & Bright, J. (2016). Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Science*, *5*(1), 1-15. ISO 690

Our proposal

- To understand what features of the news media articles can influence public opinion in an electoral context.
- We focus on 2016 USA presidential elections and analyze:
  - News articles mentioning candidates Hillary Clinton and Donald Trump
  - Electoral polls on voting intentions for each candidate.
- We analyze different features of news media articles :
  - Amount of mentions to each candidate.
  - Sentiment bias toward each candidate.
  - Topics of discussion.
- We studied correlation and causality measures between these aspects and the time series of the surveys.

Surveys as proxy of public opinion





$$CT(t) = C(t) - T(t)$$

RealClear Politics

- 263 USA national surveys
- Different pollsters: NBC, New York Times, LA Times, CBS, Fox News, Gravis, ABC, IBD (among others)

Final difference:

+3.2 Clinton according to average surveys

+2.1 Clinton according to final elections results

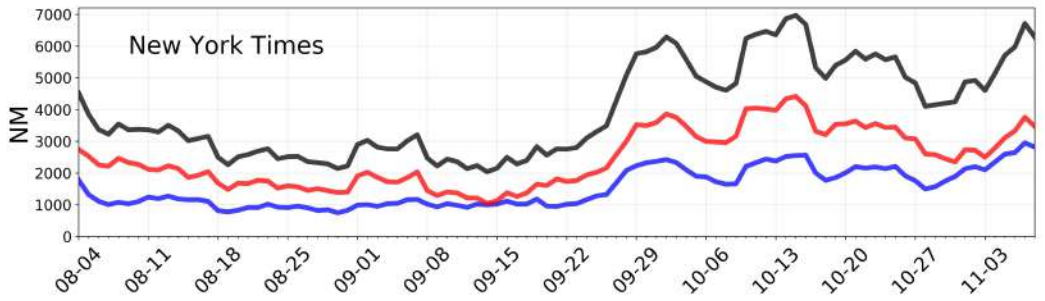
Corpus of news articles

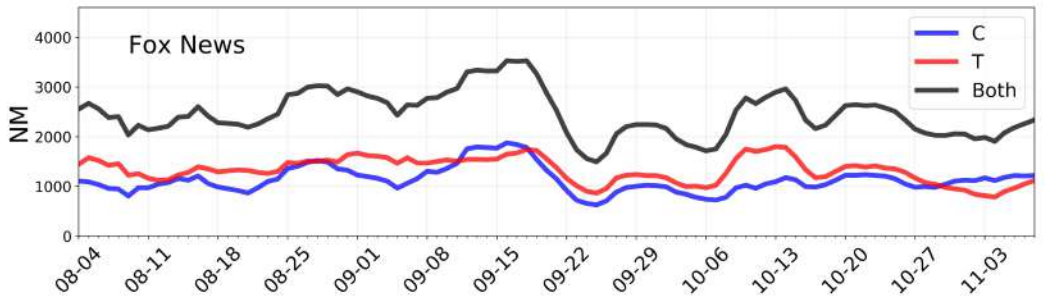
We select online articles containing at least the name of one of the two main candidates: Hillary Clinton (Democrat) and Donald Trump (Republican).

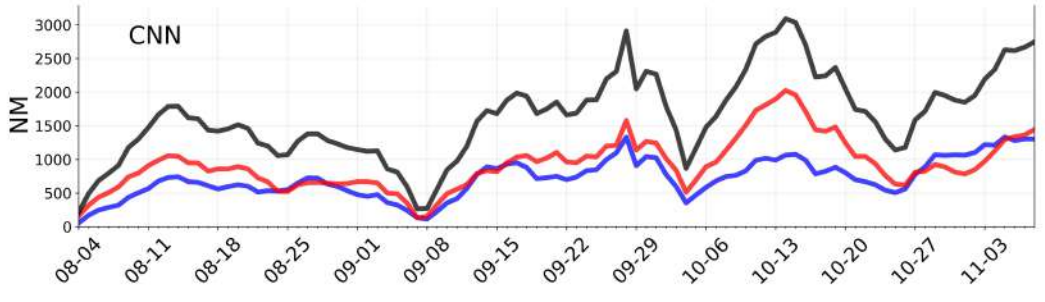
- **The New York Times** (5672 news articles)
- **Fox News** (5750)
- **USA Today** (833)
- **CNN** (2920)

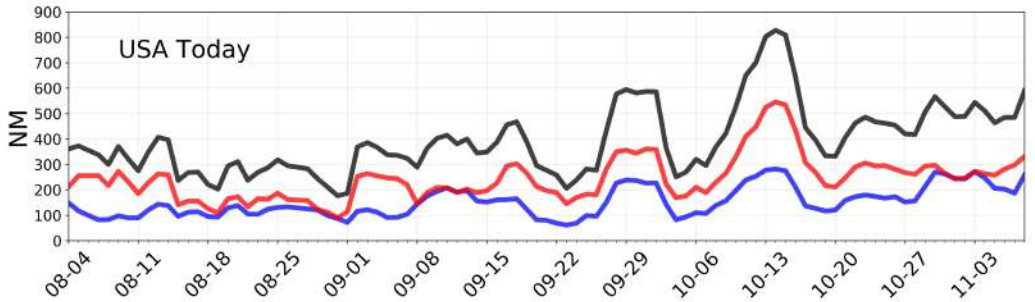
<u>At least one of the candidates is mentioned in the notes.</u>

1 – Mentions to candidates









Spearman Correlation / Lag

0.663 / 7-15 days, p<0.001

-0.476 / 12-15 days, p<0.05

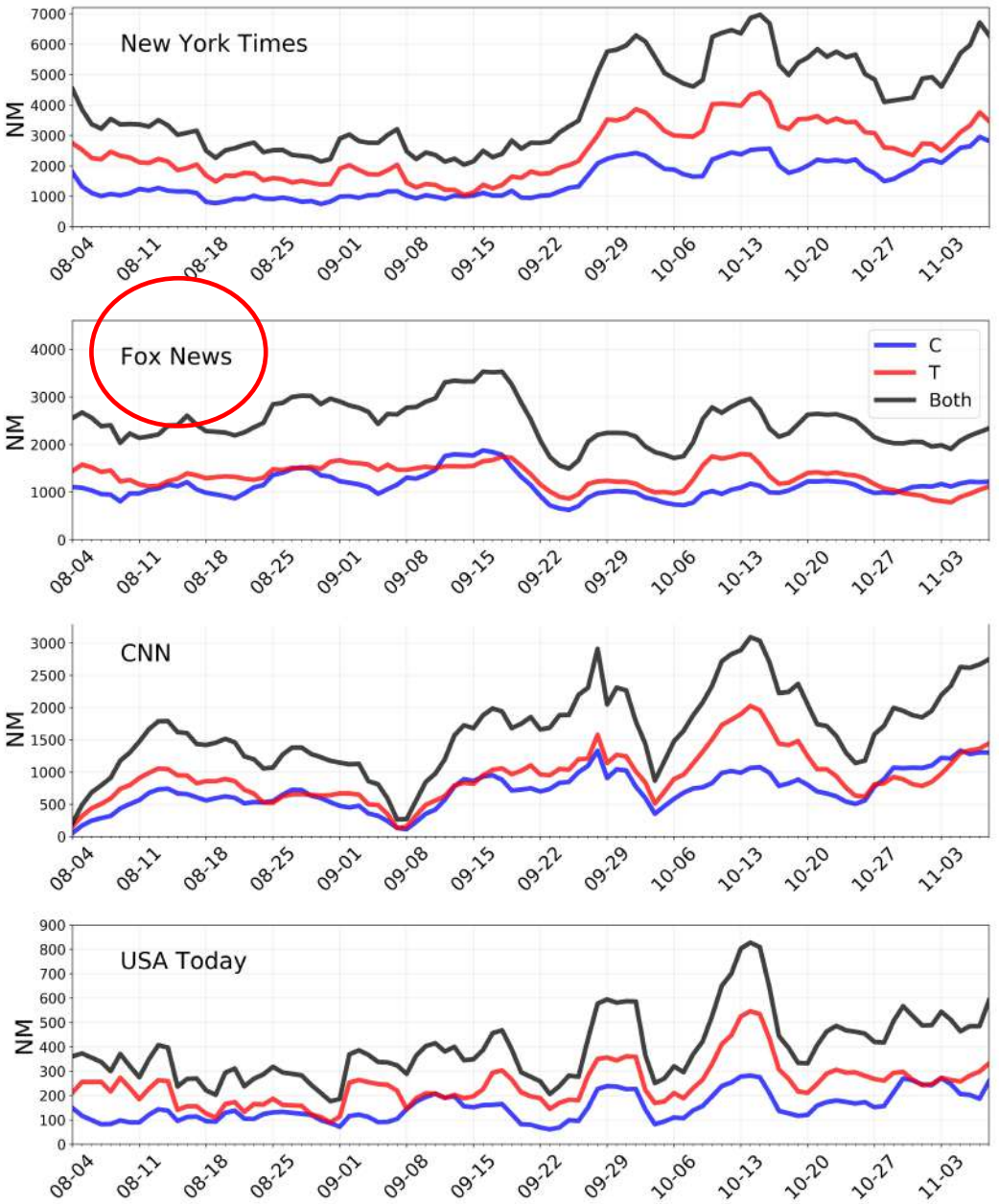
0.426 / 12-15 days, p<0.05

0.246 / 12-15 days, p<0.05

1 – Mentions to candidates

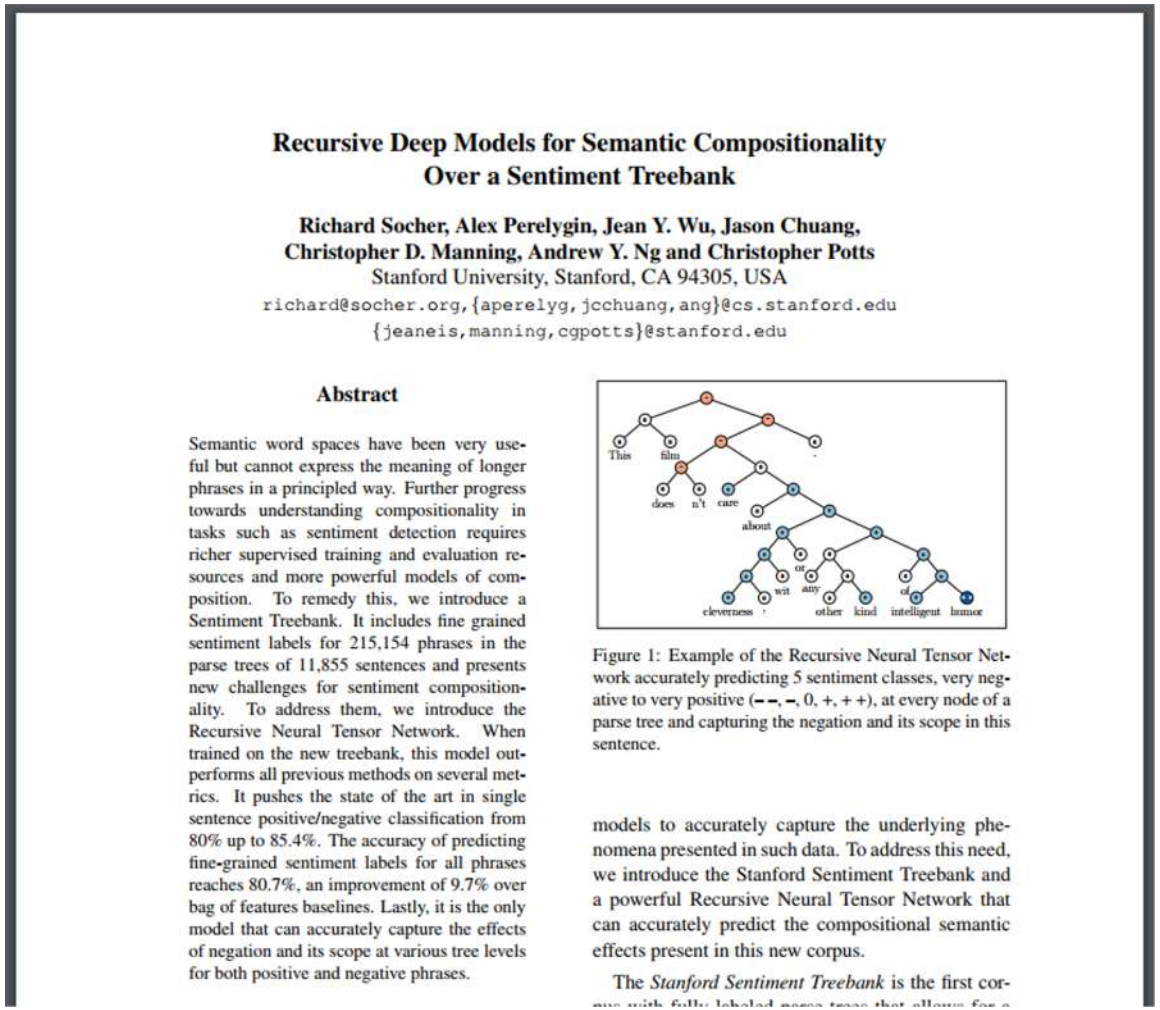
Mentions in Fox News: ⬆

Difference Clinton – Trump: ⬇



Spearman Correlation / Lag
0.663 / 7-15 dias, p<0.001
-0.476 / 12-15 dias, p<0.05
0.426 / 12-15 dias, p<0.05
0.246 / 12-15 dias, p<0.05

2 – Sentiment Analysis



Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org, {aperelyg, jcchuang, ang}@cs.stanford.edu
{jeaneis, manning, cgpotts}@stanford.edu

Abstract

Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network. When trained on the new treebank, this model outperforms all previous methods on several metrics. It pushes the state of the art in single sentence positive/negative classification from 80% up to 85.4%. The accuracy of predicting fine-grained sentiment labels for all phrases reaches 80.7%, an improvement of 9.7% over bag of features baselines. Lastly, it is the only model that can accurately capture the effects of negation and its scope at various tree levels for both positive and negative phrases.

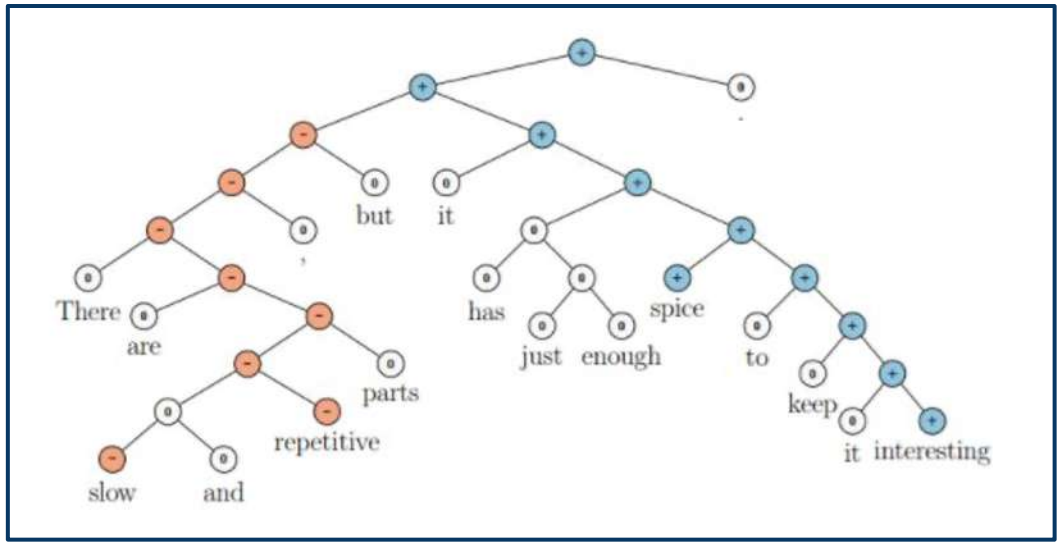
Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

models to accurately capture the underlying phenomena presented in such data. To address this need, we introduce the Stanford Sentiment Treebank and a powerful Recursive Neural Tensor Network that can accurately predict the compositional semantic effects present in this new corpus.

The *Stanford Sentiment Treebank* is the first cor-

Standford Core NLP: They build a binary tree from each phrase to determine whether the phrase has a positive, negative or neutral connotation from :

- A database with positive, negative or neutral words.
- The sentence structure.
- The presence of intensifiers, appeasers or deniers.



Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631-1642).

2 - Sentiment Analysis

Steps:

1) Phrases where Donald Trump and Hillary Clinton are mentioned are detected.

2) Sentiment analysis is run on the sentences (in the case where both are mentioned, the sentence is split and then the SA algorithm is applied).

3) The number of positive, neutral, negative and total mentions of each candidate are counted.

A Look at Clinton's Marriage Woes Raises a Question: Really?

The Public Editor
By LIZ SPAYD OCT. 4, 2016

494

Hillary Clinton talking to the media on her campaign plane. Doug Mills/The New York Times

Late last Saturday evening, The New York Times delivered an eye-popping scoop to its readers — tax documents of Donald Trump's that showed he

New: https://www.nytimes.com/2016/10/05/public-editor/hillary-clinton-bill-clinton-marriage.html

2 - Sentiment Analysis

$\#C_+$ = Fraction of positive mentions to Hillary Clinton
$\#C_-$ = Fraction of negative mentions to Hillary Clinton
$\#T_+$ = Fraction of positive mentions to Donald Trump
$\#T_-$ = Fraction of negative mentions to Donald Trump

Sentiment Bias (SB_i): measure de bias towards the candidates in the "i" media outlet. (if SB>0, the bias is towards Clinton compared with Trump and viceversa)

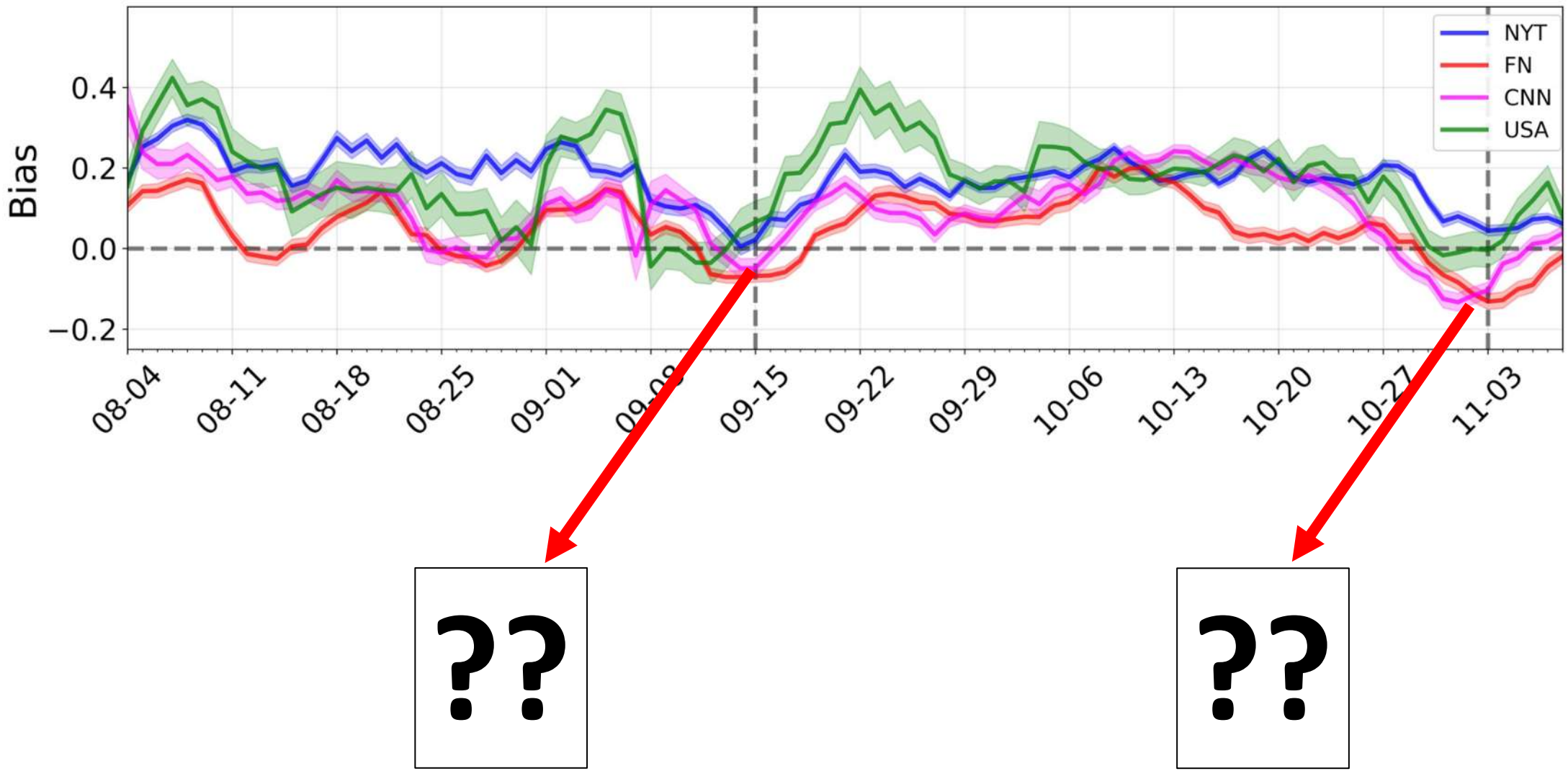
$$SB_i = (\#C_+ - \#C_-) - (\#T_+ - \#T_-)$$

Medio	The New York Times	USA TODAY	CNN	FOX NEWS channel
SB	0.162 ± 0.004	0.160 ± 0.010	0.094 ± 0.006	0.046 ± 0.005

SB_i = es nonnegative with $p < 0.001$ (boostrapping) for all media oulets

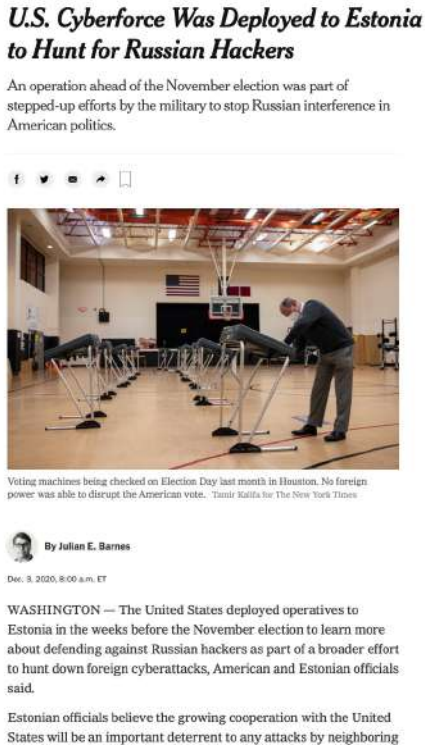
Obs: While SB is positive for all, it is higher for NYT and very close to zero for Fox News. (We will see that the time dependence of this statistic is more informative).

2 - Sentiment Analysis



3 – Topic decomposition:

We describe each news articles as a frequency vector in the semantic space (tf-idf)



U.S. Cyberforce Was Deployed to Estonia to Hunt for Russian Hackers

An operation ahead of the November election was part of stepped-up efforts by the military to stop Russian interference in American politics.

Voting machines being checked on Election Day last month in Houston. No foreign power was able to disrupt the American vote. Tamir Kalifa for The New York Times

By Julian E. Barnes

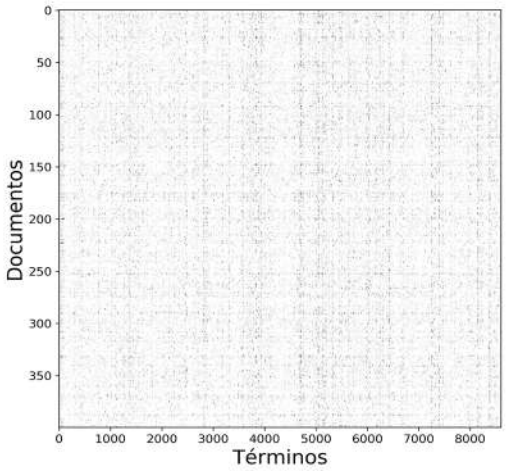
Dec. 3, 2020, 8:00 a.m. ET

WASHINGTON — The United States deployed operatives to Estonia in the weeks before the November election to learn more about defending against Russian hackers as part of a broader effort to hunt down foreign cyberattacks, American and Estonian officials said.

Estonian officials believe the growing cooperation with the United States will be an important deterrent to any attacks by neighboring

$$V=[1,\ldots,\mathbf{tf(t)*idf}(t),\ldots,F]$$

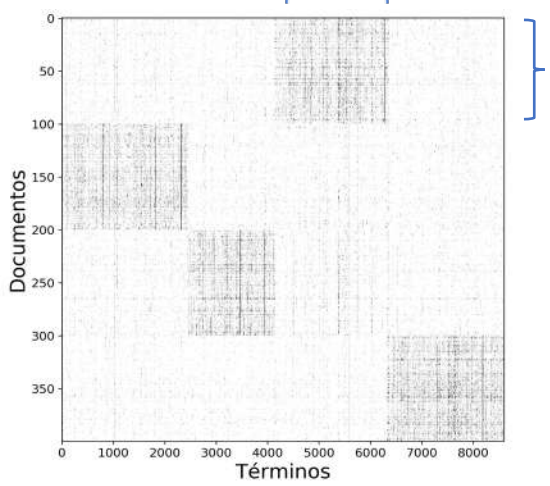
News corpus (tf-idf)



NMF

Topic's keywords

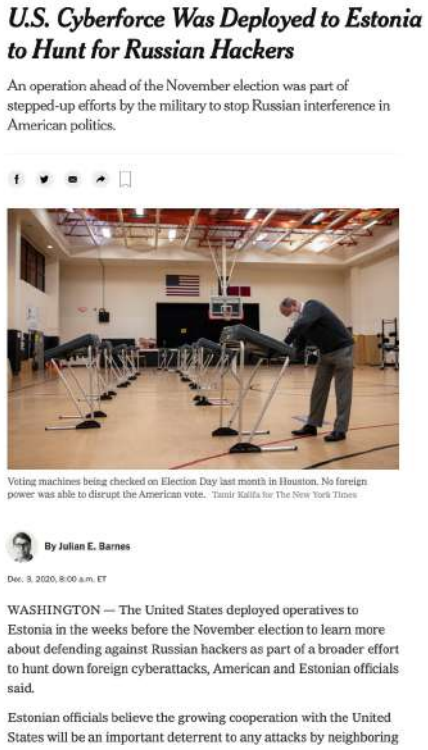
Similar articles



Topic decomposition: We look identify the group of news speaking of the same topic and described by a limited number of keywords (unsupervised clustering).

3 - Topic decomposition

We describe each news articles as a frequency vector in the semantic space (tf-idf)



U.S. Cyberforce Was Deployed to Estonia to Hunt for Russian Hackers

An operation ahead of the November election was part of stepped-up efforts by the military to stop Russian interference in American politics.

Voting machines being checked on Election Day last month in Houston. No foreign power was able to disrupt the American vote. Tamir Kalifa for The New York Times

By Julian E. Barnes

Dec. 3, 2020, 8:00 a.m. ET

WASHINGTON — The United States deployed operatives to Estonia in the weeks before the November election to learn more about defending against Russian hackers as part of a broader effort to hunt down foreign cyberattacks, American and Estonian officials said.

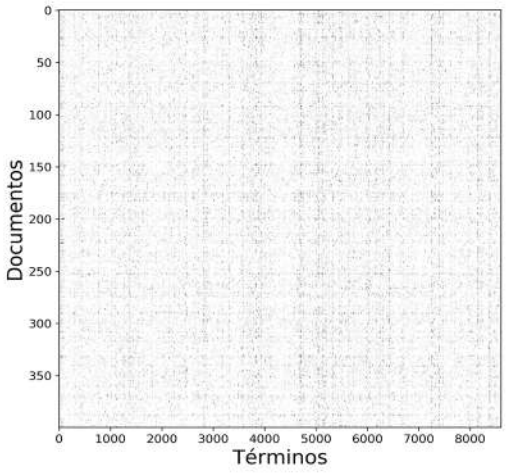
Estonian officials believe the growing cooperation with the United States will be an important deterrent to any attacks by neighboring

$$V=[1,\ldots,\mathbf{tf(t)}*\mathbf{idf}(t),\ldots,F]$$

# of time term "t" appears in the article

Specificity factor of term "t"

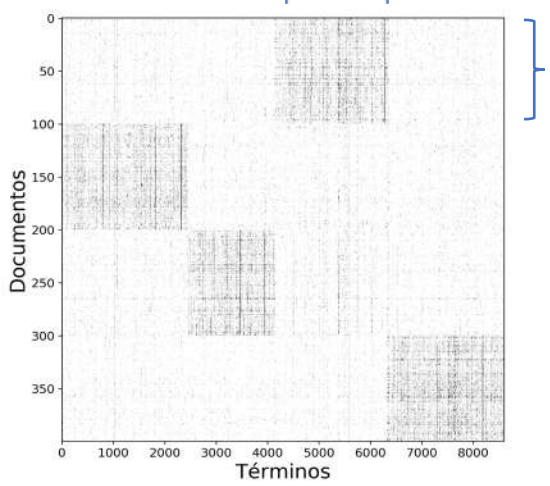
News corpus (tf-idf)



NMF

Topic's keywords

Similar articles



Topic decomposition: We look identify the group of news speaking of the same topic and described by a limited number of keywords (unsupervised clustering).

Emergent Topics

Clinton email controversy

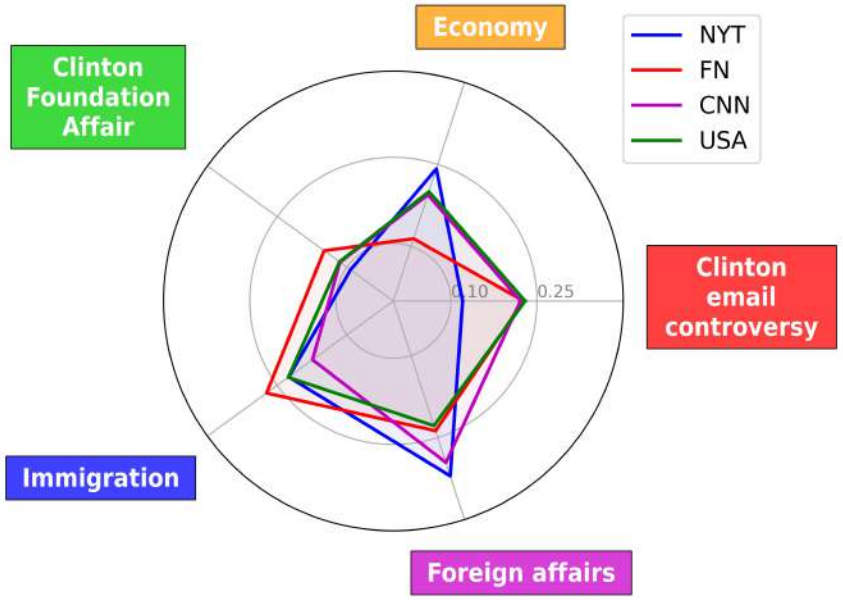
Economy

Foreign affairs

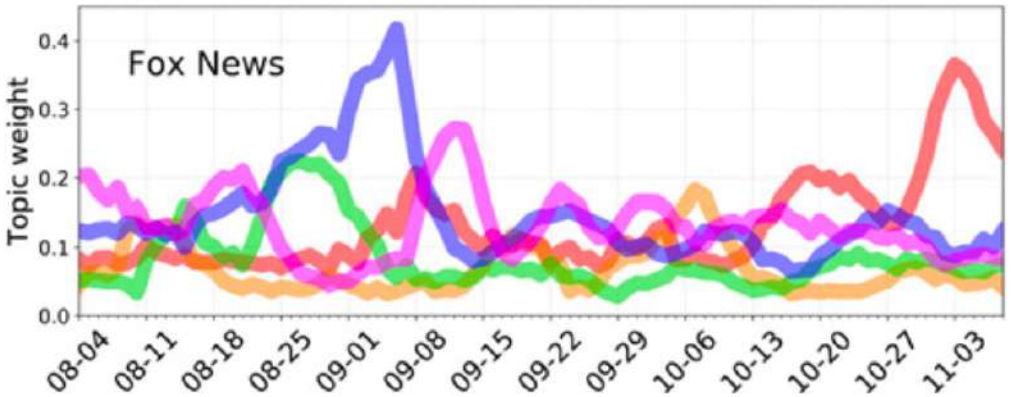
Immigration

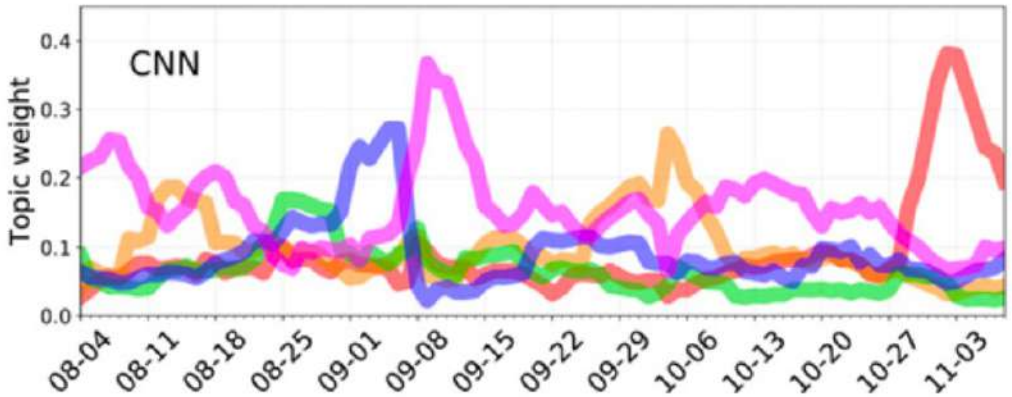
Clinton Foundation Affair

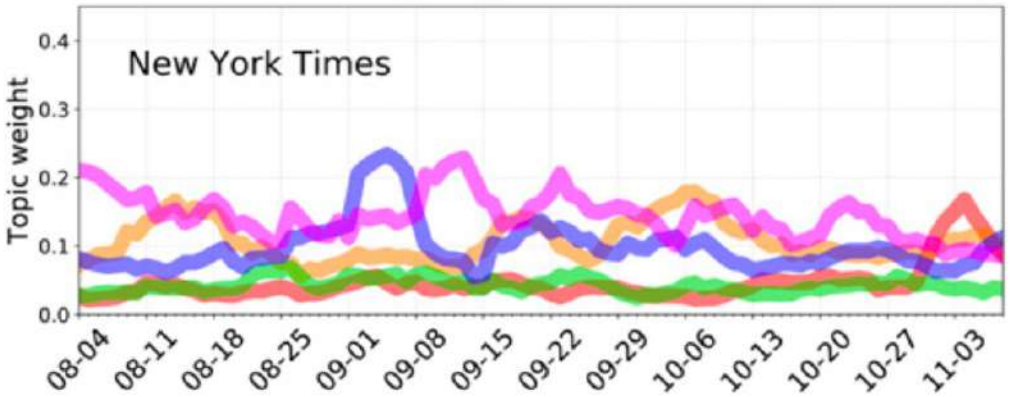
Agendas

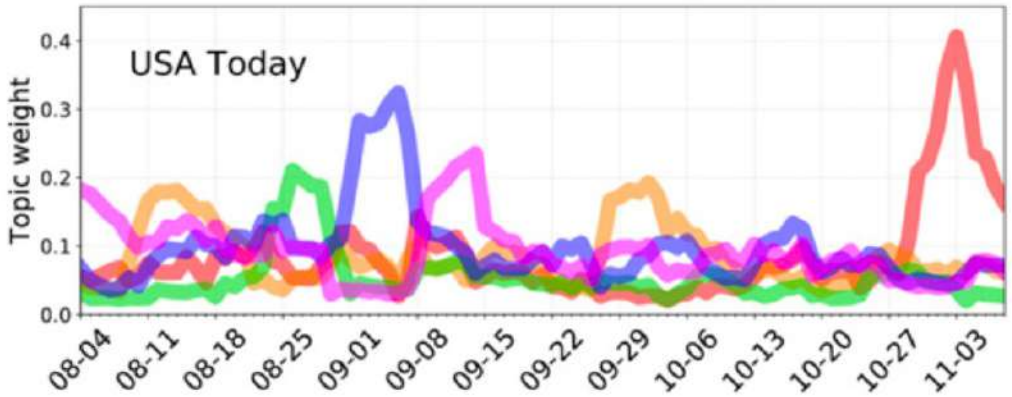


3 – Topic decomposition and media agenda

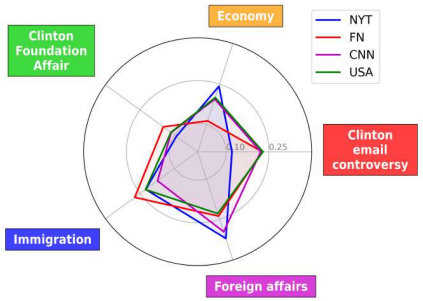




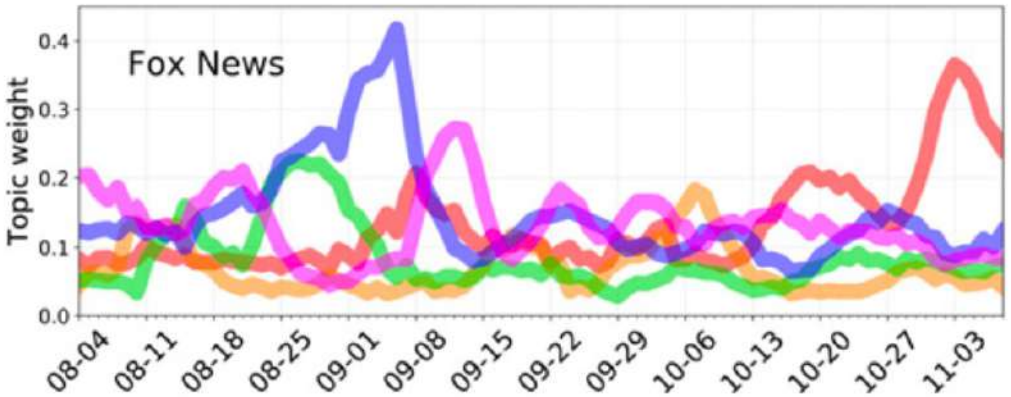


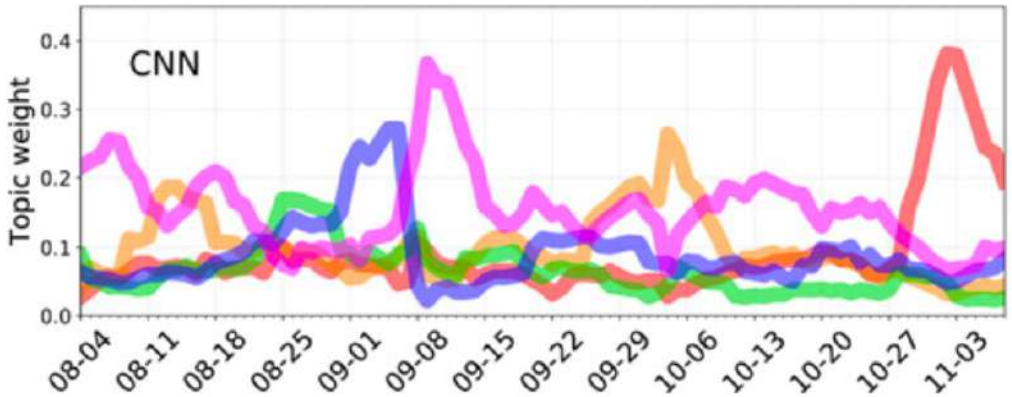


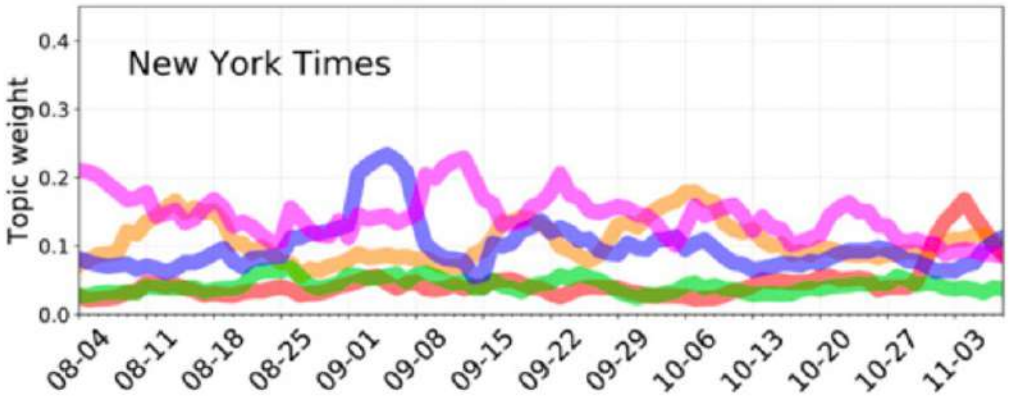
Topic	NYT (SRL)	Fox (SRL)	CNN (SRL)	USA (SRL)
Clinton email controversy	−0.46 (10–20)	−0.42 (11–20)	−0.45 (13–20)	−0.54 (10–20)
Economy	0.56 (4–20)	0.59 (8–20)	0.48 (5–18)	0.40 (10–15)
Clinton foundation affair	−0.53 (3–20)	−0.43 (15–20)	−0.53 (1–20)	−0.40 (5–12)
Immigration	−0.42 (0–12)	−0.44 (5–20)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

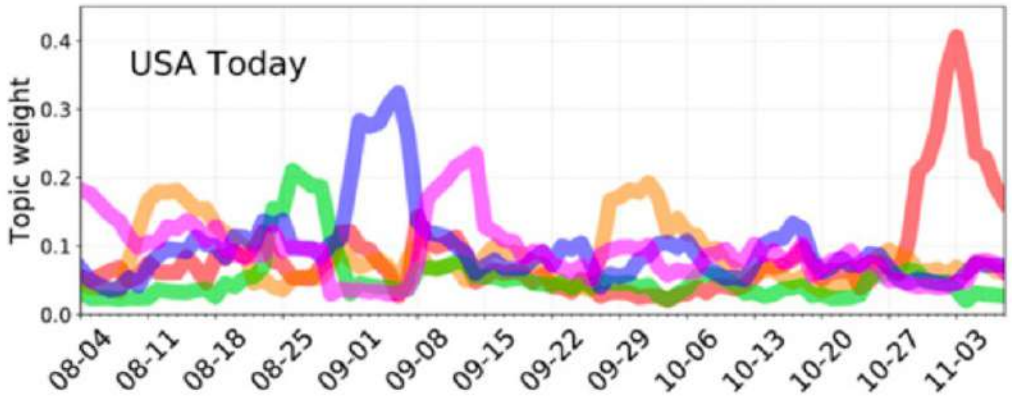


3 – Topic decomposition and media agenda

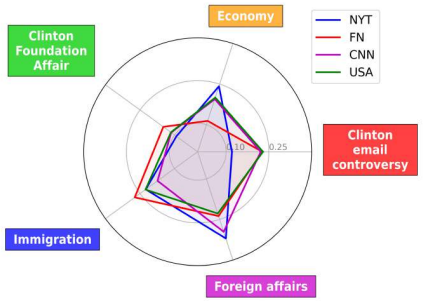








Topic	NYT (SRL)	Fox (SRL)	CNN (SRL)	USA (SRL)
Clinton email controversy	−0.46 (10–20)	−0.42 (11–20)	−0.45 (13–20)	−0.54 (10–20)
Economy	0.56 (4–20)	0.59 (8–20)	0.48 (5–18)	0.40 (10–15)
Clinton foundation affair	−0.53 (3–20)	−0.43 (15–20)	−0.53 (1–20)	−0.40 (5–12)
Immigration	−0.42 (0–12)	−0.44 (5–20)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



4 – Combined topic decomposition and sentiment analysis

Sentiment Bias (SB): relative number of positive and negative mentions of Clinton and Trump.

$$SB = (\#C_+ - \#C_-) - (\#T_+ - \#T_-)$$

4 – Combined topic decomposition and sentiment analysis

Sentiment Bias (SB): relative number of positive and negative mentions of Clinton and Trump (by topic and media outlet).

$$SB = (\#C_+ - \#C_-) - (\#T_+ - \#T_-)$$

Tópico	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 affair	-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

Bias of each topic (SB < 0 unfavorable to Clinton).

4 – Combined topic decomposition and sentiment analysis

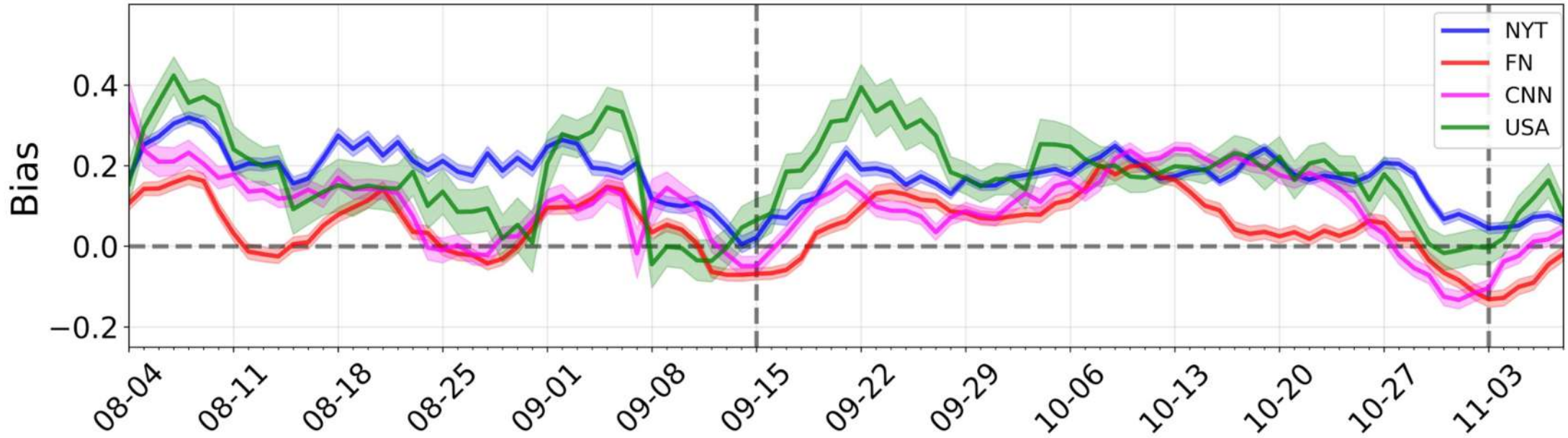
Sentiment Bias (SB): relative number of positive and negative mentions of Clinton and Trump.

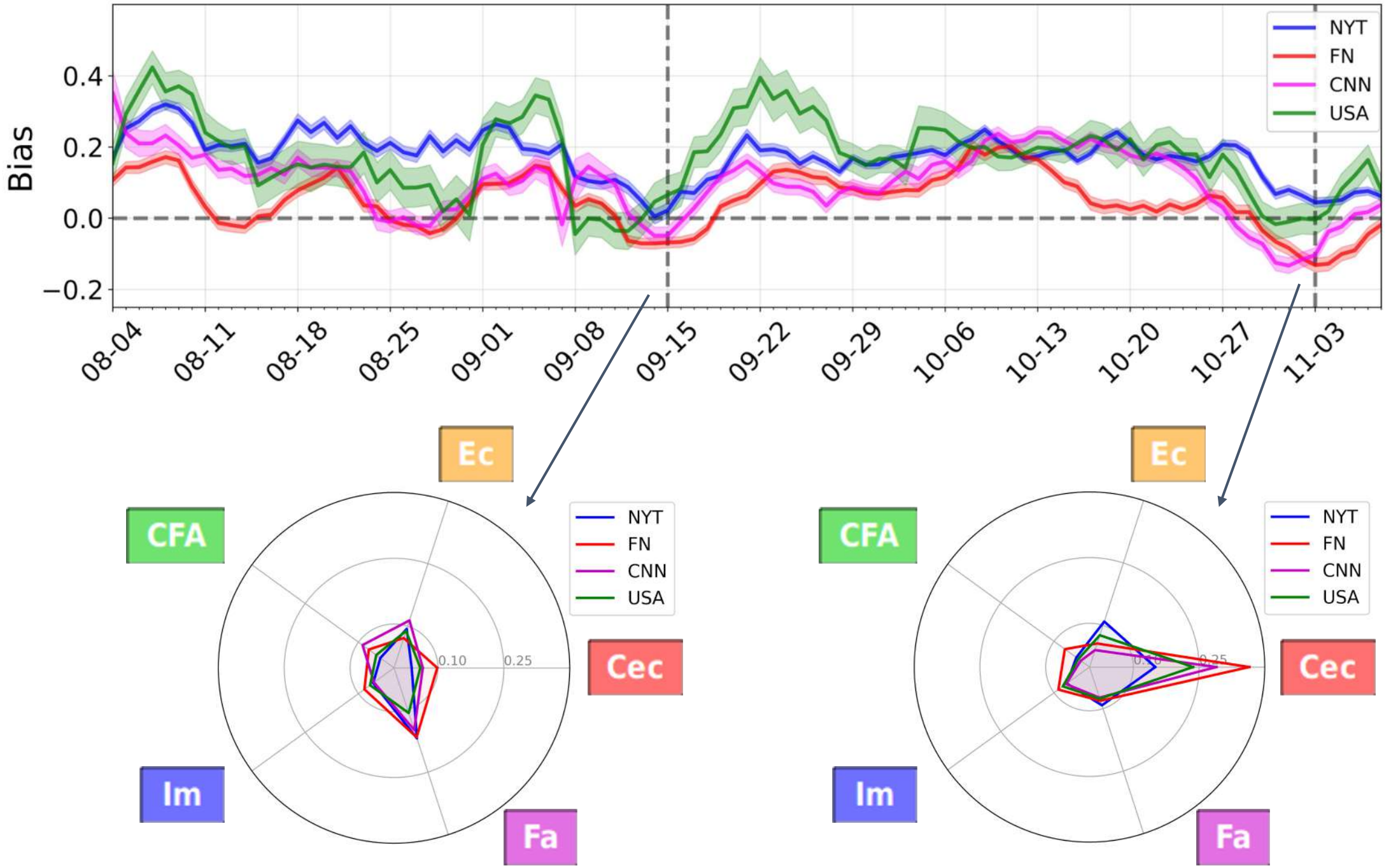
$$SB = (\#C_+ - \#C_-) - (\#T_+ - \#T_-)$$

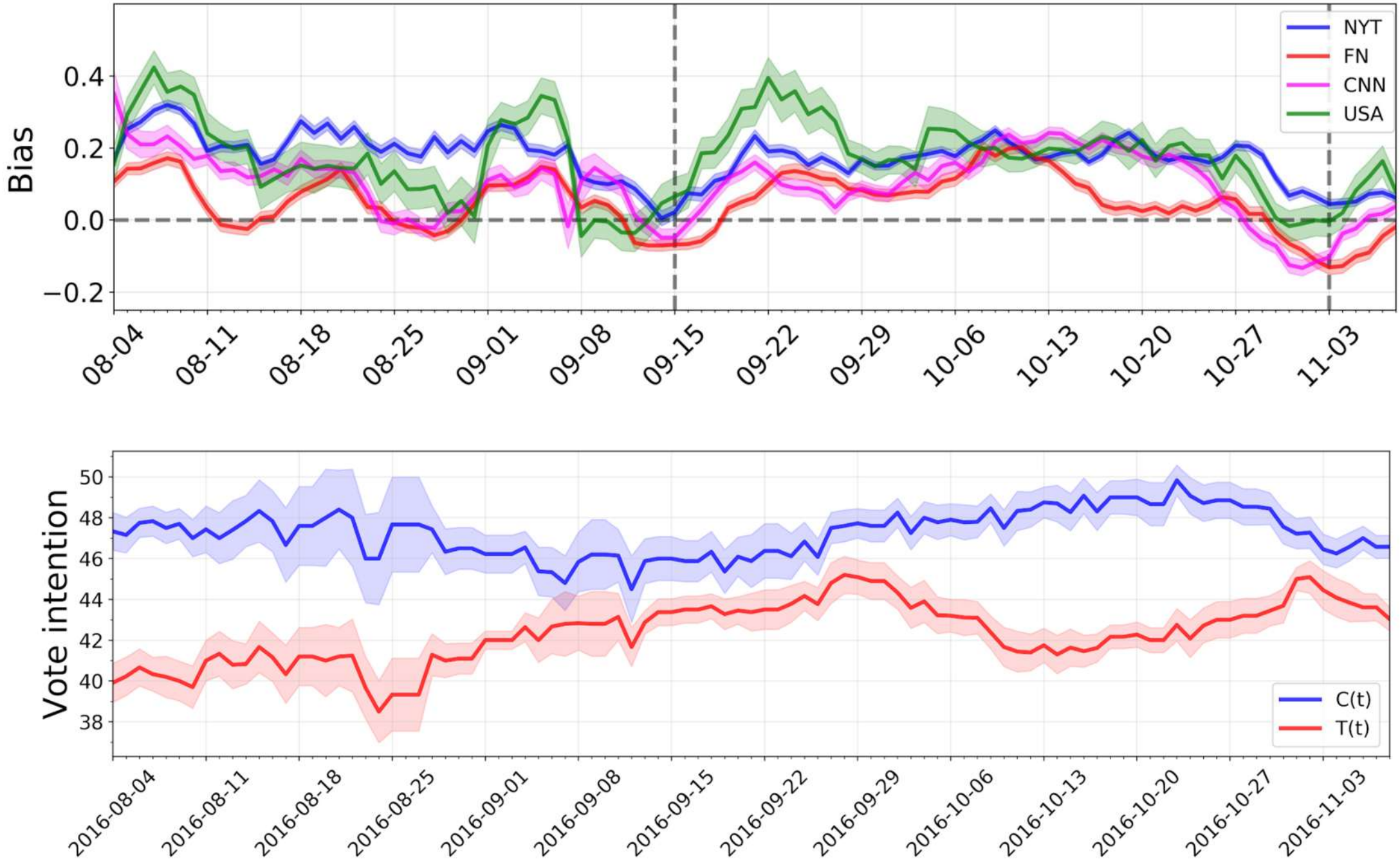
Tópico	SB_{NYT}	SB_{FN}	SB_{CNN}	SB_{USA}
Clinton email controversy	-0.475	-0.429	-0.302	-0.315
Economy	0.332	0.070	0.152	0.168
Clinton foundation affair	-0.256	-0.257	-0.304	-
Immigration	0.501	0.347	0.306	0.382
Foreign affairs	0.146	0.053	0.115	0.166

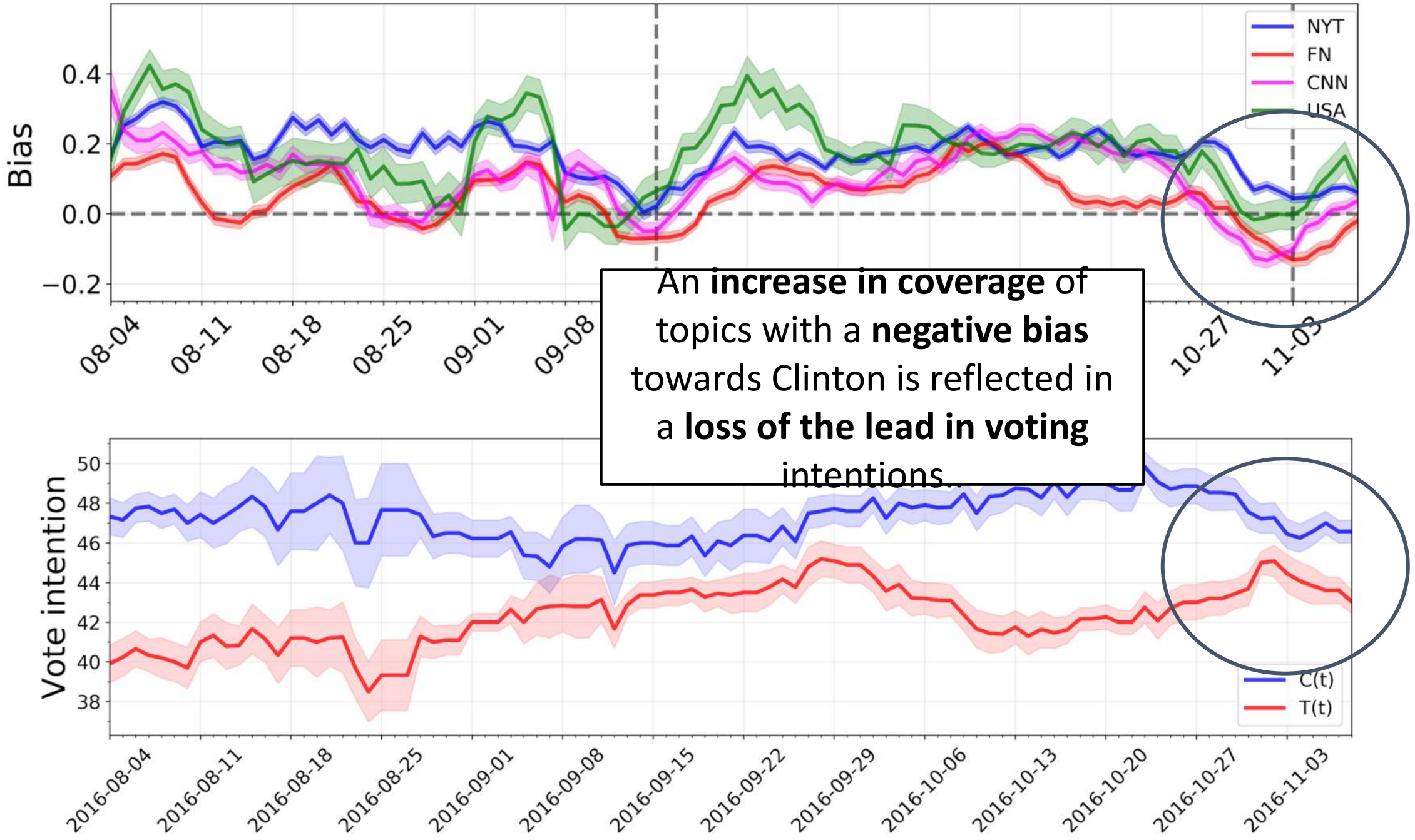
SB < 0

Bias of each topic (SB < 0 unfavorable to Clinton).


NYT
FN
CNN
USA
Bias
0.4
0.2
0.0
−0.2
08-04
08-11
08-18
08-25
09-01
09-08
09-15
09-22
09-29
10-06
10-13
10-20
10-27
11-03


Bias
0.4
0.2
0.0
−0.2
08-04
08-11
08-18
08-25
09-01
09-08
09-15
09-22
09-29
10-06
10-13
10-20
10-27
11-03
NYT
FN
CNN
USA
Ec
CFA
Cec
Im
Fa
0.10
0.25
NYT
FN
CNN
USA
Ec
CFA
Cec
Im
Fa
0.10
0.25


Bias
0.4
0.2
0.0
−0.2
NYT
FN
CNN
USA
08-04
08-11
08-18
08-25
09-01
09-08
09-15
09-22
09-29
10-06
10-13
10-20
10-27
11-03
Vote intention
50
48
46
44
42
40
38
C(t)
T(t)
2016-08-04
2016-08-11
2016-08-18
2016-08-25
2016-09-01
2016-09-08
2016-09-15
2016-09-22
2016-09-29
2016-10-06
2016-10-13
2016-10-20
2016-10-27
2016-11-03



An **increase in coverage** of topics with a **negative bias** towards Clinton is reflected in a **loss of the lead in voting** intentions..

5 –Granger causality

It reveals if the dynamics of a temporal series have a causal relationship with another time series.

Basically, it consist in determine if additional information provided by a second lagged time series, y(t − τ), improves the forecast of x(t).

$$\Delta CT(t) = CT(t) - CT(t-1) = w_t \qquad (1)$$

$$\Delta CT(t+\tau) = \beta \cdot \Delta T_i(t) + w_{t+\tau} \qquad (2)$$

We say that the **coverage** of a given issue effectively **affects** the **difference** between voting intention between candidates when the parameter **β i**n la equation (2) significatively **differs from zero.**

Tópico	NYT	FN	CNN	USA Today
Clinton email controversy	-	$\beta < 0$ (τ=19)		
Clinton Foundation affaire	$\beta < 0$ (t=19)			
Inmigration	$\beta < 0$ (t=10)			
Economy		$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)

5 –Granger causality

It reveals if the dynamics of a temporal series have a causal relationship with another time series.

Basically, it consist in determine if additional information provided by a second lagged time series, y(t − τ), improves the forecast of x(t).

$$\Delta CT(t) = CT(t) - CT(t-1) = w_t \qquad (1)$$

$$\Delta CT(t+\tau) = \beta \cdot \Delta T_i(t) + w_{t+\tau} \qquad (2)$$

We say that the **coverage** of a given issue effectively **affects** the **difference** between voting intention between candidates when the parameter **β i**n la equation (2) significatively **differs from zero.**

Tópico	NYT	FN	CNN	USA Today
Clinton email controversy	-	$\beta < 0$ (τ=19)		
Clinton Foundation affaire	$\beta < 0$ (t=19)			
Inmigration	$\beta < 0$ (t=10)			
Economy		$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)

Favorable to Donald Trump

5 –Granger causality

It reveals if the dynamics of a temporal series have a causal relationship with another time series.

Basically, it consist in determine if additional information provided by a second lagged time series, y(t − τ), improves the forecast of x(t).

$$\Delta CT(t) = CT(t) - CT(t-1) = w_t \qquad (1)$$

$$\Delta CT(t+\tau) = \beta \cdot \Delta T_i(t) + w_{t+\tau} \qquad (2)$$

We say that the **coverage** of a given issue effectively **affects** the **difference** between voting intention between candidates when the parameter **β i**n la equation (2) significatively **differs from zero.**

Tópico	NYT	FN	CNN	USA Today
Clinton email controversy	-	$\beta < 0$ (τ=19)		
Clinton Foundation affaire	$\beta < 0$ (t=19)			
Inmigration	$\beta < 0$ (t=10)			
Economy		$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)	$\beta > 0$ (τ=11-16)

Favorable to Hillary Clinton

Conclusions

- A combined approach of **topic decomposition** and **sentiment analysis** to **news stories** mentioning candidates in an election allowed us to see which relevant aspects influence **people's voting intentions**.
- The analysis of **correlations by topic** with the difference in **voting intentions** observed in the polls allowed us to understand the **fundamental role** played by the topics that **favor or disfavor** the **candidates**.
- **Sentiment analysis discriminated by topic** allowed us to understand how a given topic (Clinton email controversy) played a **key role** in **Trump's vote increase** in the days leading up to the election.
- The **Granger test** is consistent with previous results and reinforces the interpretation of a **causal relationship** between the **topic addressed** by candidates in the media and voting behavior.

SoPhy Lab: The Social Physics Group

Federico Albanese

Sebastian Pinto

IOP Publishing J.Phys.Complex. 1 (2020) 025005 (12pp) https://doi.org/10.1088/2632-072X/ab87

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, C1120AAQ, 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

Thanks!!