

report

March 7, 2021

1 Political Polarization of Major News Networks on Twitter

2 Abstract

We will construct a geometric definition of the political spectrum of major US news outlets through an unsupervised approach. We model the political alignments of the outlets in terms of pairwise political similarity among pairs of outlets using graphs and embed the graph onto a Euclidean space for result. We will be collecting hashtags used in the users' own timelines and cross reference it with the hashtags used during election period to classify their political stance as well as create a graph analysis between the news networks as a whole. Through this, we demonstrate the location where each news network lies on the U.S. political spectrum and how they lie relative in hashtag vector space to one another.

3 Introduction

Factual reporting is critical to the education of the general public on not only regular news, but more importantly to provide information paramount to understanding and interpreting the political atmosphere. Major news outlets have always played a key role in delivering important news and information to the public in a predictable and concise manner tailored to their viewers. As a result of this conformation to the preferences of their respective audiences, many news networks have developed a tendency to report news with a bias in various aspects of reporting; most notably, the most prevalent defining characteristic of a news network is its political affiliation. This often leads to skewed information motivated by viewership and rating results. This bias in reporting can often lead to creating confirmation bias in viewers who already agree with the sentiments being reported.

An example of the bipartisan split in television news networks can be seen in the contrast between CNN and Fox news. CNN is widely considered to be a left-leaning or democratic organization, while Fox is catered to a republican audience. This polarization of news is often criticized as furthering the tunnel vision in viewers by only showing them what they already agree with. Similar to "echo chambers" in *The Spread of Misinformation Online*, Vicario et al.[1], this action of reporting biased news creates isolated communities of viewers where information is often circulated within their own groups.

Although most news media outlets already have a pretty well defined political alignment, our analysis will involve further verifying this classification and comparing these news stations on different spectrums other than political affiliation. The question we want to answer is whether or not the users who actively interact with various news outlets conform directly to the political viewpoint of the outlet. Our goal in this project is to quantify the political inclination (pro-democrat

vs pro-republic) on the users of eight news stations. In other words, we want to construct a political spectrum and see how these new stations fall on the spectrum.

4 Data

4.1 Collection

Through use of the Twitter API, we are able to gather any data that was made publicly available on the Twitter platform. Per the terms of this API, we are unable to access any tweet that is protected by a private account or has been deleted. Although we are collecting tweet data from individual users, we will only be analyzing aggregated values from hashtags and will not be releasing any individual data points.

The eight news stations we chose were BBCWorld, BreitbartNews, CBS, CNN, FoxNews, MSNBC, RT_America, nytimes, and washingtonpost. We decided on these accounts as they are all relatively well-known and we wanted to have a sample of media outlets that were distributed along the political spectrum and therefore chose outlets that are left, neutral and right leaning. Since the goal of our project is to analyze the users that actively interact with each news station, we needed to gather a sizable sample of tweets to quantify the general trend of political affiliation. Our steps were as follows: sample the most recent tweets from each news outlet, gather all retweeters in each tweet, and then finally examine the retweeters by collecting the counts of hashtags used in each retweeter’s timeline.

The main portion of our data collection process was gathering the users that actively interact with the Twitter accounts of major news stations. We looked at the most recent 100 posts from each news outlet and collected every retweet and subsequently every user who retweeted the post. After collecting these users, we randomly sampled 500 users from each news station and gathered the most recent 1000 tweets from each user’s timeline. If a user did not have 1000 tweets, we simply took their entire timeline. This resulted in a minimum of 400,000 tweets in our dataset as the average user has less than 1000 tweets, we assume that the average lies around 100 per user.

From these tweets, we collected every hashtag used and took down the occurrence of each hashtag. To account for minor variations in hashtag spelling and syntax, we stored each hashtag in its lowercase form. Each news outlet now has a list of hashtags along with a mapping of the respective number of times used in a tweet. This information will be used as a hashtag vector in our graph analysis elaborated further in the methodology section.

Our definition of a user that “actively interacts” with a news outlet is someone who has retweeted one of the outlet’s tweets in the past 3 months. Although it may have been easier to collect users from the news outlets’ follower lists, we wanted to ensure our users analyzed were active on their Twitter so that we can analyze how they interact with current political accounts and tweets.

4.1.1 An overview of followers and traditionally believed political alignment

News/Media Outlet	Number of Followers	Traditional Political Alignment
FoxNews	20M	Towards Right
BBCWorld	30.9M	Middle
MSNBC	4M	Middle Left
CNN	52.6M	Middle Left

News/Media Outlet	Number of Followers	Traditional Political Alignment
Washingtonpost	17.4M	Towards Left
CBSNews	8M	Middle Left
nytimes	49.2M	Towards Left
RT_America	367.5K	Unknown
BreitBartNews	1.4M	Far Right

Data sourced from <https://guides.lib.umich.edu/c.php?g=637508&p=4462444>

4.2 Exploratory Data Analysis

As the end goal of our project is to utilize hashtag usage for comparison of news outlets, we wanted to begin by looking at the most commonly used hashtags relative to each news outlet. We believed that the most noticeable difference in hashtag trends would be politically motivated as the traditional consensus is that these eight news outlets all have some sort of political affiliation or bias in their reporting, attracting users of the same political alignment.

Hashtags only found in left	Hashtags only found in right
Blacklivesmatter	Oann
BidenHarris2020	Antifa
ConvictTrump	Maga
wearamask	pelosilovestump
	stopthesteal
	trump2020
	americafirst
	bidenlied
	bidencheated

To do this, we looked to aggregate hashtag frequencies and compare the distributions across each news outlet. There were a few common hashtags found across all eight news stations such as variations of “covid-19” and “trump”. These words have relatively neutral meaning in terms of political leaning and therefore looked into the effect of removing them in our methodology.

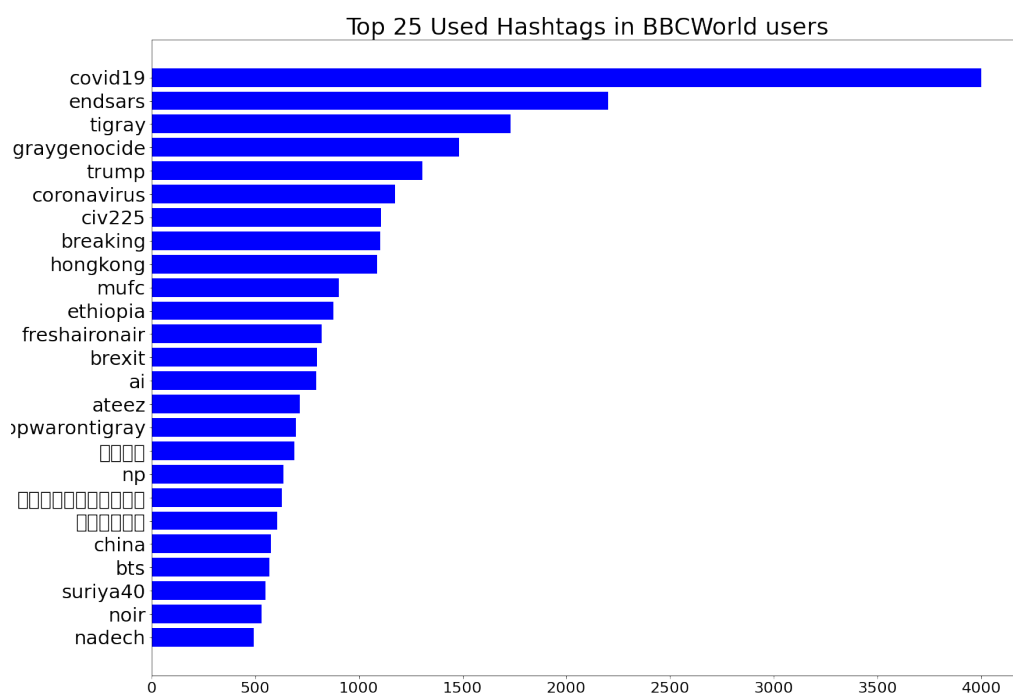
Shared hashtags with differing connotations
covid19
breaking
coronavirus
trump
fbi
election2020
biden
china
georgia

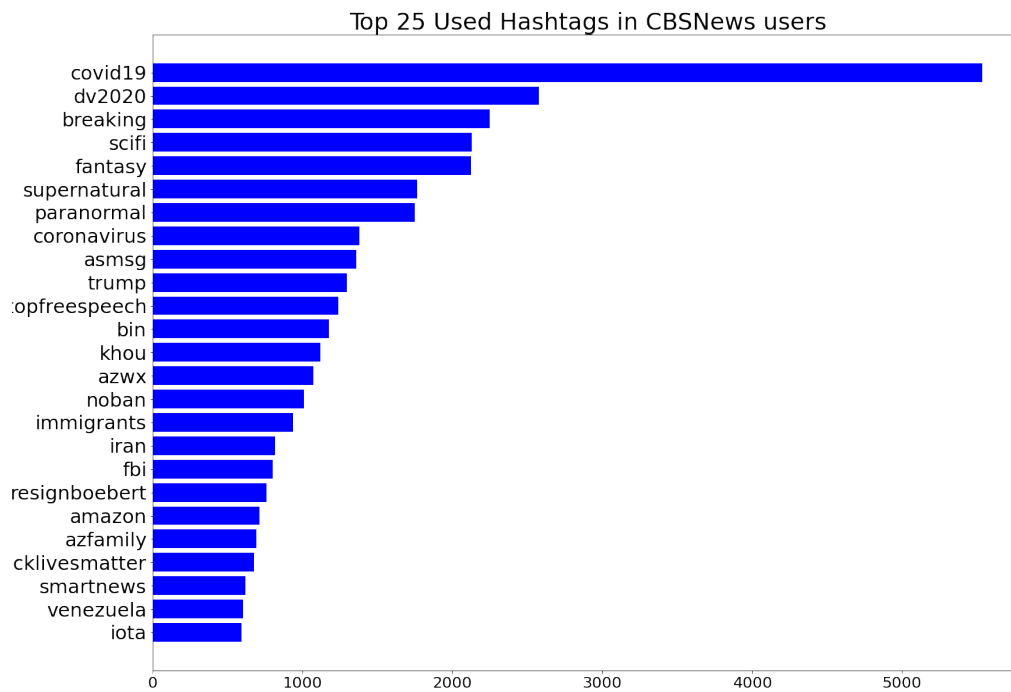
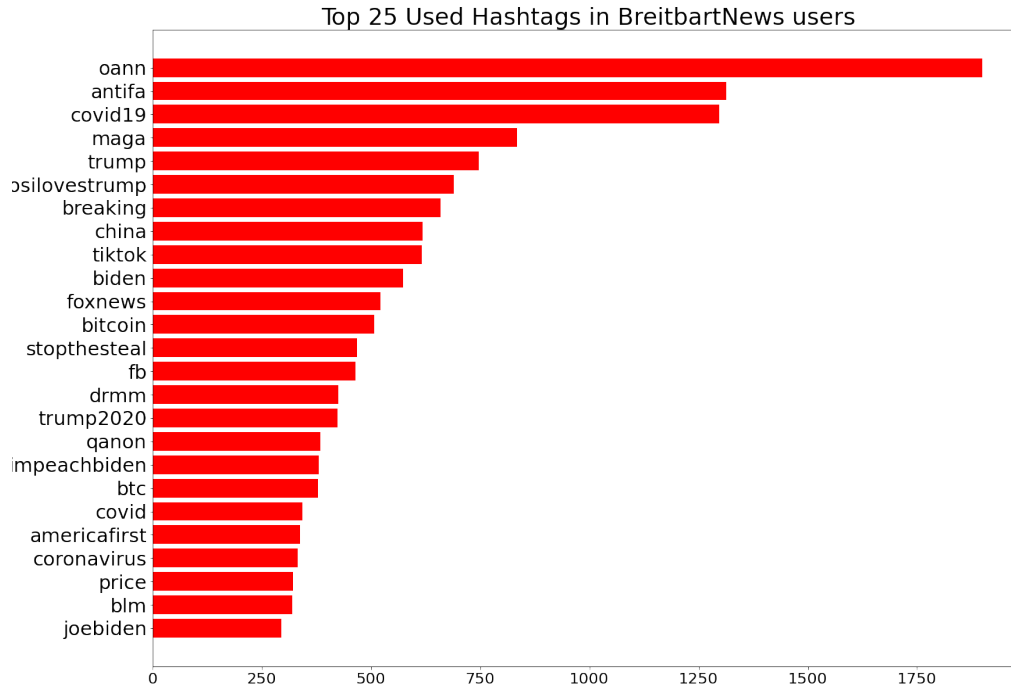
Our first visualization was a horizontal bar chart designed to display the most popular hashtags used in each news outlet. We hypothesized that there will be a quantifiable difference in the hashtags used by users of each news outlet due to the difference in population of their active users.

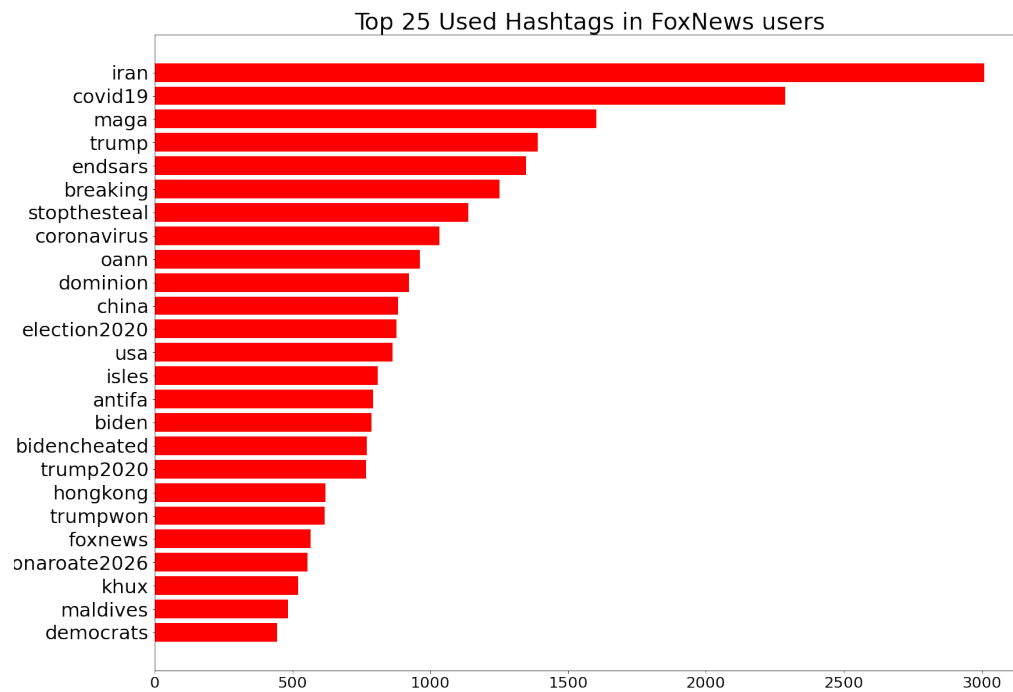
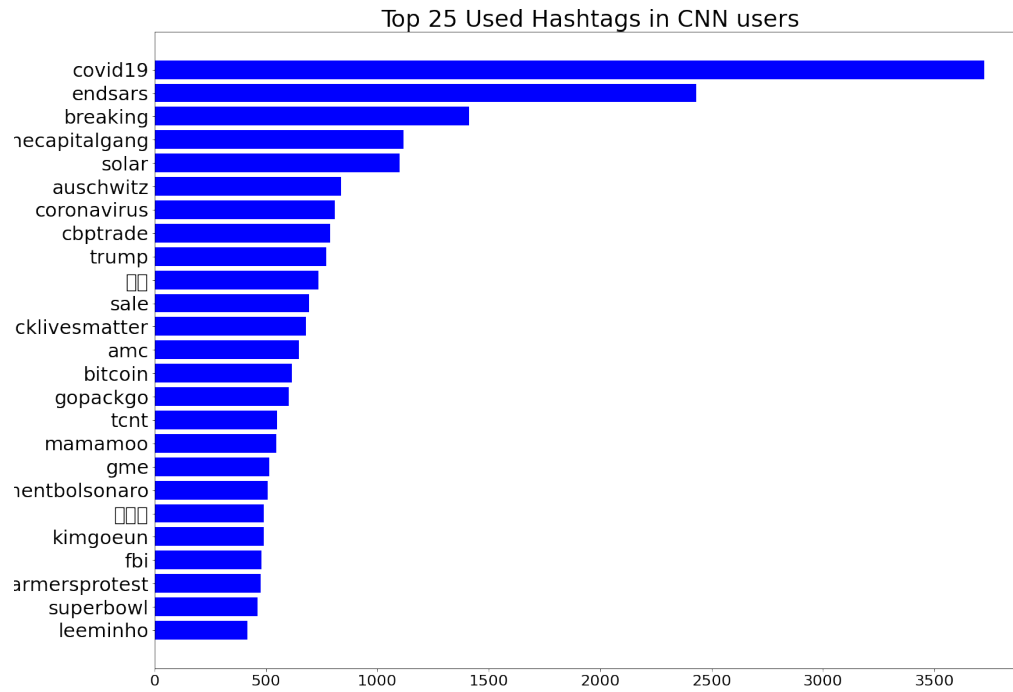
A brief glimpse into the figures below shows that there is indeed a noticeable difference in hashtag usage between users of each news outlet, more specifically with news outlets of differing political alignments.. We found that politically charged words are the most prevalent separation between each collection of hashtags.

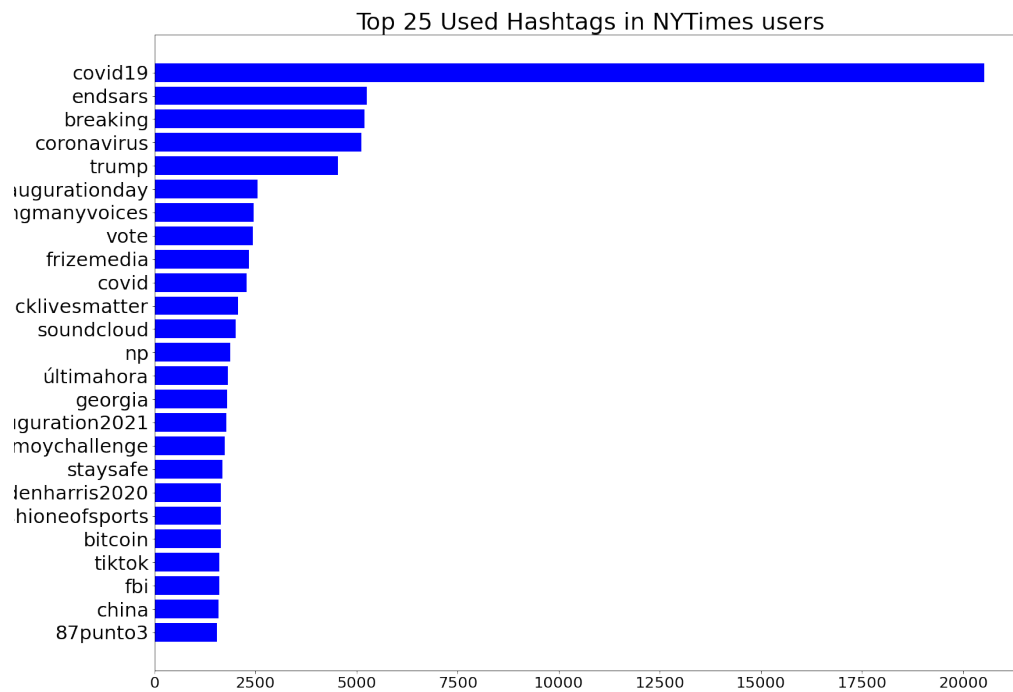
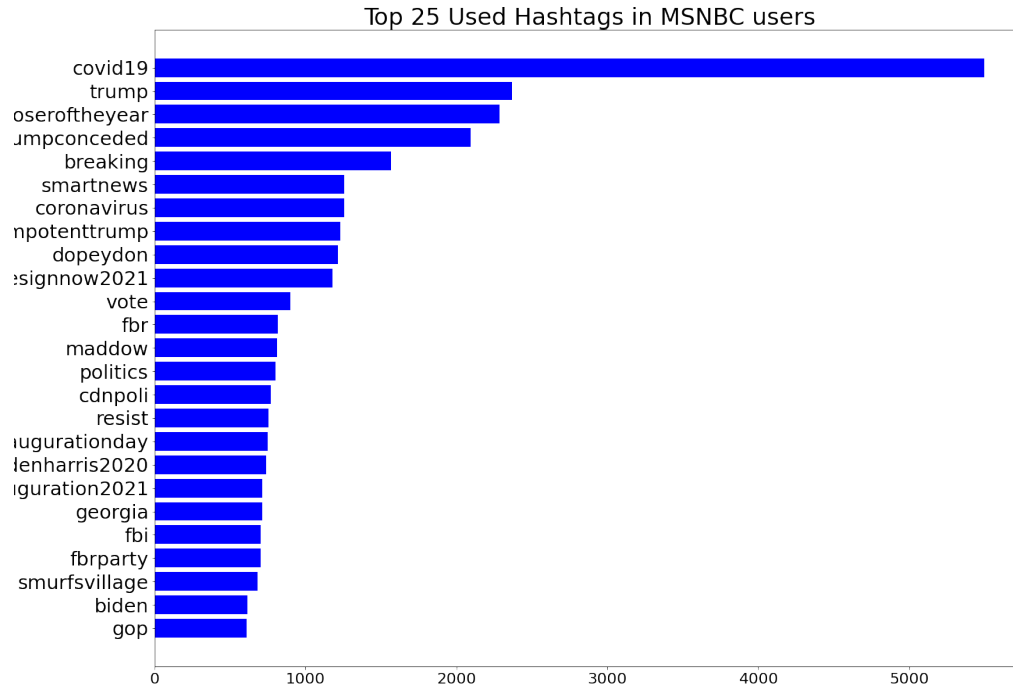
4.2.1 Hashtag Counts of Various News Outlets

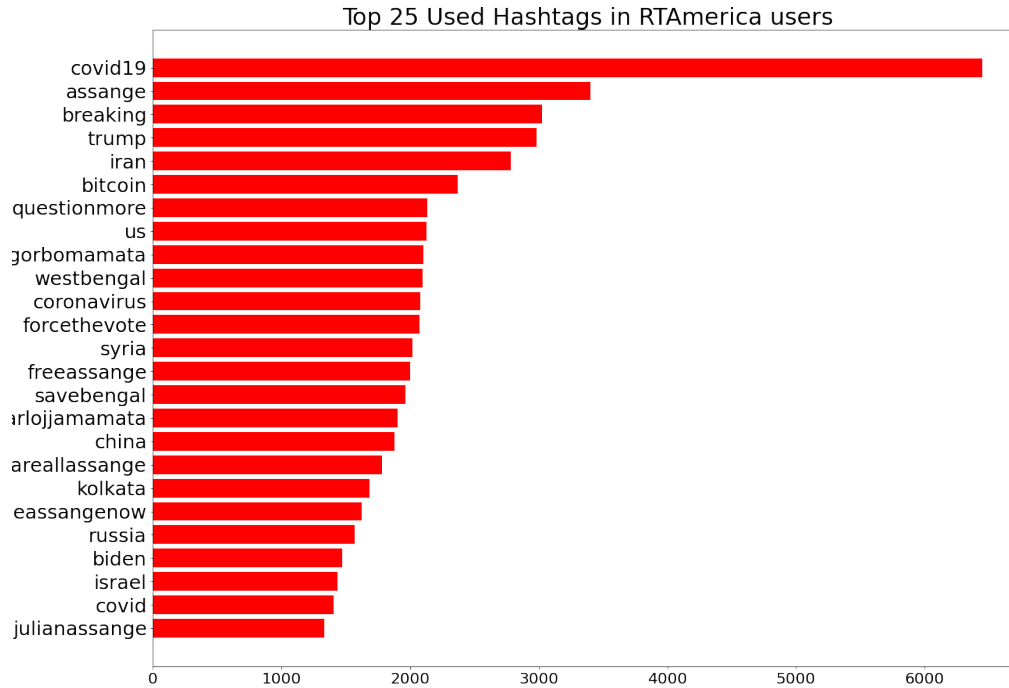
Red = Generally More Conservative, Blue = Generally More Liberal











5 Related Literature

Predicting the political alignment of users on social media has always been a topic of interest for many scholars and institutions to research. As social media grows in popularity, more and more users will begin to upload information about their personal lives, and oftentimes their political beliefs, into the public domain for others to interact with. With this ever growing plethora of information, many new approaches have been developed to better understand the characteristics of US voters as opposed to traditional census and polling practices. While our project focuses exclusively on analyzing the hashtag usage of our gathered users compared to an existing dataset of election related tweets, there are many other publications that investigate a user's actions in more detail.

In Predicting the Political Alignment of Twitter Users, Conover, Goncalves, Ratkiewics, Flammini and Menczer demonstrated several implementations of predicting the political stance of a Twitter user based on their tweets. The paper utilized the hashtags and tweet text to build a machine learning model for predicting a user's political stance. In a SVM model, the researchers were able to achieve a higher accuracy through metadata on hashtags versus tweet text. This coinsigns with our hypothesis that hashtags will provide the best viable separation in how users display their political stance. Their analysis also showed clear clusters that represented the two respective political groups, republicans and democrats. Whereas the researchers defined the political stance of hashtags through Latent Semantic Analysis to discover political affiliation of hashtags, our group will be plotting the hashtag vectors of each news outlet as a whole to demonstrate the differences of news outlets in terms of vector space.

6 Methodology

6.1 Overall Process

We are adopting an unsupervised approach towards quantifying the concept of political spectrum. In short, we define the political spectrum to be the 1-D Euclidean space where the news station lies and where outlets with similar political inclination would be close to one another. To transform the news stations into said space, we plan to construct a complete graph among the news stations - where the nodes are our news stations in question and the edges are weighted by our pairwise similarity (to be defined in the next paragraph) measure - and maps the graph onto the euclidean space through spectral embedding using Laplacian EigenMap. The resultant plot - of the nodes lying in the euclidean space in a fashion relative to their pairwise similarity - would be the main answer to our research question.

6.2 Defining Pairwise Similarity for Graph

We formally define the concept of similarity between two news stations to be the

$$1 - \frac{\sum \min(X_{1i} X_{2i})}{\sum \max(X_{1i} X_{2i})}$$

where X_1 and X_2 are feature vectors for the two news stations. The feature space of the vector is the hashtags used in tweets in the election dataset and the value for each feature is the normalized/unnormlized (dependent on configuration) number of occurrences of each hashtag in the user timeline dataset for each news outlet. Aside from the aforementioned optional normalization of count of occurrences, other configuration of the feature space includes removing overly neutral hashtags based on a set of pre-defined hashtags (such as Covid-19, coronavirus) in an attempt to remove overwhelming hashtags with overly neutral net implication. This method is our attempt to capture the similarity in political view between pairs of news stations in an unsupervised approach.

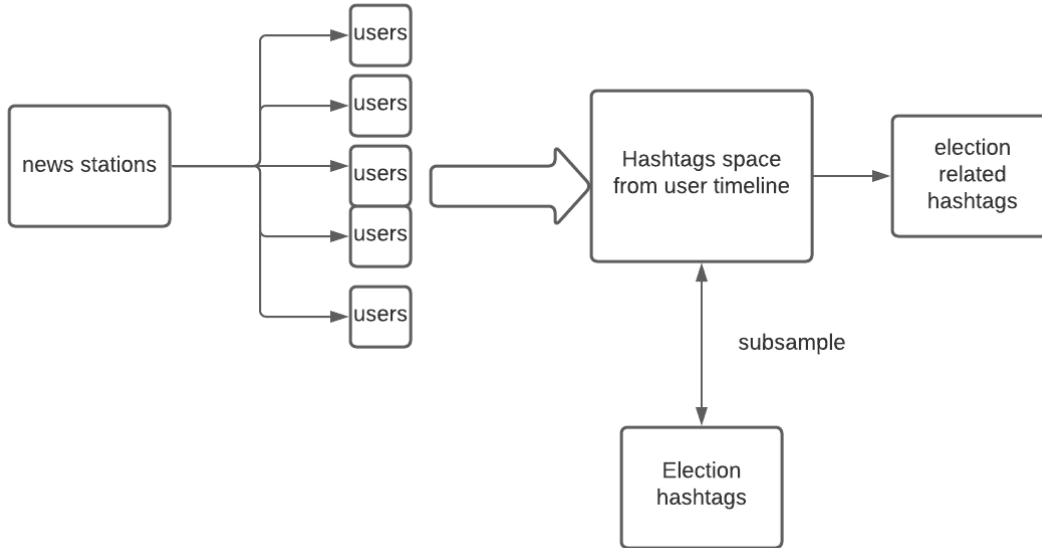


Fig.1

data pipeline for feature vector

6.3 Embedding Graph

The method we chose for graph embedding is Laplacian Eigenmap. The minimization goal of the method, which is

$$\sum_{ij} (y_i - y_j)^2 W_{ij}$$

(y denotes euclidean coordinate of a node and W_{ij} denotes the edge weight between the two nodes), rewards short pairwise euclidean distance based on edge weight. This largely coincides with our definition of a political spectrum and therefore is a sensible option for embedding.

To recap we define the position of news stations in a political spectrum as their relative position in euclidean space embedded from a graph that stores pairwise similarity, characterized as a function of two vectors of hashtags under the same feature space, as edge weights between vertices. There are a few advantages and disadvantages ostensible upon its conception.

7 Results

Through reducing the graph to a 1-D euclidean space, we can observe the pairwise similarity between the news stations. Again, it is important to note that the plot only captures the relative position of the news stations from the higher dimension euclidean space that they reside in. However, the relative distance between the news stations are still observable from the plot. The greatest distance is between Breitbart News and Fox News with the next greatest between CNN and New York Times. This finding is contrary to what is expected based on the traditional political alignments of the news stations being observed.

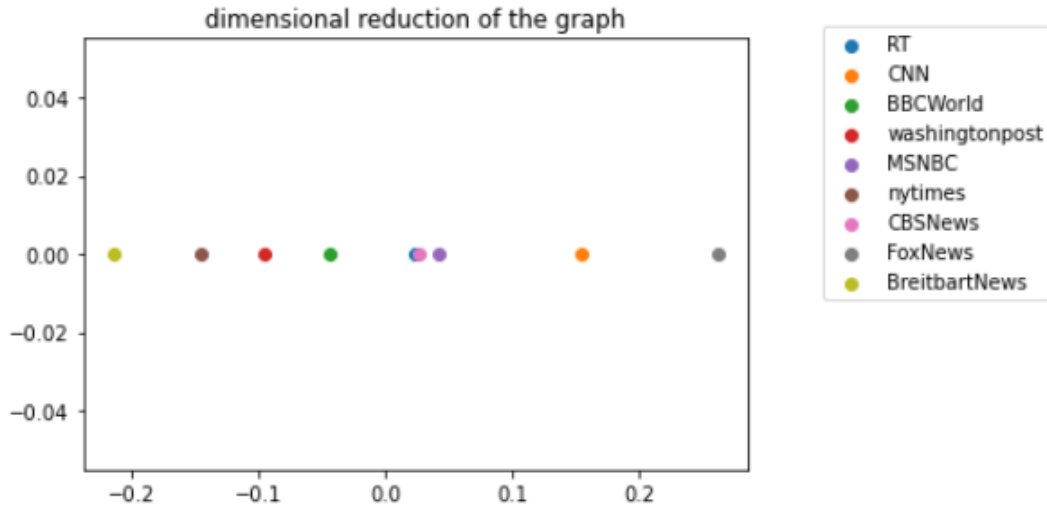


Fig 2.1

The results of the spectral embedding did not align with the expected results. This has several implications about the underlying assumptions made which were discussed in the methodology section. That would be that the use of hashtags is sufficient in representing the user's political affiliation and in turn, the news station as well. However, if taking the results at face value, this brings up further questions about the nature of the relationship between the viewer base and the news networks as well as the difference in behavior caused by the setting of social media and an online world.

8 Discussion

The results of the spectral embedding did not align with the expected results. This has several implications about the underlying assumptions made which were discussed in the methodology section. That would be that the use of hashtags is sufficient in representing the user’s political affiliation and in turn, the news station as well. However, if taking the results at face value, this brings up further questions about the nature of the relationship between the viewer base and the news networks as well as the difference in behavior caused by the setting of social media and an online world.

9 Conclusions

Initially, we hoped to discern the discrepancy between traditional notions of news outlets’ political affiliation and the political affiliation of their viewer base on social media. Our initial results did not align with our expectations which instead raised further questions about our methodology but also the nature of social media and the dynamics of online presentation. Though there were a number of limitations our team faced. To begin with, there was space and time limitations to collecting a more extensive dataset as well as utilizing a larger amount of users per news station partially due to api call limits.