Jordan G. Rohrlich

University of Virginia

Corcoran Department of Philosophy

PHIL 4500: Data & Culture

April 25th, 2017

News Source Social Media and Content Analysis

ABSTRACT

This project hopes to dive into two primary questions: Does the social media content of news sources imply a bias toward maximizing user engagement? And, do news sources show bias in the coverage of particular topics? Toward the first, this study analyzes dispersion plots and sentiment of source tweets and articles, finding meaningful difference in the frequency of keyword references between source social media and site content, but no relationship between tweet sentiment and follower engagement. Toward the second, this study uses source clustering by tweets and article texts, as well as topic modelling, to show that sources can be grouped noticeably by frequencies of word use and do show measurable preferences in regards to topic coverage.

KEYWORDS – News, Media Bias, Social Media, Twitter, Politics, Topic Modelling

PROJECT OVERVIEW

This project seeks to better understand media bias in the digital age. It examines the relationship between news source social media and site content, as well as the interrelation of sources by preferred topics of coverage. The first thesis of this study supposes that news source social media content differs significantly from actual article content, an idea suggesting that news organizations attempt to attract more viewers by modifying the appearance of their content on social media platforms. The second thesis supposes that news sources show noticeable preference for particular topics, also suggesting a content bias catering to audience interests as well as political motivations. Altogether, this study should help understand the degree of media bias in today's digital information spaces.

By pulling 2000 tweets from 15 news source Twitter timelines, then extracting 3000 URLs to scrape article text, tweet and article corpora were built to examine these content and source relationships. First, mentions of keywords over time for tweets and articles were examined for each source using dispersion plotting. Next, sentiment of individual tweets and the collection of tweets for each source were correlated with two engagement metrics: retweets and followers, respectively. Third, news source word frequencies were calculated to generate a dendrogram that clusters news sources together by frequency of word use. Finally, the combined article corpus was used to create a topic model, and sources' word frequencies of topic top words were used to plot a heat map indicating degree of news source bias toward particular topics.

DATA COLLECTION AND PREPARATION

Data curation relied on an R package, "rtweet", to fetch 2000 tweets for each of 15 news sources[1]. Fetching tweets from Twitter timelines, the resulting tweet data frame was trimmed of retweets, then truncated to only include tweets that fit in the same time period for each source –

[1] Sources included: Occupy Democrats, Huffington Post, The Atlantic, Vox, The New York Times, Associated Press, CNN, Politico, The Wall Street Journal, The Economist, Fox News, RealClear Politics, Breitbart News, BuzzFeed, and The Onion.

April 8th to 21st, 2017. Then, article links were extracted from the tweet database using a regular expression, then organized into an article link list. Because this link list object had differing numbers of links for each article (with a lower bound of 200 and an upper bound of 1600), a trimming function combed through the object, selecting articles at a constant interval decided by the excess ratio of that source's link number divided by 200. The resulting object, with approximately 200 links per source and no more than 250, was iterated through, having an article scraping algorithm pull article text from news site pages according to each source's respective web markup. The resulting objects, from which all subsequent analysis began, included a data frame with source tweets and related data, as well as a corpus of article texts for each of the 200+ articles for each source.

PROCESSING AND ANALYSIS

The first part of this analysis consisted of dispersion plotting. In order to compare the dispersions of two keywords, "Trump" and "Russia", within the narrative time of each source's tweet and article corpus, the lines of the corpora were expanded into a vector of words. From this, a corresponding vector of NAs was created, of equal length to the vector of words, wherein 1 replaced NA if the word matched the keyword, for each word in the corpus vector of words. When the resulting vector was plotted, a barcode graph identified the distribution of the keywords over the narrative time of each news source, from April 8th to 21st, for both tweet and article corpora. The dispersion plots were saved externally as png image files.

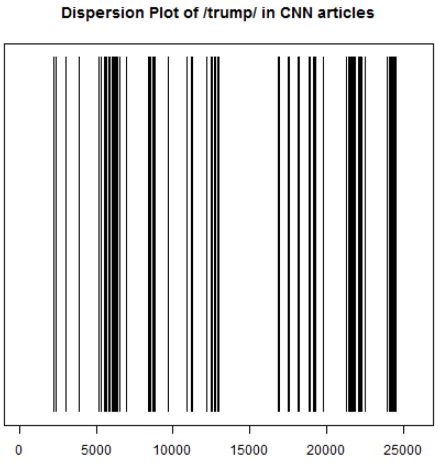
The second part of the study involved sentiment analysis of news source tweets. For each source, the R package "Syuzhet" was used to score each tweet based on positive- or negative-sentiment words it contained. A correlation coefficient was then produced to relate the sentiment score of each tweet with the corresponding number of retweets, a metric fetched using rtweet in the data collection process. For the same source, the total sentiment of all tweets combined was scored, and the total number of followers of that source was added. The resulting data frame displayed the news source, the total tweet corpus sentiment of that source, the individual tweet sentiment-to-retweet correlation coefficient, and the number of followers of that source's twitter page. The mean correlation coefficient of the sentiment-to-retweet values was -0.046; the correlation between total source sentiment and source follower count was -0.25, with an insignificant p-value of 0.37.

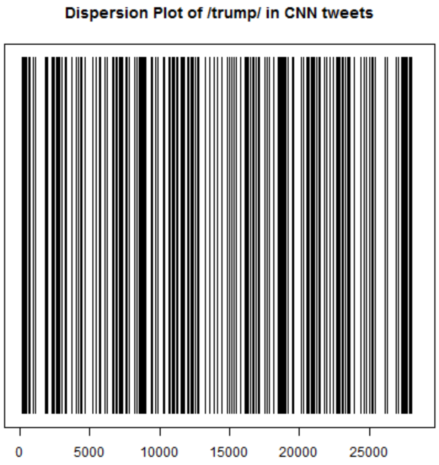
Moving into analyses of news source interrelation, the third section of this analysis attempted to cluster news sources together by word frequency, across both tweet and article corpora, as well as both full and keyword-only word frequencies. The tweet and article corpora were broken apart into word vectors, which were then tabulated to calculate relative word frequencies. For the cluster analysis that relied exclusively on keywords, this processing stage also included a filter to remove stopwords in the corpus word vectors. After converting this frequency table into an xtabs object, then reshaping it into a wide matrix of corpus word frequencies by source and word, a threshold frequency of 0.05 was applied to discard words with low average relative frequencies across the 15 news sources. Next, the Euclidean distances were calculated for the 15 sources, then fed into a cluster plot. The 4 plots, corresponding to tweet versus article corpora, for both full and keyword-only frequencies, were saved externally as png image files.

Finally, a topic model was created for the article data corpus. Using the R package "Mallet", an LDA topic model was created, taking a chunk size of 400 words for the article corpus, a topic number of 20, an optimization interval of 40, an iteration burn-in of 80, and a set number of 400 iterations. This produced a set of top words and corresponding frequencies for each of 20 topics, which were renamed subjectively by approximate descriptive fit. Next, for each topic, the relative frequencies of topic top words were found within the word frequency table of each news source and, for each source, a Euclidean distance was calculated to estimate the degree of proximity that each source's word frequencies have to the topic's top words. Organizing these scores into a data frame, a heat map was produced using R package "ggplot2" to help visualize the topic affinities of each news source.

INTERPRETATION OF RESULTS

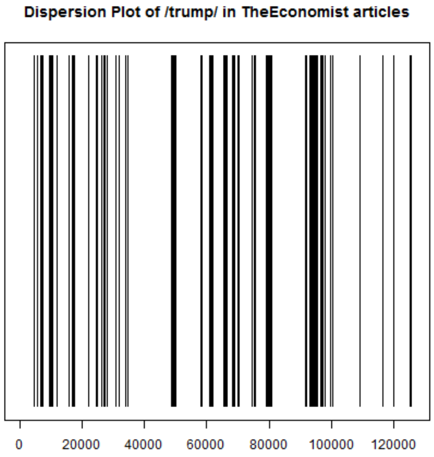
First, the dispersion plots showed remarkable difference, when comparing keyword use across tweet and article timelines. Take CNN for example:

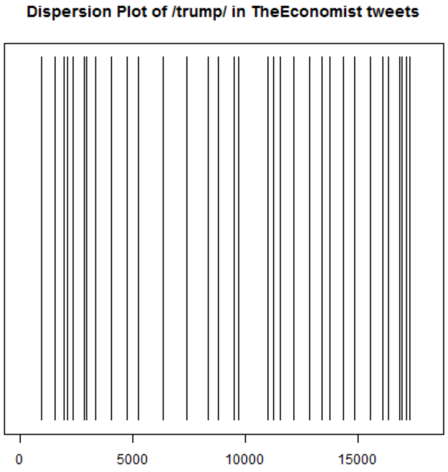




CNN writes about Trump quite frequently in its articles – this is made clear by the discrete lines or bars in the graph to the left. However, on Twitter, its tweets constantly mention Trump, even if there is no apparent article content related to him on the CNN site. The fact that these two media, despite reflecting the same period of time, show very different use (or maybe abuse) of the "Trump" keyword may suggest that CNN crafts its social media content to reference Trump disproportionately often, as this may attract more readers interested in hearing about hot topics related to the new US president.

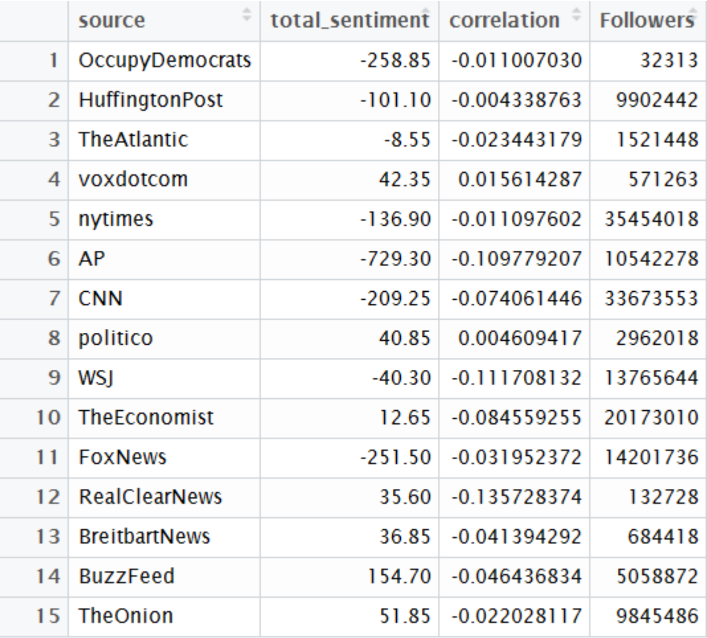
As a contrasting example, take The Economist:





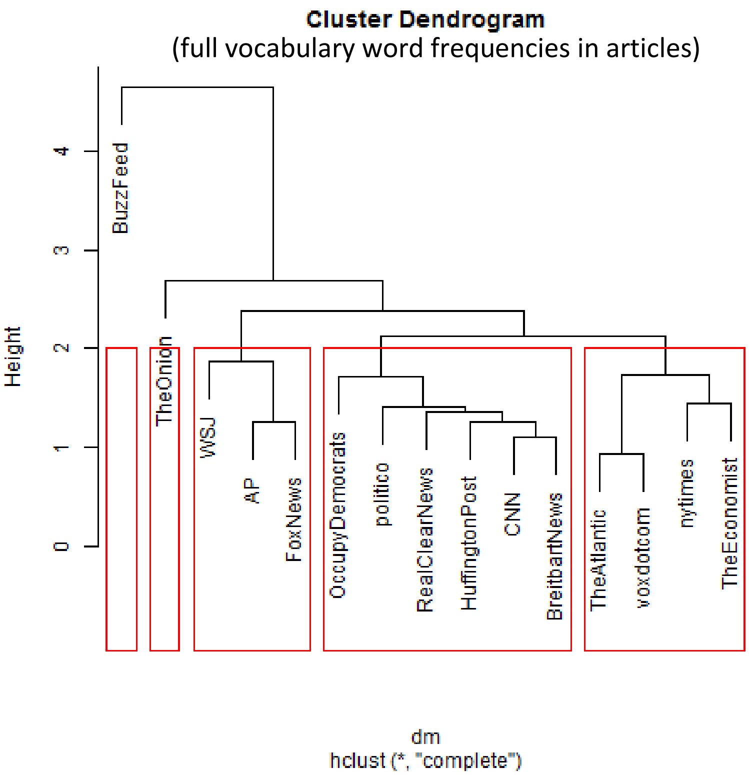
The data shows that The Economist, while writing frequently about Donald Trump, tweets about him modestly, in apparent proportion with their site content. In contrast with CNN, The Economist reflects what one would expect of a comparison between social media and article keyword use, suggesting a low degree of reader-targeted keyword abuse. Taking these two examples, this dispersion activity shows that news sources have the tendency to (or not to) overuse hot topics for the sake of attracting readers. However, seeing as this analysis is merely exploratory, it produces no definitive proof of the phenomenon.

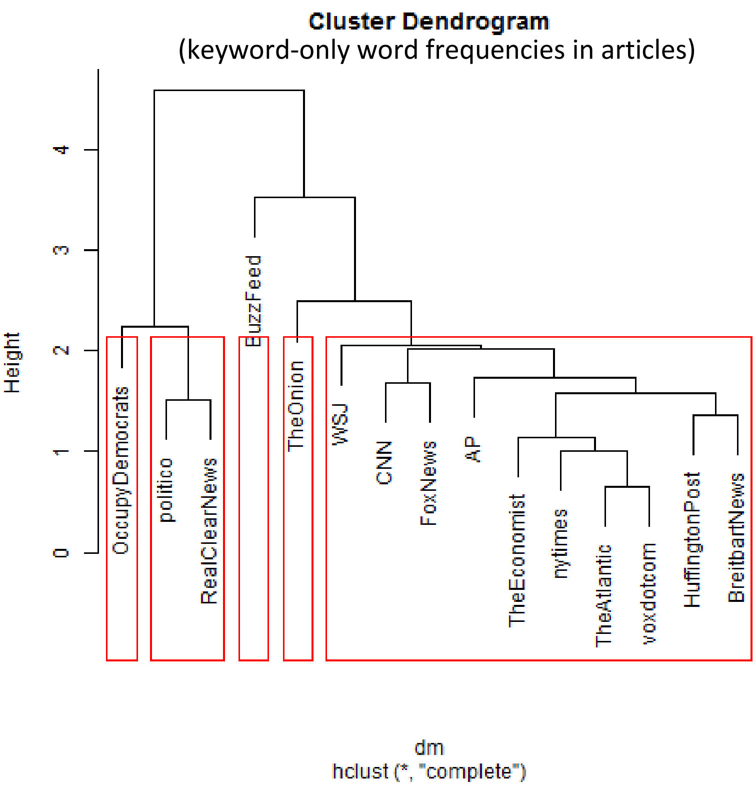
The second aspect of the analysis relates tweet sentiment to follower engagement. By tabulating total sentiment scores, correlation coefficients between individual source tweets and follower retweets, and source follower counts, one can numerically estimate the role of tweet sentiment in audience engagement. Here, the mean correlation coefficient is -0.046, meaning sentiment matters very little in explaining variation in retweets of source social media content. Further, the correlation between total source sentiment and source follower count is -0.25. With a p-value of 0.37, one concludes that the total sentiment cannot explain the popularity and following of particular news sources. Altogether, these two data imply that sentiment plays little to no role in predicting follower engagement, so manipulating the emotion of social media content may not be a plausible way for news sources to attract readers, as one may have expected. However, it can be noted that source following arises from perceptions of a source's legitimacy, which remain independent of sentiment given that older and more established sources will naturally have a greater following.



	source	total_sentiment	correlation	Followers
1	OccupyDemocrats	-258.85	-0.011007030	32313
2	HuffingtonPost	-101.10	-0.004338763	9902442
3	TheAtlantic	-8.55	-0.023443179	1521448
4	voxdotcom	42.35	0.015614287	571263
5	nytimes	-136.90	-0.011097602	35454018
6	AP	-729.30	-0.109779207	10542278
7	CNN	-209.25	-0.074061446	33673553
8	politico	40.85	0.004609417	2962018
9	WSJ	-40.30	-0.111708132	13765644
10	TheEconomist	12.65	-0.084559255	20173010
11	FoxNews	-251.50	-0.031952372	14201736
12	RealClearNews	35.60	-0.135728374	132728
13	BreitbartNews	36.85	-0.041394292	684418
14	BuzzFeed	154.70	-0.046436834	5058872
15	TheOnion	51.85	-0.022028117	9845486

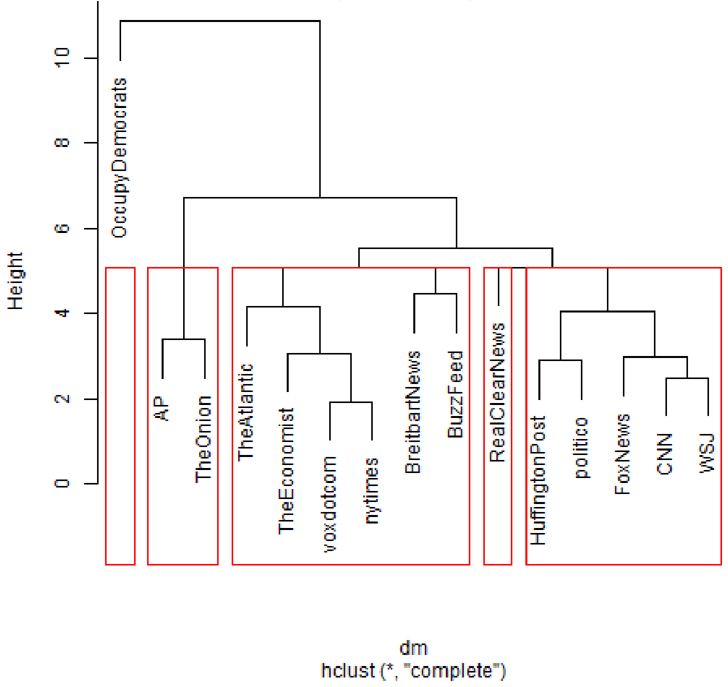
Moving on to the clustering analysis, we see interesting patterns for the article corpus data and more muted patterns when it comes to tweet data. These first two describe the former:

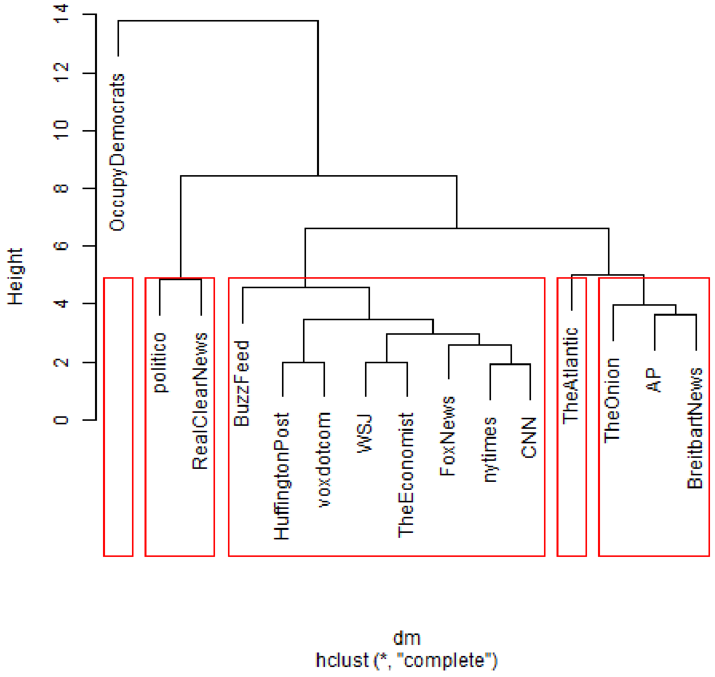




The full vocabulary data makes interesting groupings. The Atlantic, Vox, the New York Times, and The Economist, all sophisticated news sources, come grouped together, as one would expect. The more easily accessible news sources, ranging for far left source Occupy Democrats to alt-right source Breitbart News, come in a middle group, suggesting meaningfully that they share a common and accessible vocabulary. On the left, The Onion is duly separated – it is, after all, a satirical magazine that would have quite different article content – and so is BuzzFeed, which is commonly recognized as a pop-culture media organization with little informational content and mostly “click-bait” headlines. When it comes to the keyword plot on the right, we begin to see segregation by topic. Occupy Democrats, Politico, and RealClear Politics all write extensively on political issues, whereas other sources have wider scopes. While the Onion and BuzzFeed remain isolated because of the uniqueness of satirical and pop-culture topics, most of the other sources get grouped together, reflecting a wide and common variety of news topic coverage. This data suggests that, firstly, news sources can be grouped sensibly by their word distributions into classes that reflect writing style and, secondly, they can be grouped by their keywords, a useful proxy for understanding topic preferences.

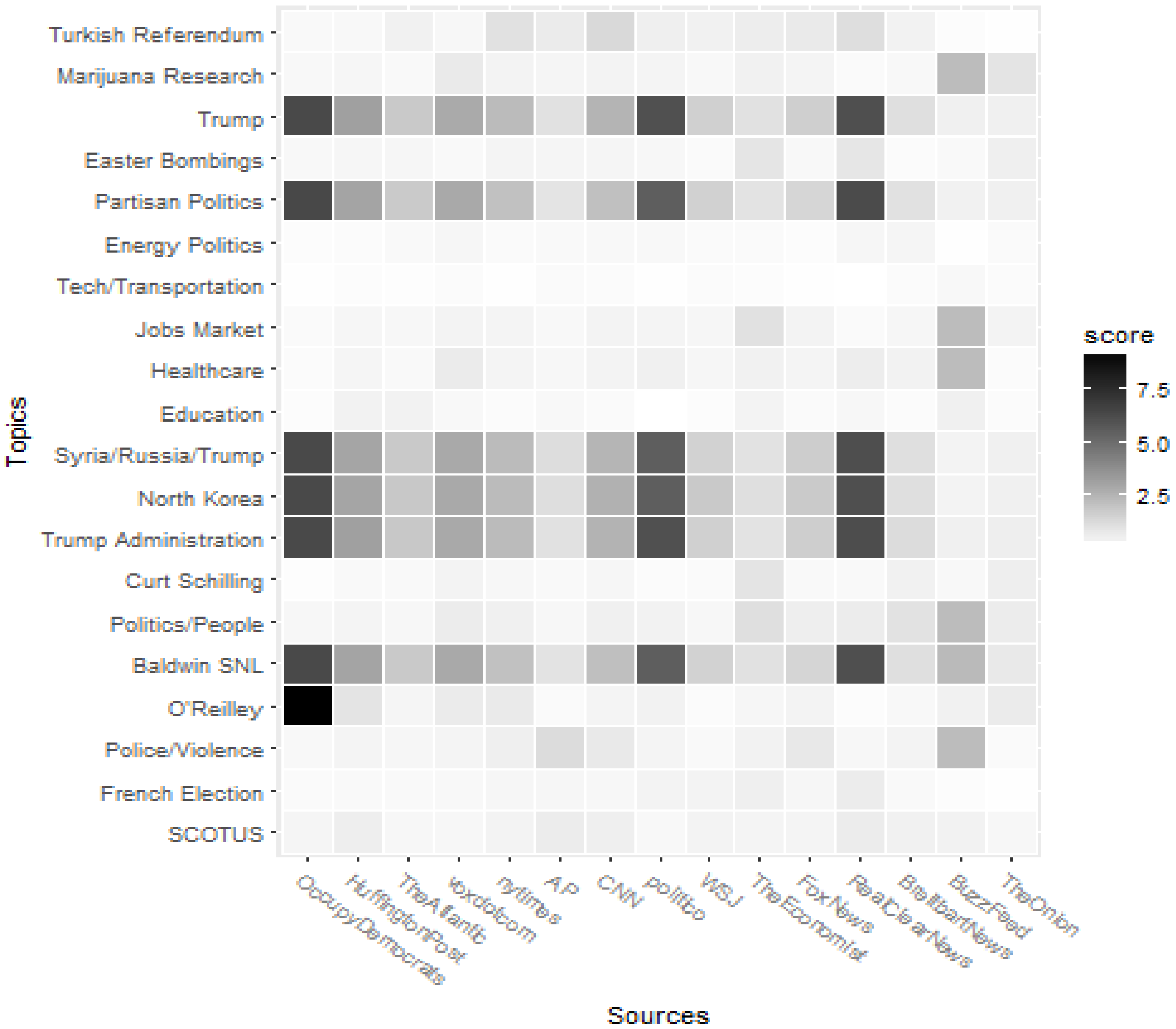
The cluster analysis is less clear for tweet data:


Cluster Dendrogram
(full vocabulary word frequencies in tweets)


Cluster Dendrogram
(keyword-only word frequencies in tweets)

In this case, we see more muddling of the previous groups, for both full vocabulary and keyword-only plots. Seeing as these will relate tweets of each news source to each other, these curious grouping changes hint that variety of tweet style and keyword use has less room for variation, as the information medium is condensed into 140 characters. Interestingly, this suggests that news sources sound alike when posting on Twitter, and topic preferences becomes less distinguishable, perhaps as a result of the previously mentioned trend of audience-targeted content. Given the same, information-hungry Twitter audience, all sources will tailor their tweeted content to the same topics most desired by users, with the purpose of attracting the most potential readers.

Finally, the topic model provides the strongest evidence of topic bias among news sources. The initial output of the model, combining article corpora of each source, produces 20 topics (See Appendix, A.1), which have been labeled to best describe the group of top words. These topics are plotted below in a heat map that measures closeness of news source word frequencies to each topic:



Many observations arise from this plot. First, the extremely political sources – Occupy Democrats, Politico, and RealClear Politics – show a clear inclination toward political topics such as Trump, North Korea, chemical weapons in Syria, etc. Since these sources are already known to be more political, they serve as a litmus to validate the method. Moving along, one sees BuzzFeed's clear preference for lifestyle topics – police violence, healthcare, jobs, Marijuana, etc. – over political ones. In between these two poles, mainstream news sources display varying degrees of political topic preference, from Huffington Post and Vox, which have relatively high political affinity, to The Economist and Associated Press, which better balance their topic biases.

On the whole, this map shows us that news sources certainly prefer some topics over others, and their degrees of bias are quantifiable.

CONCLUSION

This study seeks to better understand the content manipulation of news sources online. First, it studies the way these organizations communicate information differently whether using social media or their site as a platform. Toward this question, the findings suggest that news organizations, although paying little attention to sentiment in tweets, can disproportionately mention keywords or hot topics on social media. This represents a tactic to draw in viewers, who are naturally more interested in particular media topics, and so flock to sources that seem to cover them. The Twitter data topic cluster further supports this interpretation, as it shows that sources of all types tweet in similar styles about similar topics, despite clear differences in political leaning and news source category. So, news sources express demonstrable content manipulation on social media. But, do their sites show coverage bias toward particular topics?

The second series of analyses concludes yes. By looking at the news source article data cluster, different sources can meaningfully be grouped by the keywords they use in their articles. Concretely, the heat map shows that some sources' keyword frequencies much more closely match word frequencies of some key topics over others. Further, the mapping of sources like the Associated Press and The Economist contrast topic-biased content, showing what it looks like to have unbiased topic coverage.

Ultimately, this data provides a compelling commentary on the economic motivations of news sources. Although the methodology described above remains exploratory and unrefined, it does produce enough evidence to make one think harder about the constraints of today's news companies. More and more, legacy news organizations suffer increasing losses in print advertising. Having to rely more heavily on digital media for viewership, and digital advertising for income, these organizations find themselves competing in the fierce and chaotic market for digital attention, relying on audience engagement and viewership in order to stay afloat. Although this coercive power of viewership has always existed for print advertising, the legitimacy of print news allowed for unbiased topic coverage. Today's news industry is structured differently. The consumption of digital content, often through media platforms like Google, Facebook, or Twitter, places content sources neck and neck on search results or social media feeds to compete for audience attention. This structure allows those with lower media literacy to give less consideration to the originator of the content, and more to the subject of the content itself. This behavior forces news sources to cater toward these mass topic preferences, resulting in biased coverage.

These trends will only worsen with the increasing reliance on digital information platforms. People's attention will be pulled in more and differing directions, and attention-seeking media organizations will more aggressively seek viewership. But the future of digital information need not be so dramatically grim. Improved media literacy will increasingly support unbiased news sources; and the cooperation of digital content platforms will further reward unbiased sources with higher search or feed priority. There is a future where content is not a function of politics or audience preference. Although we may not be there yet, critically thinking about the information we consume, and how we consume it, is an important step in the right direction.

Appendix

A.1 - METHODS AND TOOLS

- **Collection of Twitter data** – “rtweet” R package, Michael Kearney
- **Web Scraping** – “rvest” R package and “stringr” R package, Hadley Wickham
- **Dispersion plotting** – method from “Dispersion Plots”, Chapter 4.1, *Text Analysis with R for Students of* Literature, Matthew Jockers (2014)
- **Sentiment Analysis** – “syuzhet” R package, Matthew Jockers
- **Clustering** – unsupervised clustering method from Chapters 11.6-11.9, *Text Analysis with R for Students of* Literature, Matthew Jockers (2014)
- **Topic Modelling** – “mallet” R package, David Mimno
    - Mallet LDA topic modelling method from Chapters 13.2-13.6, *Text Analysis with R for Students of* Literature, Matthew Jockers (2014)
- **Heat Mapping** – “ggplot2” R package, Hadley Wickham
- **Helper Functions for Topic Modelling** – “textman” text manipulation functions based on Jockers 2014, R. C. Alvarado

A.2 - MALLET LDA MODEL - TOPICS

```
[[1]]
      words     weights
1     court 0.028073608
2    united 0.017519847
3   supreme 0.010180430
4   justice 0.009269699
5     judge 0.009162554
6   federal 0.009055409
7    flight 0.008305396
8  thursday 0.007608955
9   gorsuch 0.007555382
10 arkansas 0.007233948

[[2]]
         words     weights
1       french 0.016341594
2       france 0.014387929
3           le 0.014387929
4          pen 0.014151121
5  immigration 0.010539801
6       macron 0.010006983
7   immigrants 0.007402097
8        paris 0.006869279
9     election 0.006810077
10       party 0.006573269

[[3]]
          words     weights
1        police 0.021169295
2        people 0.007556707
3          told 0.006730775
4        killed 0.006149564
5        attack 0.005629532
6      shooting 0.005323631
7   authorities 0.005109501
8      officers 0.005017730
9         found 0.004803600
10     security 0.004742420

[[4]]
         words     weights
1         news 0.023999635
```

```
2         fox 0.015383776
3      reilly 0.014375090
4     twitter 0.009247603
5       women 0.008365002
6       media 0.007272259
7        sign 0.007062116
8  newsletter 0.006515745
9      sexual 0.006095459
10     follow 0.006053430

[[5]]
     words     weights
1  baldwin 0.014606815
2    trump 0.011796430
3     time 0.007623435
4     film 0.005920172
5    night 0.004898214
6     life 0.004685306
7   people 0.004642724
8       tv 0.003578184
9      day 0.003535603
10    love 0.003407858

[[6]]
       words     weights
1     people 0.017845030
2   american 0.006311670
3      world 0.006255818
4        don 0.005376155
5  political 0.005278415
6       time 0.005236526
7    country 0.004775750
8     change 0.004691973
9      media 0.003979865
10     human 0.003854199

[[7]]
       words     weights
1  schilling 0.010817582
2       time 0.008775325
3     season 0.007943294
4       game 0.006581789
5     series 0.006241413
6      world 0.005485021
7      girls 0.005182465
8       team 0.004842089
9       poem 0.004728630
10    boston 0.004501712

[[8]]
            words     weights
1           trump 0.043536955
2       president 0.024300407
3           house 0.021461913
4           white 0.016130148
5  administration 0.012217630
6           obama 0.008976377
7      government 0.008305112
8             tax 0.007384519
9        congress 0.006521464
10         policy 0.005965272

[[9]]
       words     weights
1      north 0.036806631
2      korea 0.030563271
3      china 0.023403053
4      trump 0.021263719
5  president 0.013579583
6    nuclear 0.013535923
7      south 0.011745869
8     korean 0.010959991
9    chinese 0.009519216
10   missile 0.007947461

[[10]]
       words    weights
1      trump 0.02500149
2      syria 0.01917946
3     russia 0.01550566
4      assad 0.01354422
5  president 0.01148938
6   military 0.01142711
7   chemical 0.01127144
8    weapons 0.01102237
9     attack 0.01058650
10    syrian 0.01049310

[[11]]
        words     weights
1      school 0.014839198
2        city 0.010454202
3    students 0.010403214
4          ms 0.009179494
5       women 0.008975540
6    children 0.008720599
7     schools 0.007037984
8     college 0.006986996
9  university 0.006681066
10  education 0.006171182

[[12]]
       words     weights
1     health 0.016047617
2       care 0.013141943
3     people 0.011016897
4        pay 0.010800056
5      money 0.010062795
6    program 0.007720909
7  obamacare 0.007113753
8       cost 0.006940280
9    federal 0.006810175
10      plan 0.006593334

[[13]]
       words     weights
1     cities 0.007658748
2    workers 0.007103409
3       city 0.006606526
4   economic 0.005963502
5  companies 0.005203564
6     market 0.004969737
7       jobs 0.004823595
8     people 0.004618997
9      world 0.004122115
10    growth 0.004063658

[[14]]
```

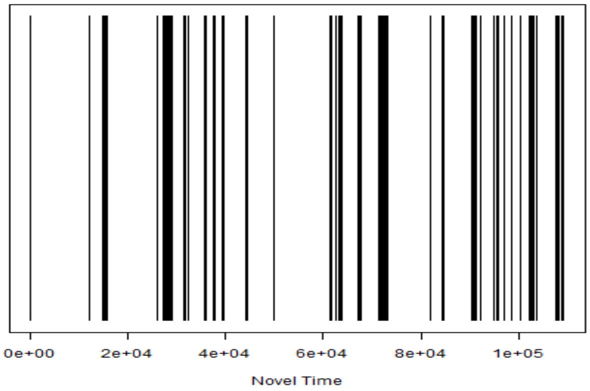
```
        words     weights
1       music 0.006875796
2        uber 0.006791961
3       songs 0.005785939
4     protest 0.005366763
5     parking 0.004570328
6        cars 0.004402658
7         car 0.004276905
8        song 0.004234988
9       space 0.004234988
10 technology 0.004193070

[[15]]
        words     weights
1      energy 0.010200969
2  department 0.008773857
3     million 0.007598588
4     company 0.007262797
5       water 0.006968979
6        coal 0.006675162
7     federal 0.006129501
8   companies 0.006087527
9         law 0.005793710
10 government 0.005248050

[[16]]
         words     weights
1   republican 0.018535715
2    democrats 0.017979019
3  republicans 0.017553311
4        trump 0.016734641
5     election 0.014638845
6     district 0.013754682
7       ossoff 0.010676482
8      percent 0.010447254
9        party 0.008940901
10         won 0.008646180

[[17]]
       words     weights
1     church 0.009157451
2     easter 0.008236458
3      world 0.006394471
4  christian 0.006394471
5      death 0.006340295
6        god 0.006286119
7        day 0.006123591
8  religious 0.005961063
9    suicide 0.005636006
10      life 0.005094245

[[18]]
       words     weights
1      trump 0.052714444
2  president 0.015833775
3   campaign 0.009202705
4    russian 0.008345882
5     donald 0.006706741
6  political 0.005775411
7       page 0.005700905
8      house 0.005365626
9      april 0.005291119
10     white 0.005291119

[[19]]
       words     weights
1    science 0.011445349
2   research 0.011113637
3  marijuana 0.009869719
4     people 0.007672129
5    percent 0.007506273
6      found 0.007091633
7      study 0.007050169
8       drug 0.007008705
9     health 0.006759922
10  evidence 0.006718458

[[20]]
        words     weights
1     erdogan 0.015802111
2  referendum 0.015045457
3      turkey 0.013131567
4        vote 0.012552949
5   president 0.011974331
6  government 0.011262186
7       party 0.010372005
8    minister 0.009036733
9    election 0.008858697
10 opposition 0.008769678
```

A.3 - SOURCE DISPERSION PLOTTING

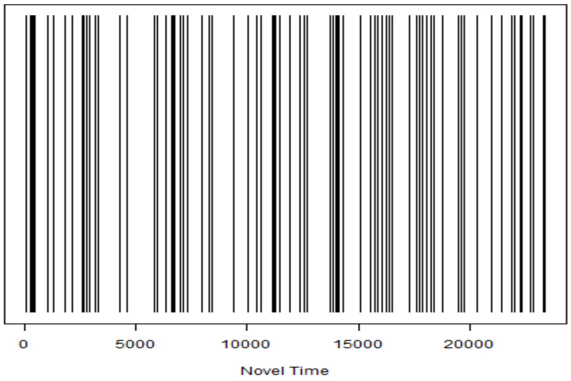
Dispersion plots comparing keyword use in tweets versus articles of selected news sources, limited to news sources with over 1000 tweets, after date and retweet trimming.

Keyword: "trump"

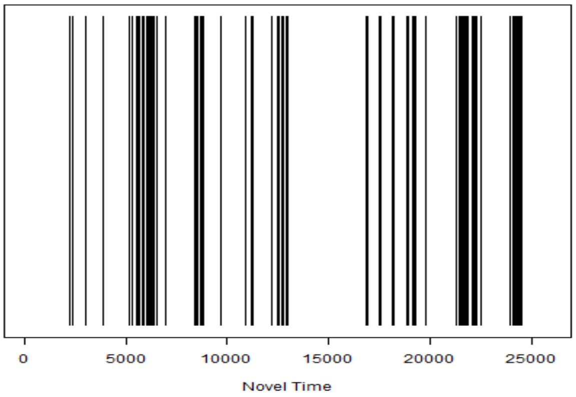
Dispersion Plot of /trump/ in AP articles



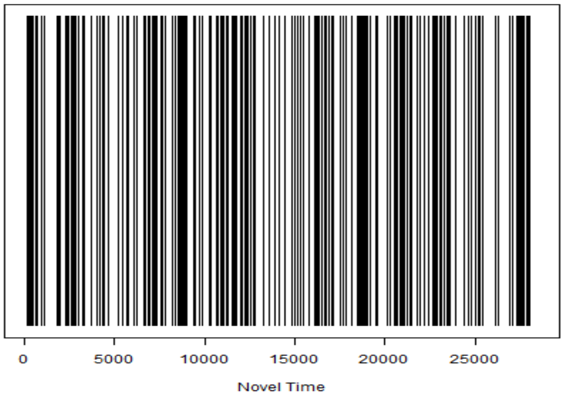
Dispersion Plot of /trump/ in AP tweets



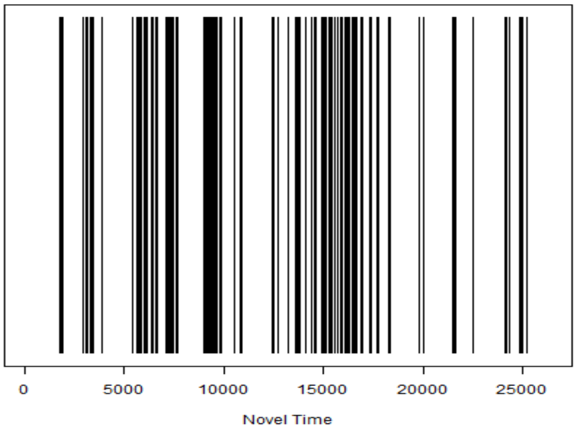
Dispersion Plot of /trump/ in CNN articles



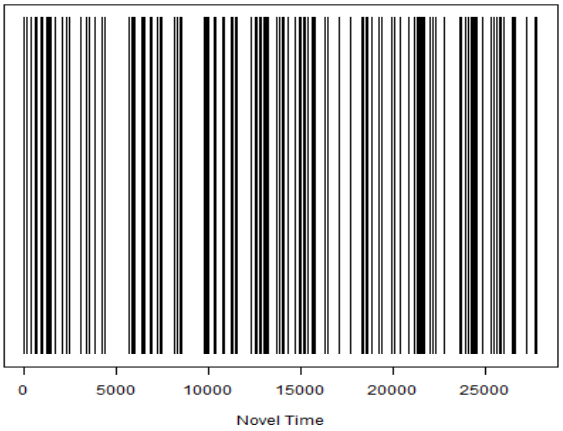
Dispersion Plot of /trump/ in CNN tweets

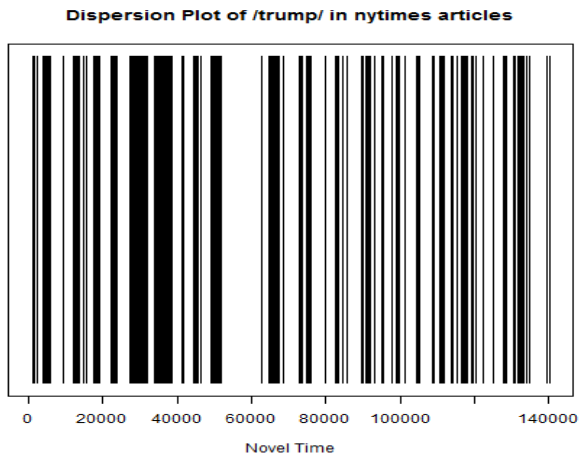


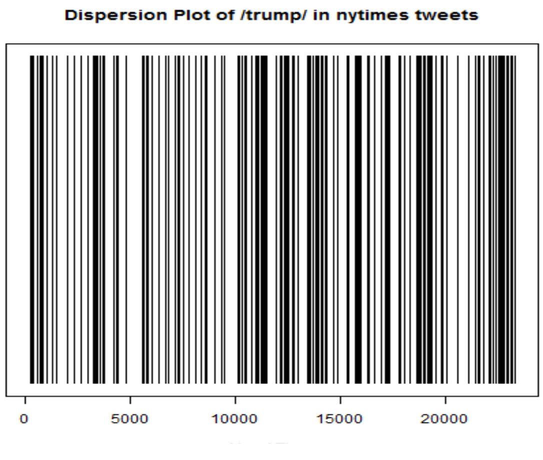
Dispersion Plot of /trump/ in FoxNews articles

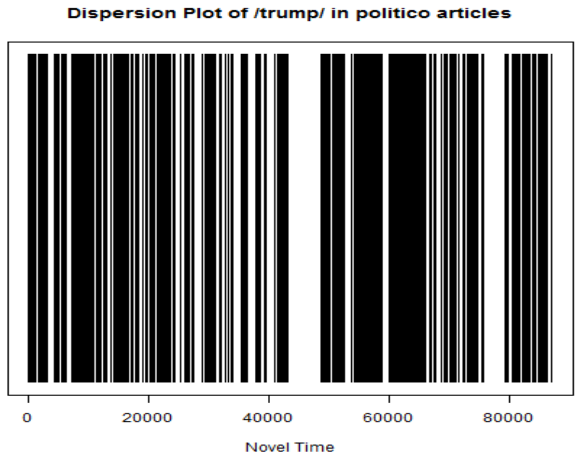


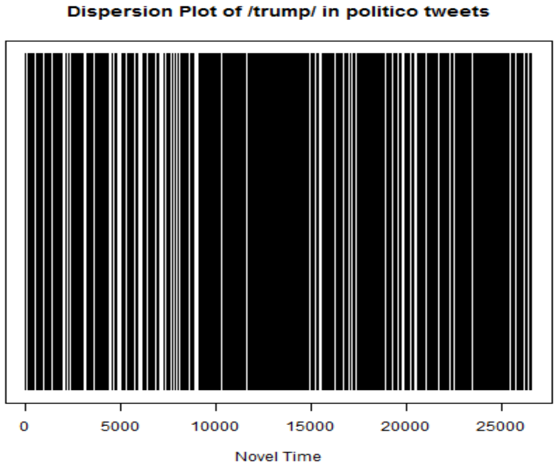
Dispersion Plot of /trump/ in FoxNews tweets

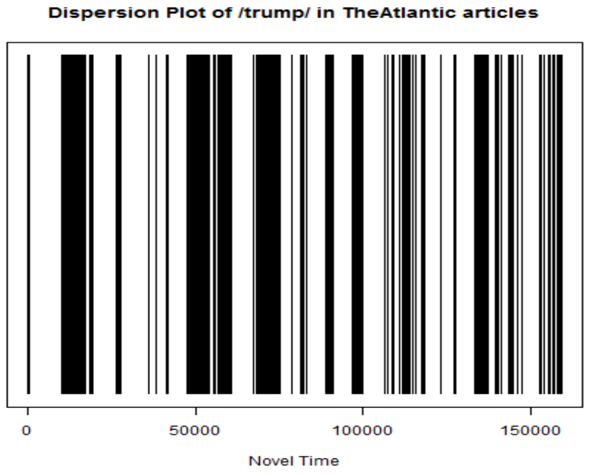


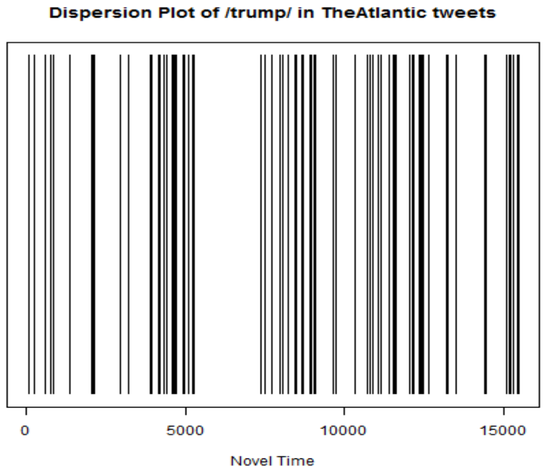

Dispersion Plot of /trump/ in nytimes articles

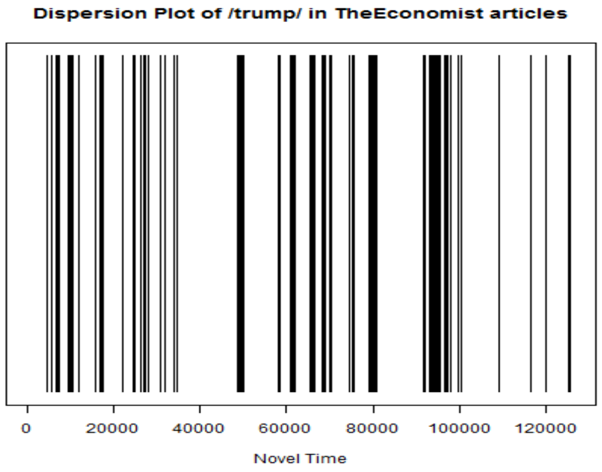

Dispersion Plot of /trump/ in nytimes tweets


Dispersion Plot of /trump/ in politico articles

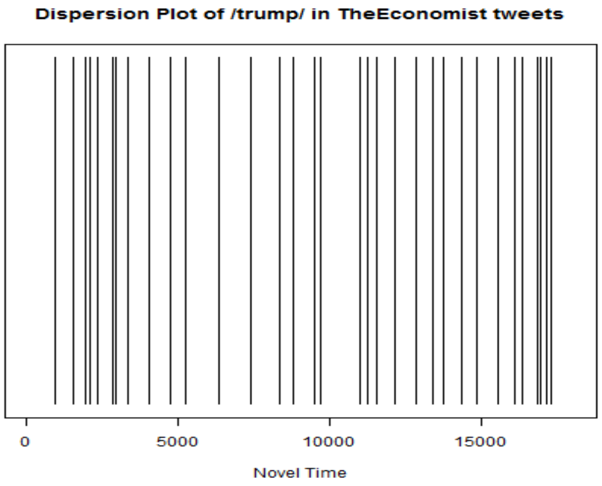

Dispersion Plot of /trump/ in politico tweets


Dispersion Plot of /trump/ in TheAtlantic articles


Dispersion Plot of /trump/ in TheAtlantic tweets

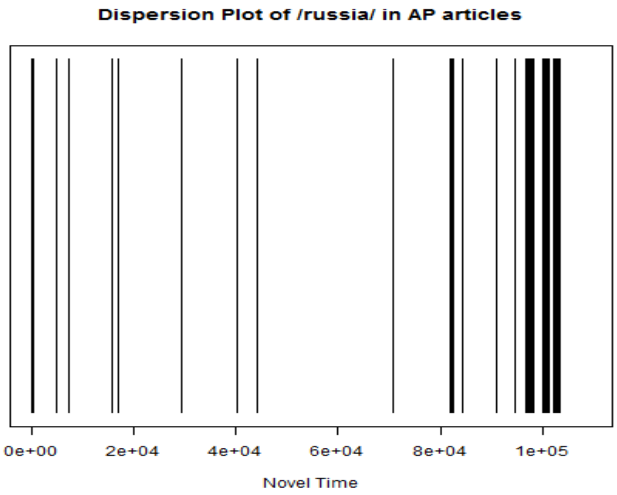


Dispersion Plot of /trump/ in TheEconomist articles

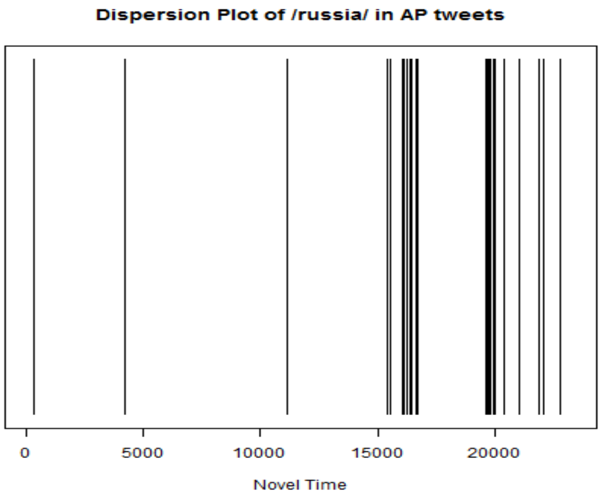


Dispersion Plot of /trump/ in TheEconomist tweets

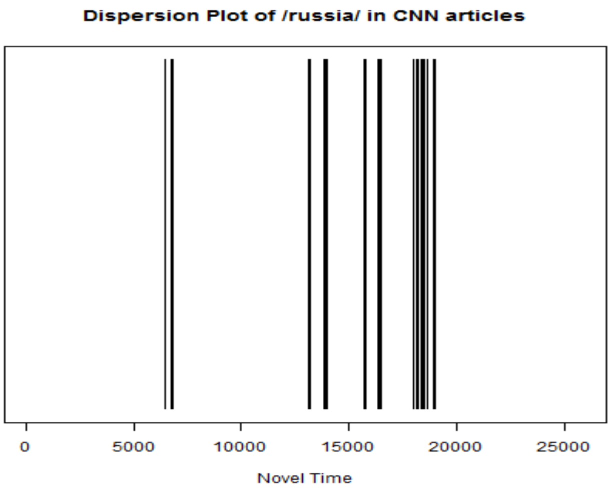
<u>Keyword</u>: “Russia”



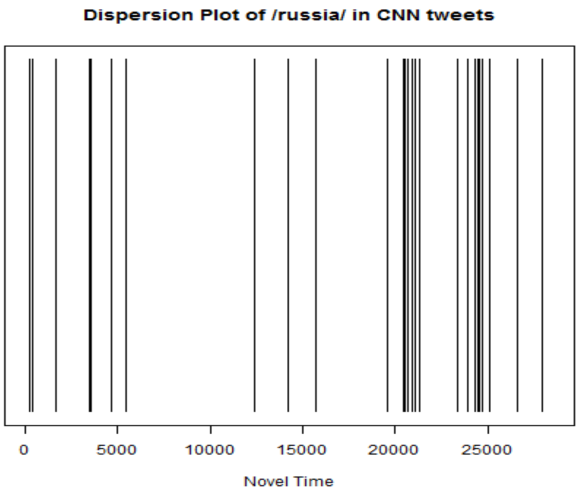
Dispersion Plot of /russia/ in AP articles



Dispersion Plot of /russia/ in AP tweets

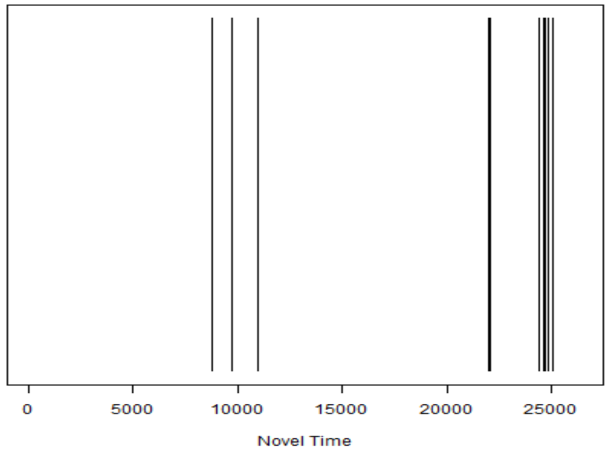


Dispersion Plot of /russia/ in CNN articles

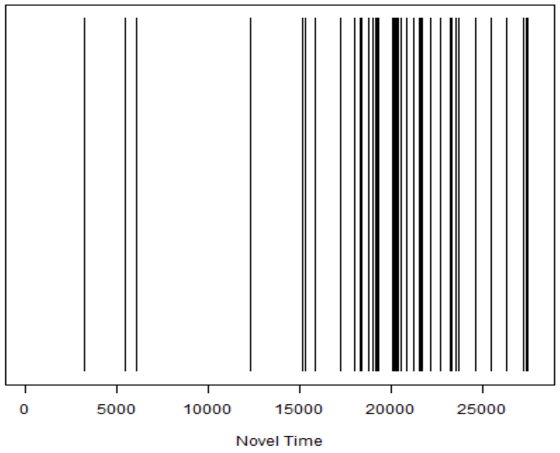


Dispersion Plot of /russia/ in CNN tweets

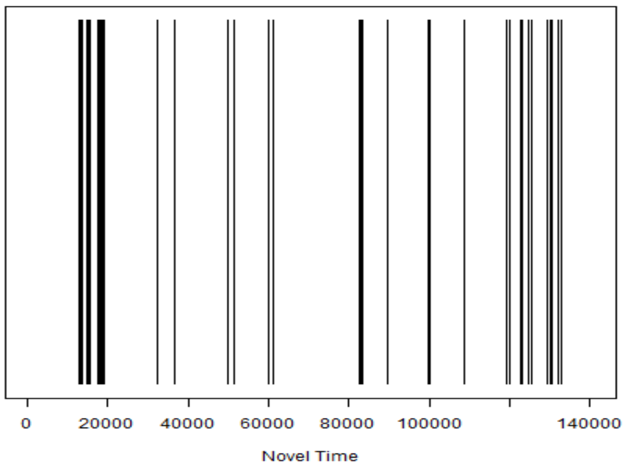
Dispersion Plot of /russia/ in FoxNews articles



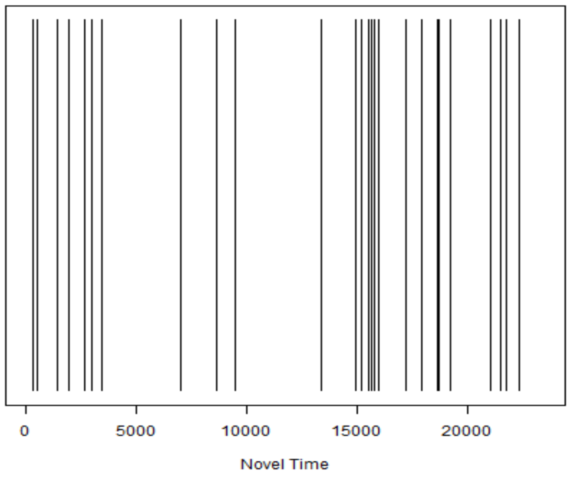
Dispersion Plot of /russia/ in FoxNews tweets



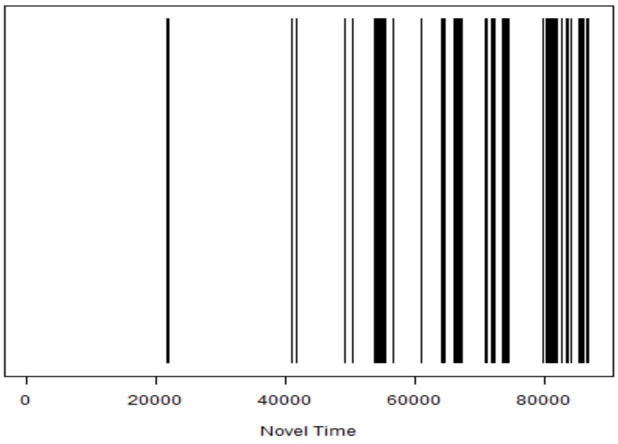
Dispersion Plot of /russia/ in nytimes articles



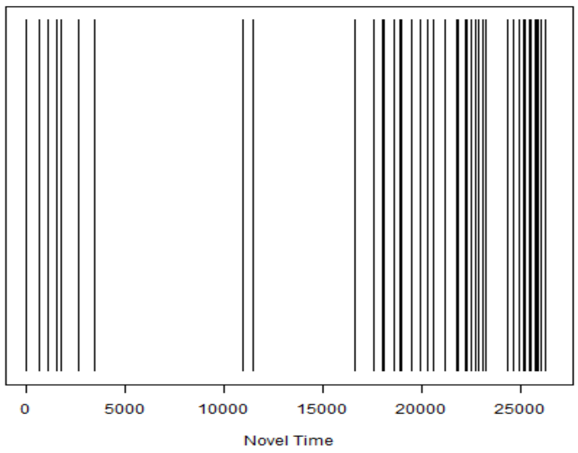
Dispersion Plot of /russia/ in nytimes tweets

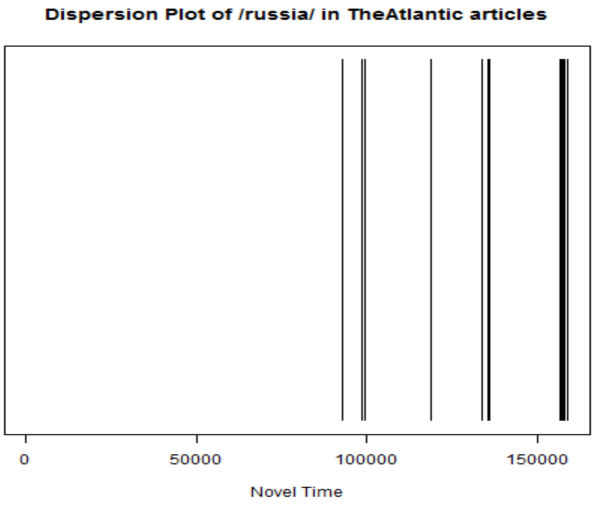


Dispersion Plot of /russia/ in politico articles

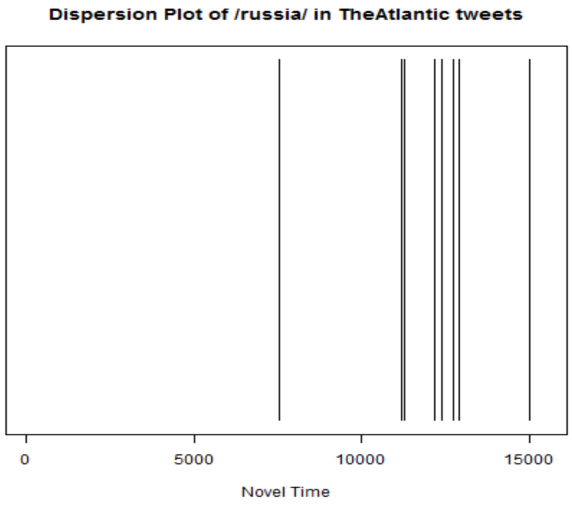


Dispersion Plot of /russia/ in politico tweets

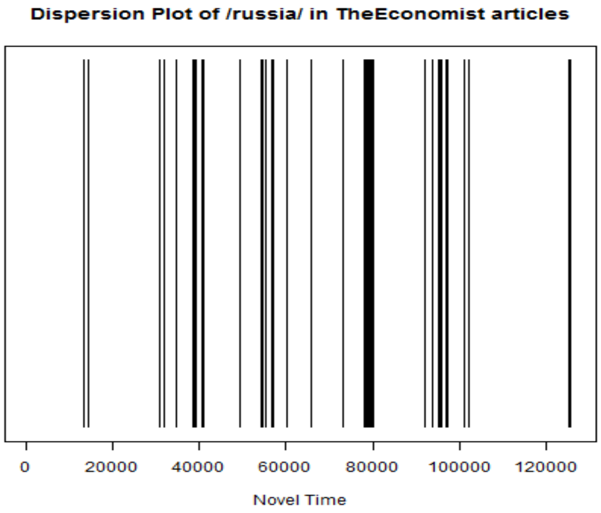




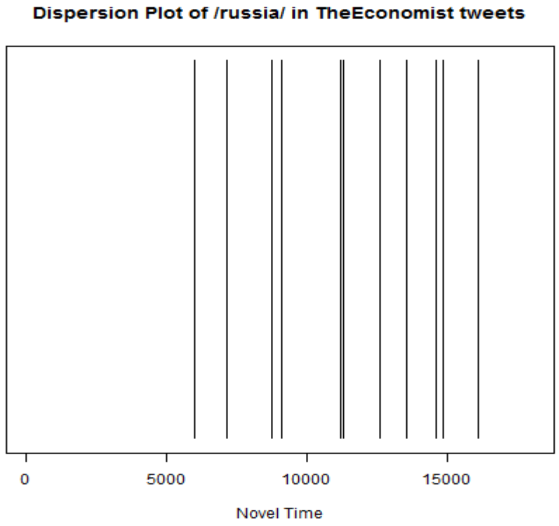
Dispersion Plot of /russia/ in TheAtlantic articles



Dispersion Plot of /russia/ in TheAtlantic tweets



Dispersion Plot of /russia/ in TheEconomist articles



Dispersion Plot of /russia/ in TheEconomist tweets