report

March 7, 2021

1 Political Polarization of Major News Networks on Twitter

2 Abstract

We will analyze the users who interact with the Twitter accounts of various popular news networks to compare their alignment across the U.S. political spectrum. We will be collecting hashtags used in the users' home timelines 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 politcal 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
Washingtonpost	17.4M	Towards Left
CBSNews	8M	Middle Left
nytimes	49.2M	Towards Left
RT_America	367.5K	Unknown

News/Media Outlet	Number of Followers	Traditional Political Alignment
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.

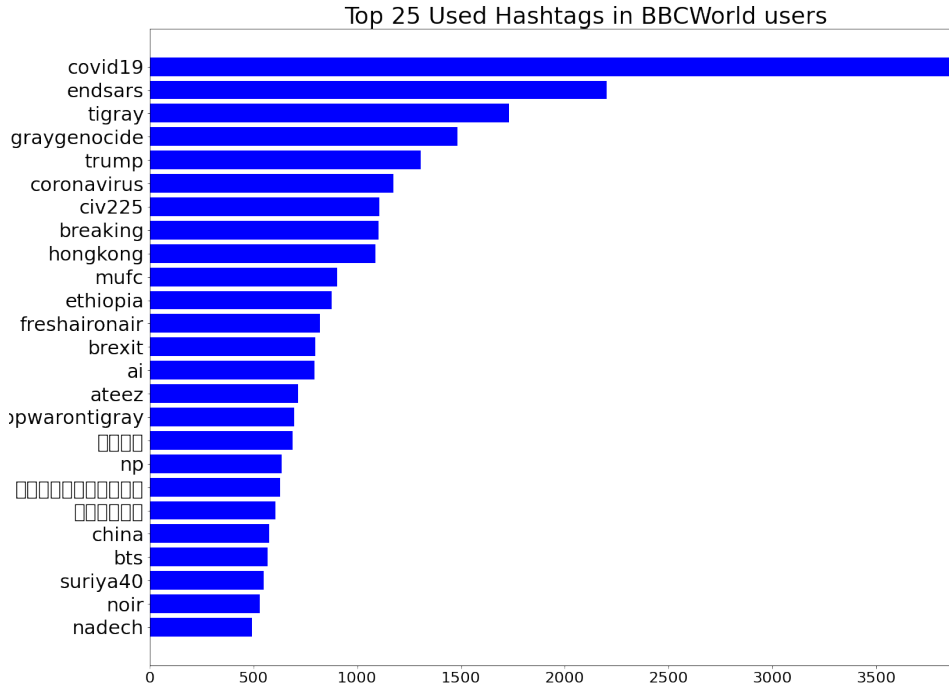
To do this, we looked to aggregate hashtag frequencies and compare the distributions across each news outlet. There were a few common hashtags found across all eight news stations such as variations of "covid-19" and "trump". These words have relatively neutral meaning in terms of political leaning and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

Our first visualization was a "Word Cloud" 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.

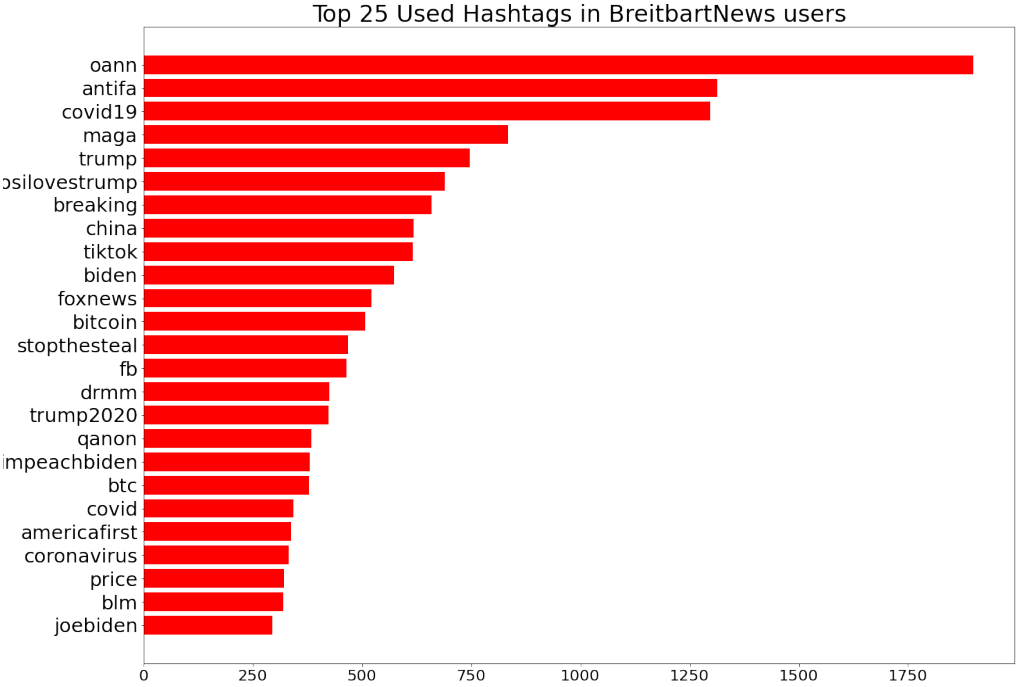
A brief glimpse into the figures below shows that there is indeed a noticeable difference in hashtag usage between users of each news outlet, more specifically with news outlets of differing political alignments.. We found that politically charged words are the most prevalent separation between each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

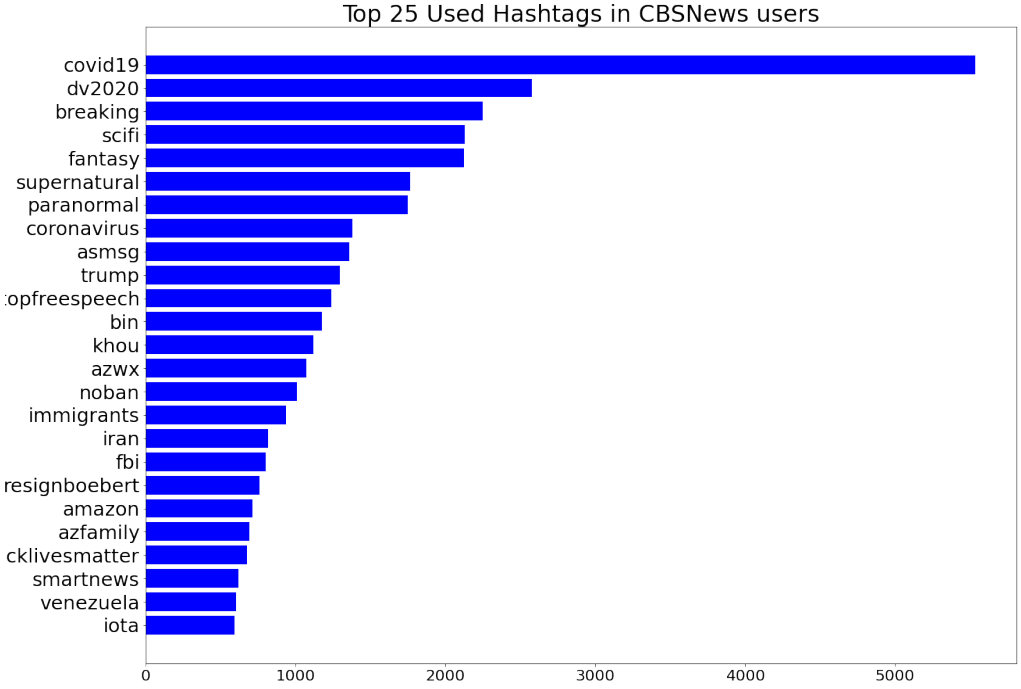
4.2.1 Hashtag Counts of Various News Outlets

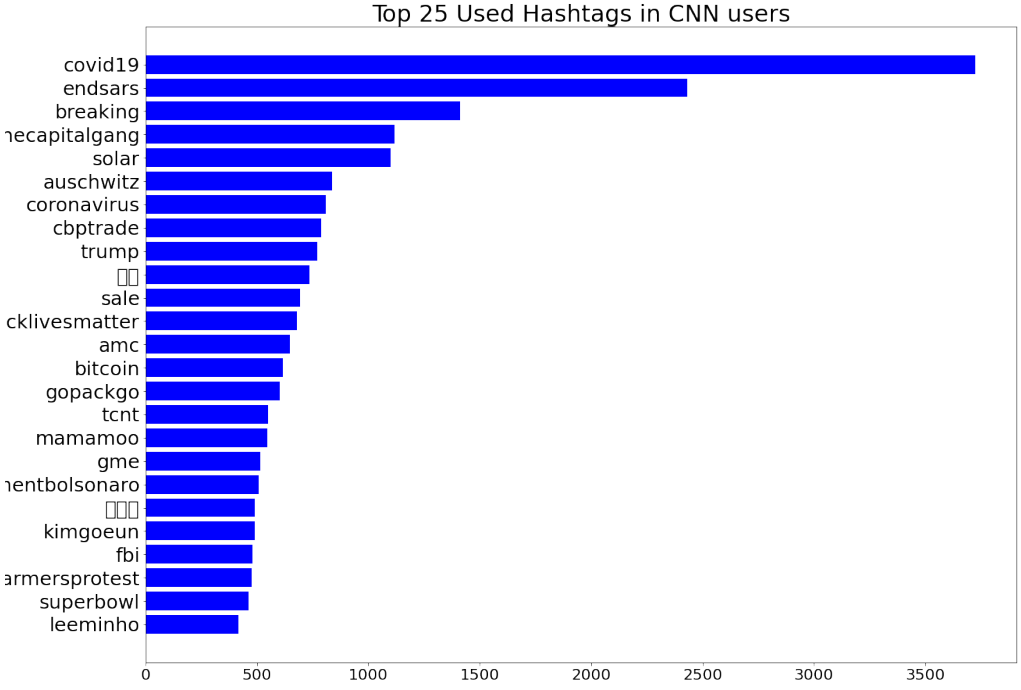
Red == Generally More Conservative

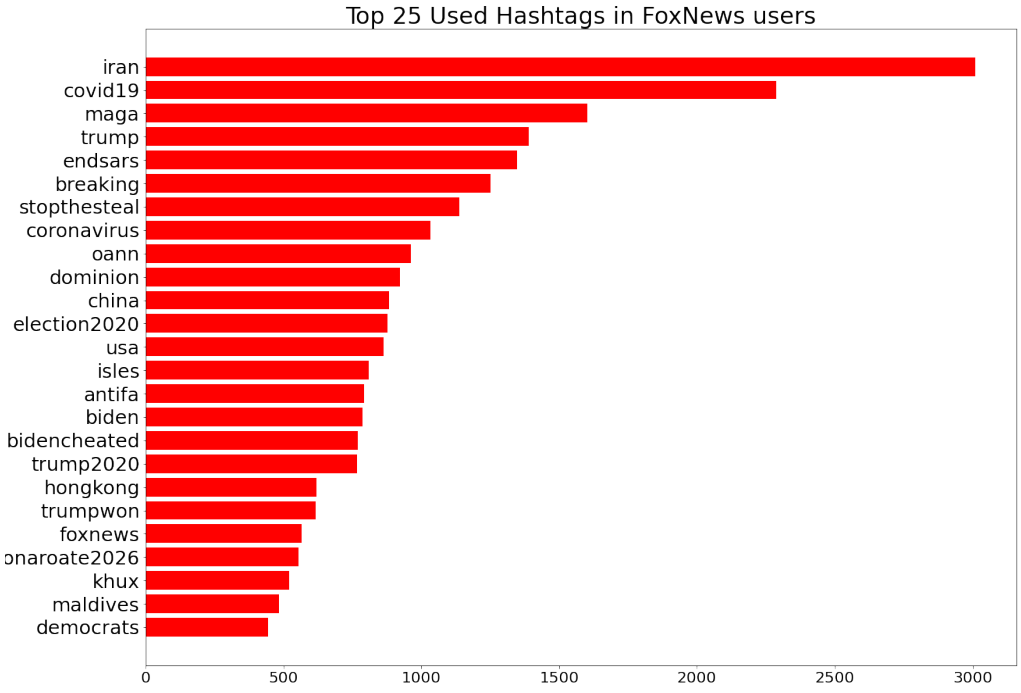


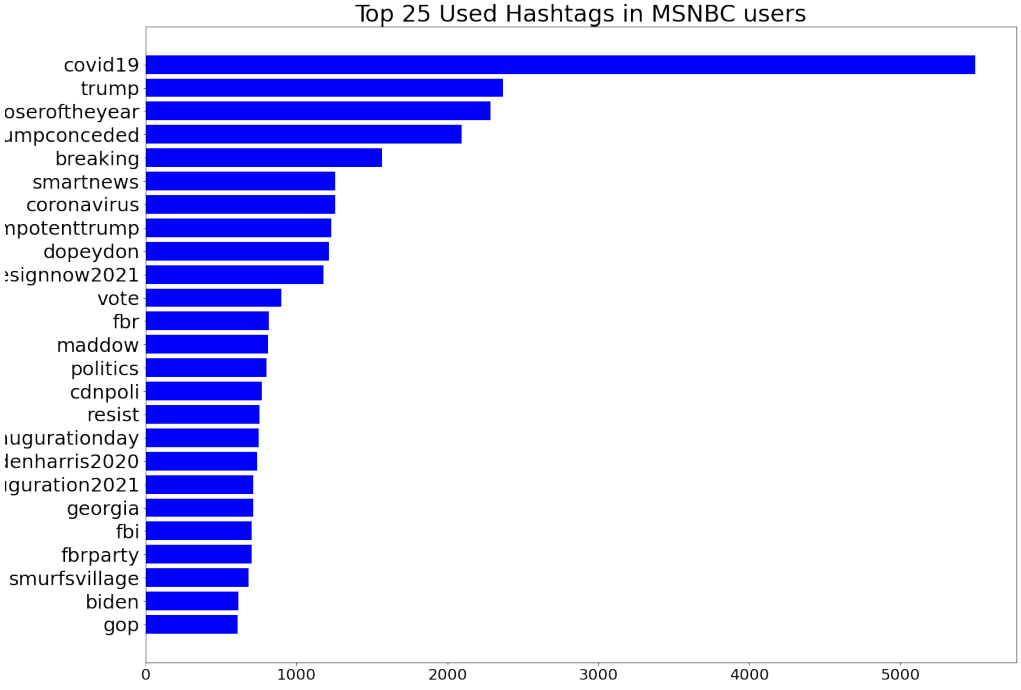
Blue == Generally More Liberal

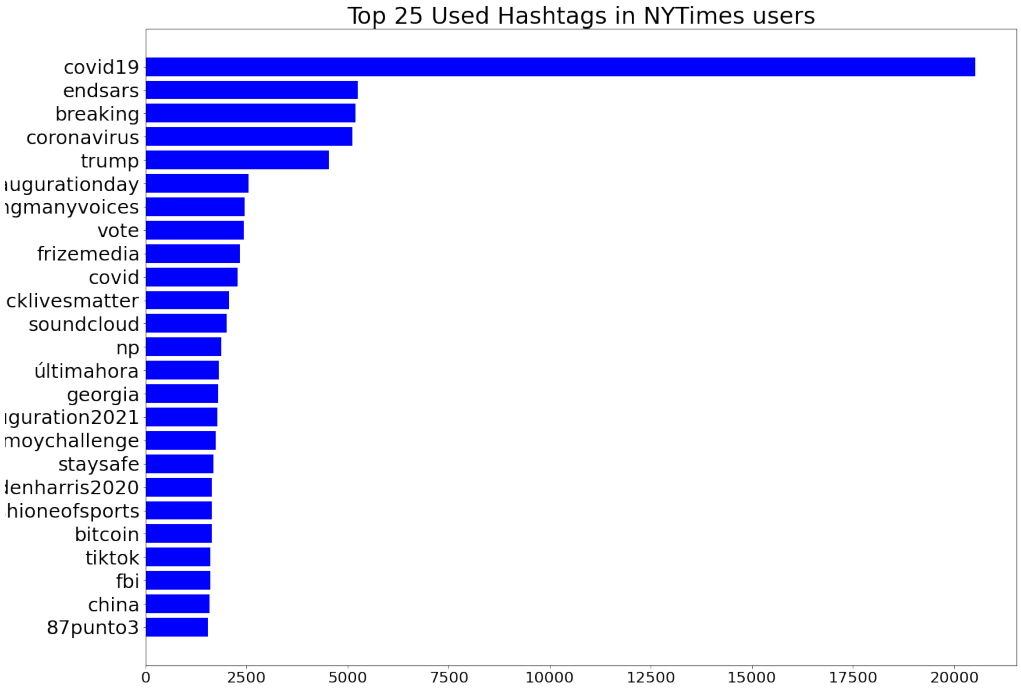


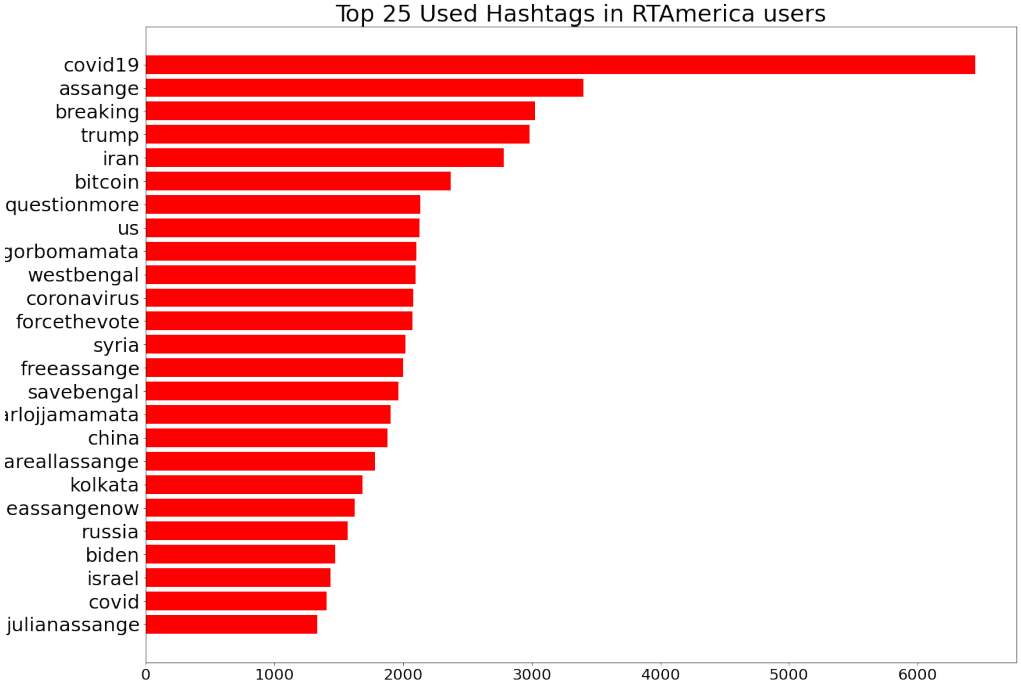

Top 25 Used Hashtags in CBSNews users
covid19
dv2020
breaking
scifi
fantasy
supernatural
paranormal
coronavirus
asmsg
trump
opfreespeech
bin
khou
azwx
noban
immigrants
iran
fbi
resignboebert
amazon
azfamily
cklivesmatter
smartnews
venezuela
iota
0
1000
2000
3000
4000
5000


Top 25 Used Hashtags in CNN users
covid19
endsars
breaking
necapitalgang
solar
auschwitz
coronavirus
cbptrade
trump
□□
sale
cklivesmatter
amc
bitcoin
gopackgo
tcnt
mamamoo
gme
nentbolsonaro
□□□
kimgoeun
fbi
armersprotest
superbowl
leeminho
0
500
1000
1500
2000
2500
3000
3500


Top 25 Used Hashtags in FoxNews users
iran
covid19
maga
trump
endsars
breaking
stopthesteal
coronavirus
oann
dominion
china
election2020
usa
isles
antifa
biden
bidencheated
trump2020
hongkong
trumpwon
foxnews
onaroate2026
khux
maldives
democrats
0
500
1000
1500
2000
2500
3000


Top 25 Used Hashtags in MSNBC users
covid19
trump
oseroftheyear
umpconceded
breaking
smartnews
coronavirus
npotenttrump
dopeydon
esignnow2021
vote
fbr
maddow
politics
cdnpoli
resist
ugurationday
lenharris2020
guration2021
georgia
fbi
fbrparty
smurfsvillage
biden
gop
0
1000
2000
3000
4000
5000


Top 25 Used Hashtags in NYTimes users
covid19
endsars
breaking
coronavirus
trump
ugurationday
gmanyvoices
vote
frizemedia
covid
cklivesmatter
soundcloud
np
últimahora
georgia
guration2021
moychallenge
staysafe
lenharris2020
hioneofsports
bitcoin
tiktok
fbi
china
87punto3
0
2500
5000
7500
10000
12500
15000
17500
20000


Top 25 Used Hashtags in RTAmerica users
covid19
assange
breaking
trump
iran
bitcoin
questionmore
us
gorbomamata
westbengal
coronavirus
forcethevote
syria
freeassange
savebengal
rlojjamamata
china
areallassange
kolkata
eassangenow
russia
biden
israel
covid
julianassange
0
1000
2000
3000
4000
5000
6000

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

We are adopting an unsupervised approach towards quantifying the term political spectrum. In short, 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 some similarity measure between every pair of news stations - and maps the graph onto the euclidean space through graph embedding. The resultant plot - of the nodes lying in the euclidean space (1-D or otherwise) in a fashion relative to their pairwise similarity - and the analysis of which would be the main answer to our research question.

The question then largely boils down to the definition of similarity between news stations. We formally define the concept of similarity between two news stations to be the

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

where

$$X_{1i}$$

and

$$X_{2i}$$

are vectors of hashtag occurrences constructed from the timeline of users who recently retweeted news from the corresponding news station. To make the hashtags political in nature, the hashtag vectors are all subsampled under the same feature space as that obtained from the election dataset. In other words, we record every hashtag that occurred in the election dataset, and count the total

occurrences of these hashtags in the timelines of users that interacted with each news station. For every pair of hashtag vectors constructed in this manner, where every element corresponds to the occurrence of a hashtag in a fixed hashtag space, the similarity is calculated according to the above formulation and the resultant value is assigned as the weight to the edges among nodes.

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 the 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.

The advantages is that the definition of political spectrum is free of heuristics and thus bias. Unlike the project replication done last quarter, we do not preconceptually determine what hashtags implies what semantic meaning under the political context but rather let the occurrence and absence of vectors capture what it means to be politically similar by themselves. Additionally, the method of quantification transforms the similarity in ideology between two sets of tweets in a bounded manner (between 0 and 1) and does not necessitate skewing in the distribution of values.

On the other hand, some caveats are equally worth noting. The first of which is that an extremely popular hashtag could skew the results. Trivializing the difference in pairwise similarity among pairs of news stations. Secondly, though we are not introducing biases in the procedure in methods, the formulation of the methods rides on one crucial presupposition - that tweets with hashtags are sufficiently representative of tweets in general, with or without hashtags. Last but not least, due to the nuance in semantics in human language in general, some hashtags were brought up to promote a point while others were points brought up to be criticized. In other words, we are assuming, amidst the hodgepodge form of interactions on twitter among users, that the hashtag space is sufficient for representing a user's, and subsequently the news station's, political stance.

To account for these considerations, we will construct two (or potentially three) types of hashtag vectors. The first is the raw count of occurrences of every hashtag in the feature space. The second is the normalized vector (every element between 0 and 1 and sums up to 1) of the occurrences. The potential third vector would be occurrences constructed from only original tweets, precluding retweets. These three types of vector would hopefully help us understand and reduce the effect of skewing, and complexity in tweet interactions.

7 Results

8 Conclusions

[ ]: