COVID 19 TWITTER MESSAGES ANALYSIS: A COMPREHENSIVE APPROACH

KaiFu Ren
New York University
Center for Urban Science and Progress
kr2516@nyu.edu
NetID: kr2516

Eric Zhuang
New York University
Center for Urban Science and Progress
eric.zhuang@nyu.edu
NetID: yz2936

May 7th, 2020

Final Project Paper for Text As Data

Word Count: 2688

ABSTRACT

The world has been struck by the CoronaVirus. As the awareness of this crisis increased drastically, the social media platform became a hodgepodge of mixed opinions and emotions. Twitter, one of the biggest and most used social media platforms, was showing an unprecedented COVID19 related topic growth in a short period. This research demonstrates a comprehensive and clear analysis of tweets related to the disease posted from March 04 to April 15 within the U.S. Three approaches of analysis were conducted for the exhaustiveness of the analysis: a general exploratory representation, a supervised learning approach and an unsupervised learning approach. March 19, the date the U.S. government officially announced the highest level national travel ban, was selected as a breakpoint for the class of dataset: pre travel ban and post travel ban, which is used as a baseline comparison for population sentiment and topic difference. The results show that there is a significant difference in topics and sentiment change before and after the U.S. travel ban. We concluded from our research that major events during global crises do affect people's sentiment and their focus on topics. Moreover, news and media tend to have an incendiary effect on public emotions during special times.

Keywords: supervised, unsupervised, coronavirus, twitter, sentiment, major events

1. INTRODUCTION

COVID19 was first originated from Wuhan, China, in early 2020. The disease has collectively affected the entire world from many social and economic aspects. As the infected cases skyrocketed for the last four months, many countries proposed national emergency measures. Covid19 and the chain effect it created, will also alter people's mental health status. Currently, there are estimated over 2.6 billion individuals are under a lockdown or quarantine condition. (WeForum) One research from the Lanchet indicates that quarantine can cause stressors both during and post quarantine including post-traumatic stress symptoms, avoidance behaviours, and anger. (The Lancet, 2020) In addition to the quarantine, news and media also bring side effects to the people's mental status. Consuming too much COVID19 related news, either positively or negatively, will impact people's sentiment. (verywell mind) Therefore, understanding people's emotions over the timeline of a pandemic could prevent collective panic and anxiety. Media companies and health organizations can send messages accordingly based on people's reactions to certain events during the global crisis. At this stage, the U.S. has the highest confirmed cases and deaths among all other countries in the world and is doing everything it possibly can to combat the disease. This research will be performed mainly with tweets posted within the U.S.

2. LITERATURE REVIEW

A group of researchers from the University of Texas Southwestern Medical Center applied the sentiment analysis approach using twitter data from January 14 to January 28, 2020, to discover how people react to COVID19. The results demonstrate about 49.5% of the tweets expressed fear and nearly 30% expressed surprise emotion. Moreover, during this half of months,

people are more regularly discussing topics related to the economic and political impact of the COVID19 rather than public health risk and prevention methodologies. (Medford, 2020) Dr. Dubey from Jaipuria Institute of Management published research about twitter sentiment analysis for 12 countries using tweets collected from March 01 to March 31. The study intends to understand how citizens from different countries are dealing with the situation and the results show that most of the countries are taking positive and hopeful approaches to Covid19. (Dubey, 2020) However, no significant events were taken into account to observe the change of sentiment. From those two research, it is reasonable to infer that people tend to panic during the early phase of the spread of the disease and gently calm down as time passes.

3. THEORY AND HYPOTHESIS

We assume people's sentiments and topics discussed tend to be significantly distinguishable before and after the travel ban. Thus we believe the differences in their tweets can be used to demonstrate the impacts of government measures and the changes in people's mental status.

Hypothesis 1: We assume people's sentiment fluctuates as time goes by. Sentiments are expected to get more negative as the situation gets worse and when the government announces emergency measures.

Hypothesis 2: As a supervised learning problem, we assume that the model can successfully predict the class of tweets for two time periods.

Hypothesis 3: As a clustering problem, we assume the topics being discussed are different for pre and post travel ban.

4. DATA

We selected our dataset "COVID 19 tweets" from the open-source website Kaggle. This data contains features such as User ID, the time, the content, whether is a retweet or a quote, the country code and the language. Tweets are collected using hashtags including #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19, #epitwitter, #ihavecorona from March 1st to Apr 15th, 2020. In total, we collected 237,770 tweets and Figure 1 shows the distribution by day.

5. METHODS

The research consists of four parts, as shown in Figure 4.

Data Preprocessing

There are two parts of data preprocessing involved in this research. The first part of the process is the general data cleaning which includes:

- Remove English stop words, numbers, emojis, links, punctuations, messy codes existed in Tweets
- Convert all words to lowercase and stem the words
- Remove all pound signs in the data, then remove terms directly related to COVID-19 since the data was collected via those terms. (The reason we do not eliminate hashtags is that some of the hashtag words could be used to decipher sentiment and topics. Exp: #lockdown, #toiletpaper, #love)
- Set up the break-up timeline on March 19 and separate the data into two classes and then perform sentiment analysis, supervised and unsupervised learning on them

The second part of preprocessing will be discussed separately in the result part for each supervised and unsupervised learning since the data need to be manipulated differently.

Sentiment Analysis

Since human psychology is far more complicated than just a dichotomous measurement of positive or negative, we are trying to capture the underlying emotions of human beings during this particular period of time. In Mohammad and Turney's research paper, they examined Plutchik's eight basic emotion models, including sadness, joy, surprise, anticipation, fear, trust, anger, and disgust. (Mohammad) The NRC emotion Lexicon containing over 14000 lexicon terms was created for capturing the emotion other than just positiveness and negativeness. Figure 3 demonstrated the steps from data collection to NRC sentiment analysis.

Supervised Learning

For this part, we plan to use three machine learning models to perform the classification tasks to predict the class of tweets: Naive Bayes, Support Vector Machine and Random Forest. The accuracy of each model's performance on the test dataset will be used as a performance indicator. Then the best-performed model will be selected.

Unsupervised Learning

Latent Dirichlet Allocation (LDA), Structural Topic Model (STM) and WordFish will be used to discover how the topics change over the course. We will pay extra attention to the weighted terms for each topic and the proportion of topic distribution before and after the travel ban being announced.

7. RESULTS

7.1 Word Cloud

Word clouds are generated with a minimum word frequency of 70 in both pre and post travel ban corpus, as shown in figure 4 & figure 5. During the pre travel ban social media text generation, words like pandemic, quarantine, hands, close, working, emergency, Donal Trump were frequently mentioned. It is reasonable to guess people were discussing school closures, food buying spree, Covid19 prevention strategies, government policy, etc. The distinction is clear for post travel ban tweets. Besides the topic of coronavirus itself, crowd attitude is driven towards a positive and hopeful tone as words like support, care, love, life, stay home, together, friends, family, thank, etc. start to appear more often.

7.2 Sentiment Distribution

Figure 6 shows negative and positive emotions are dominant among the tweets followed by trust, fear, sadness, anticipation, anger, joy, disgust, and surprise. It seems like people are more likely to possess conflicting emotions during COVID19. Even though there are only small differences, the graph still captures a slightly differentiated emotion from pre and post travel ban tweets. The population shows a bit more anger after the national travel ban, but at the same time, negative and fear emotions also decreased. A reasonable guess is that people believe the travel ban will reduce the spread of the virus.

7.3 Sentiment time series

The sentiment time series analysis was conducted daily during March-04 and April-15. The original dataset was aggregated by day, converted into a document feature matrix and then the sentiment score was computed using the positive and negative dictionary. Table1 shows a

summary of people's emotions and the distinction is quite clear. Since the beginning of March, people's sentiment towards COVID19 has changed pretty drastically, as shown in figure 7. Negative emotions were captured in the earlier phase of coronavirus outbreak and decreased gradually as time went by. U.S. officials announce the suggestions of banning gathering events with more than 10 people following the avoidance of public restaurant and bar gathering on March 16. From the chart, this date is the first time overall positiveness overruns the negativeness and then entered into a relatively balanced status. People's emotions seemed to fluctuate based on news and events. A brief conclusion can be derived from the chart: our initial hypothesis of human emotion tends to become more negative as the situation continues is overturned. People's emotions towards COVID19 have been pacified over time and they tend to fluctuate based on specific news.

7.4 Supervised learning

Data Preprocessing

We have selected date and text as our reference columns and created a binary column with 0 if the post date is prior to march 19th, 1 if the post date is after. Then we apply the random sample to all the datasets to mix up the rows for further analysis. Since there are over 22k tweets in the dataset, which will overload the system if all tweets are used as inputs, only the first 5000 tweets were converted into Document-feature matrix and selected for modeling.

Model Results Matrix

The accuracy score tells that radial kernel SVM and Random Forest has the best performance in predicting the class of tweets (whether a tweet is posted before or after the travel ban measurement). Radial Kernel and random forest both return an extremely high number of

close to 99% truly relevant results, but only around 70% was accurately predicted. After all, we achieved a 72% accuracy with random forest. Although this is not a perfect number, it certainly indicates that some distinct topics are being discussed for those two time periods.

Feature importance

Mean decrease Gini in Figure 8 summarizes the most important features that contribute to the prediction of tweets class from the Random Forest classifier. Obviously, after the U.S. level 4 travel ban, many flights got canceled and suspended, which makes related words often discussed. The mean decrease accuracy expounds on how much accuracy of the prediction will be decreased if the listed features are removed from the tweets. Interestingly, toilet paper seems to be very important in terms of its predictive power.

7.5 Unsupervised Learning

Applying unsupervised techniques over the twitter datasets can discover the hidden dimensions of groups of words that are likely to occur together. Topics can be summarized by analyzing those groups of words.

Data Preprocessing

Tweets are concatenated by the date of creation for topic modeling. Thus, there are 43 documents as a corpus. Then we create the Document-feature matrix. For STM and LDA models, words that occur fewer than 20 times or appear in fewer than ten documents are removed. This is because by default, topic modeling algorithms assume words from each document are randomly drawn from a randomly sampled distribution of topics. Those rare terms will not be captured by the topic selection, which will lead to inaccurate topic assignments.

Wordfish

Firstly, we assume that tweets grouped by pre and post travel ban can be represented by a single hidden dimension. We use Wordfish, the Poisson scaling model of unsupervised one-dimensional document positions (Slapin, 2008) and set index text as March 12 for pre-closure and March 20 for post-closure representations. The result of Wordfish (Figure 9) presents a unique curve among tweets made in pre and post closure periods. Pre-closure tweets all stay on the left side with theta below zero, while post-closure ones are the opposite. From the box plot (Figure 10), pre-closure tweets are more diverse concerning the hidden dimension. The guitar plots (Figure 11) with Y-axis representing the word frequency and X-axis representing the word weight show that tweets made in the post-closure period have more distinctive words. The top 20 word weights (Table 3) that place each document differently show that in the pre-closure period, people discussed lockdown USA, wearing masks, politics, exercise at home, food and education. However, in the post-closure period, more uniform words were concentrated on staying home, quarantine, social distancing, health, family.

STM

Discussions in tweets tend to be in clusters and groups. The Structural Topic Model (STM) is a framework of topic modeling with document-level covariate information (Margaret), which in our case, is the date information for topical prevalence to affect the frequency with which a topic has been discussed. The topic number algorithms of Lee and Mimno (2014) identified 106 topics; however, most of them cannot be distinguishable with given top frequency words. We decided to use 25 topics to reduce redundancy, but still, most of the topics in Figure 12 contain the same words as "will," "get," "people." The topical prevalence contrast (Figure 13)

can further classify documents. Topic 20 and 17 have high distribution proportions in the post-closure period, while topic 18 has been frequently discussed in the pre-closure period. The others remain neutral. A closer look at highest probability and FREX words between topic 18 and 20 (Figure 14 and 15) can conclude that before travel ban people are discussing work, friends and situations in foreign countries like Italy, while after travel ban home, quarantine life, city lockdown and effects on workers have been focused.

LDA

With the limitations in STM, we decide to use the Latent Dirichlet Allocation model (LDA), the well-known unsupervised statistical topic model that can extract topics from a collection of documents (David, 2003). The matrix in calculating the optimal number of topics for LDA shows the range of 15-30 (Figure 16). In alignment with STM, we chose 25 topics. In order to track the change of topic over time, we plotted the top 2 topics per day based on the weight of each topic for each document. Initially, the top 2 topics per day were occupied by topic 1 and 23 which tend to be general descriptions about people and life (Figure 17). We decided to drop these two topics for further exploration. The result (Figure 18) demonstrates a clear pattern that the first topic changes from early March in topic 10 (previous 11), to March 19 the announcement of travel ban in topic 18 (previous 19), and then the start of April in topic 6 (previous 7). Meanwhile, there is a clear distinction among the contribution of those three topics over the two classes (Table 4). Starting from early March to the period after the travel ban, people's opinions about COVID-19 changed from "spreading flu and "accusing politicians" towards "staying at home," "quarantine and social distancing," "wearing masks."

8. DISCUSSION

The sentiment analysis of tweets from March 04 to April 15 shows that people's negative emotion has been gradually alleviated. This finding corresponds to the two research mentioned in the literature review. Besides the understanding of general sentiment change, we also want to comprehend whether major events deviate people's attention, namely how the topic people have been discussed change. The findings support our hypothesis that news, especially those related to government measures, shift people's attention. For future work, incorporating news events on tweets with an extended timeline may demonstrate a more definite pattern of the sentiment change and topic difference. In such a way, news and media companies may release news based on public sentiment to avoid collective panic.

Citation

1. Samantha K Brooks, Rebecca K Webster, Louise E Smith, Lisa Woodland, Simon Wessely, Neil Greenberg, Gideon James Rubin, The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020; 395: 912–20
2. Richard J. Medford, Sameh N. Saleh, Andrew Sumarsono, Trish M. Perl, Christoph U. Lehmann, 2020 .An "Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak
3. Akash D Dubey, 2020. Twitter Sentiment Analysis during COVID19 Outbreak,. 1-9
4. Hoof, Elke Van, and Vrije Universiteit Brussel. "Lockdown Is the World's Biggest Psychological Experiment - and We Will Pay the Price." *World Economic Forum*, www.weforum.org/agenda/2020/04/this-is-the-psychological-side-of-the-covid-19-pandemic-that-were-ignoring/.
5. Lindberg, Sara. "Is Watching the News Bad for Mental Health?" *Verywell Mind*, Verywell Mind, 14 Apr. 2020, www.verywellmind.com/is-watching-the-news-bad-for-mental-health-4802320.
6. Le T. Nguyen, Pang Wu, William Chan, Wei Peng, Ying Zhang, 2020. Predicting Collective Sentiment Dynamics from Time-series Social Media
7. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Processing, pages 1–6.
8. Karl-Michael Schneider, 2005, Techniques for Improving the Performance of Naive Bayes for Text Classification. LNCS, volume 3406
9. Jonathan B. Slapin, Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. American Journal of Political Science 52, 705–722 Wiley, 2008
10. Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley. stm: R Package for Structural Topic Models. Journal of Statistical Software.
11. Lee M, Mimno D (2014). "Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1319–1328. Association for Computational Linguistics, Doha, Qatar.
12. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022

Appendix A (results graphs):

date	Counts
03-04	129
03-05	246
03-06	133
03-07	95
03-08	132
03-09	648
03-10	1960
03-11	3305
03-12	8049
03-13	14211
03-14	6191

date	Counts
03-15	6874
03-16	8364
03-17	11035
03-18	8050
03-19	10337
03-20	9594
03-21	9512
03-22	8777
03-23	7643
03-24	7364
03-25	7372

date	Counts
03-26	8962
03-27	7420
03-28	6847
03-29	6231
03-30	5623
03-31	6890
04-01	6219
04-02	5619
04-03	5752
04-04	5572

date	Counts
04-05	4823
04-06	5453
04-07	4108
04-08	4939
04-09	4002
04-10	3841
04-11	3158
04-12	2245
04-13	1841
04-14	4000
04-15	4199

Figure 1. Number of Tweets by Day

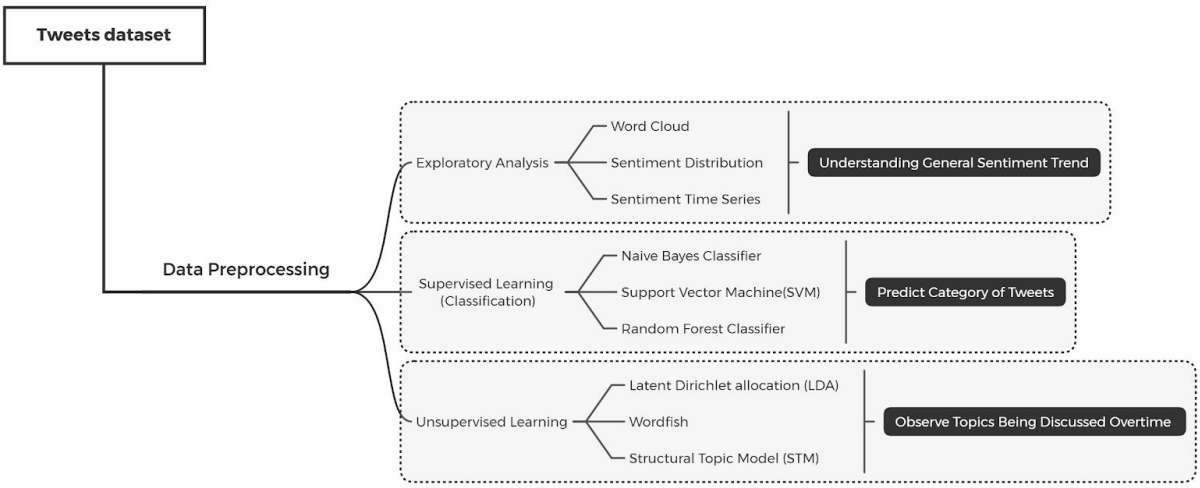

Figure 2. Summary for Research Methods

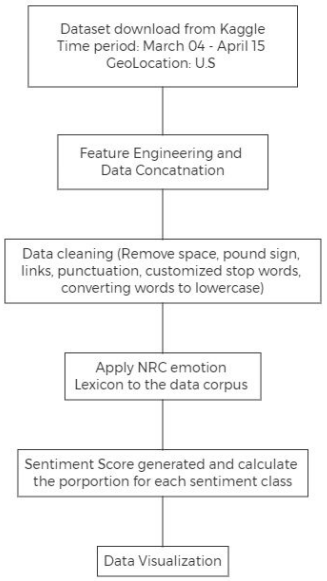

Figure 3. Steps for NRC Sentiment Analysis

Figure 4: Pre Travel Ban Word Frequency

Figure 5: Post Travel Ban Word Frequency

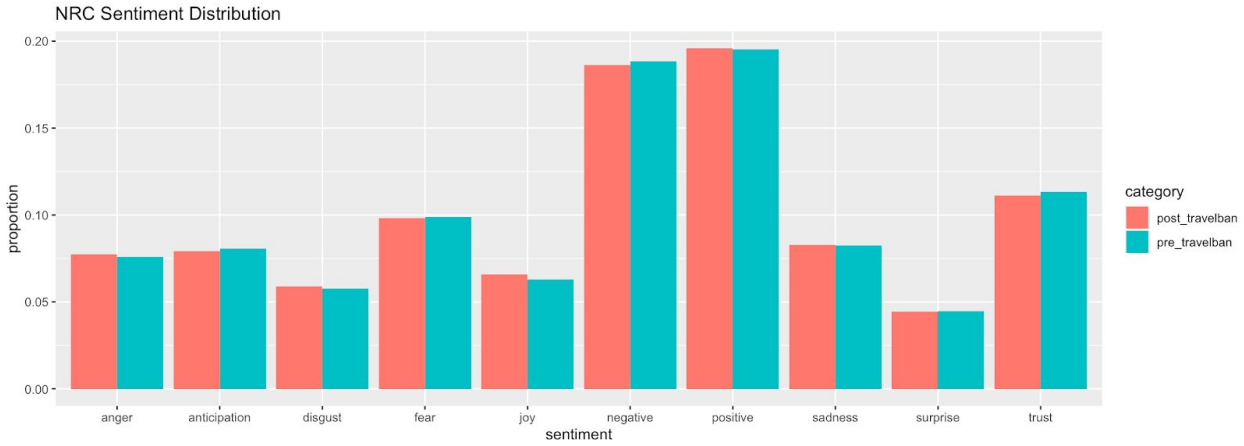

Figure 6: NRC sentiment distribution

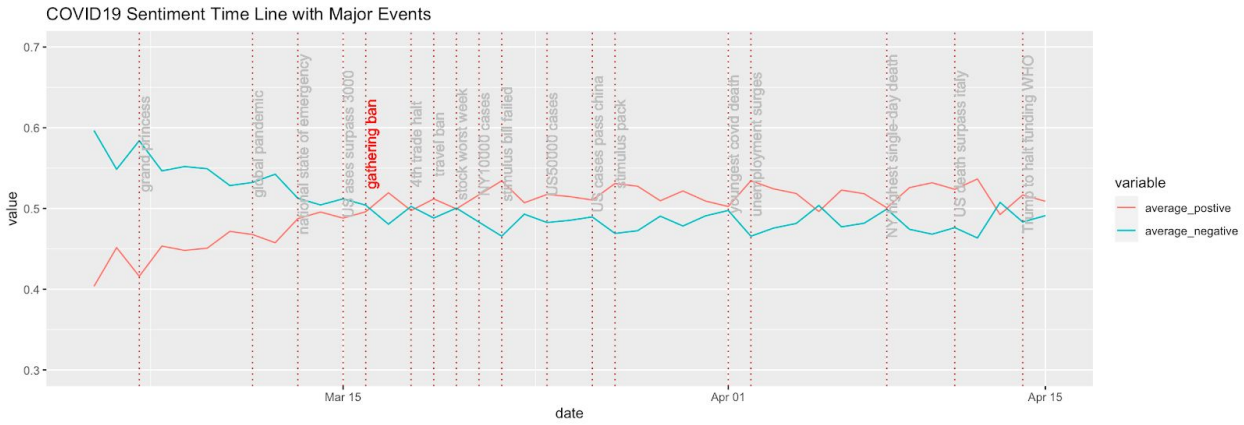

Figure 7: Sentiment Time Series

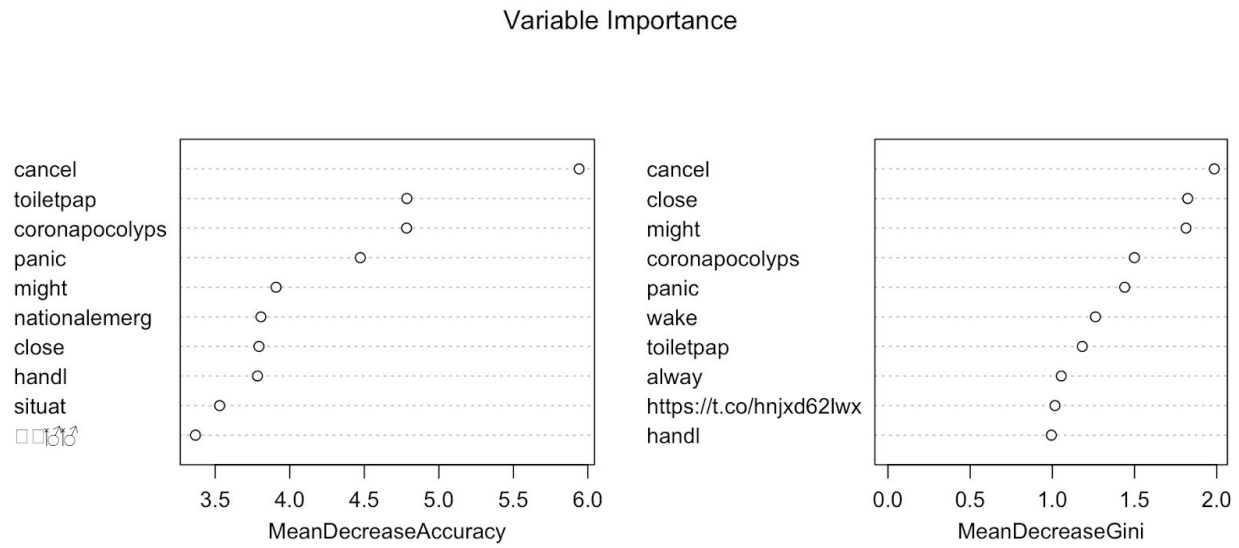

Figure 8: Feature importance

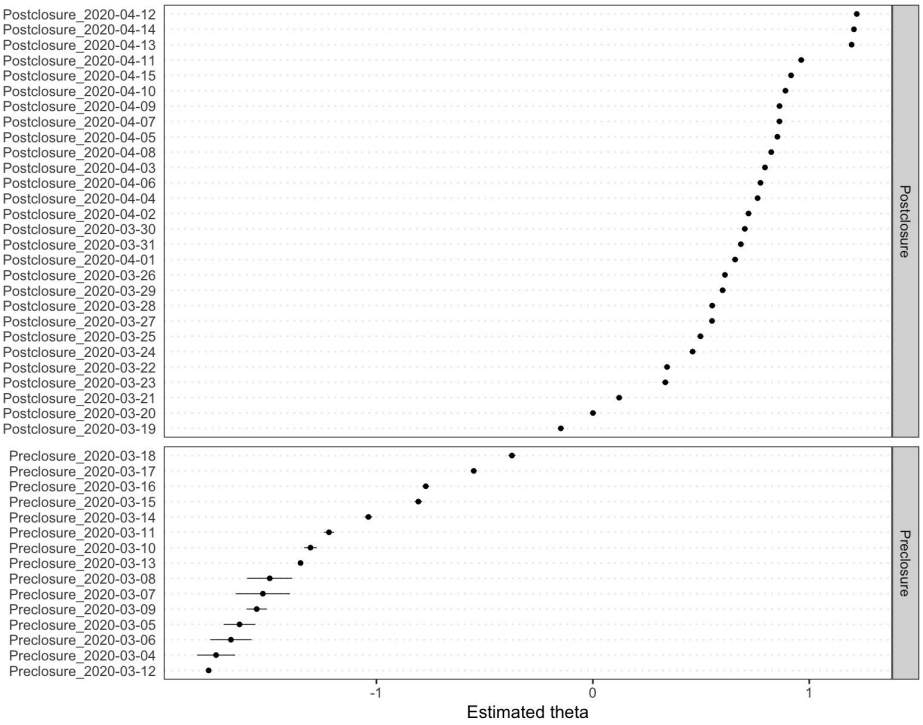

Figure 9: Wordfish Estimated Document Position

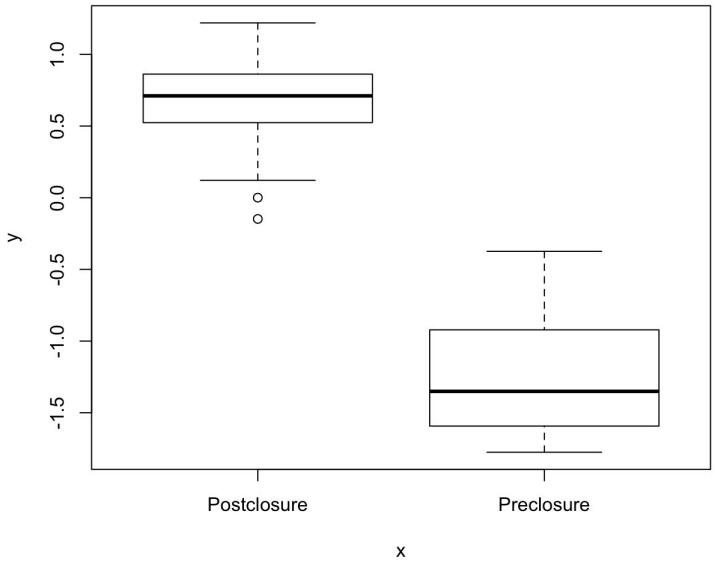

Figure 10: Wordfish Estimated Document Position Box Plot

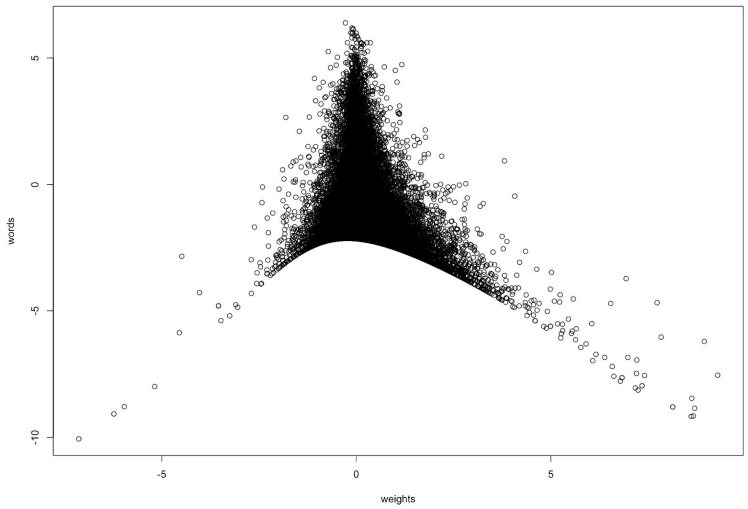

Figure 11: Wordfish Word Weight Guitar Plot

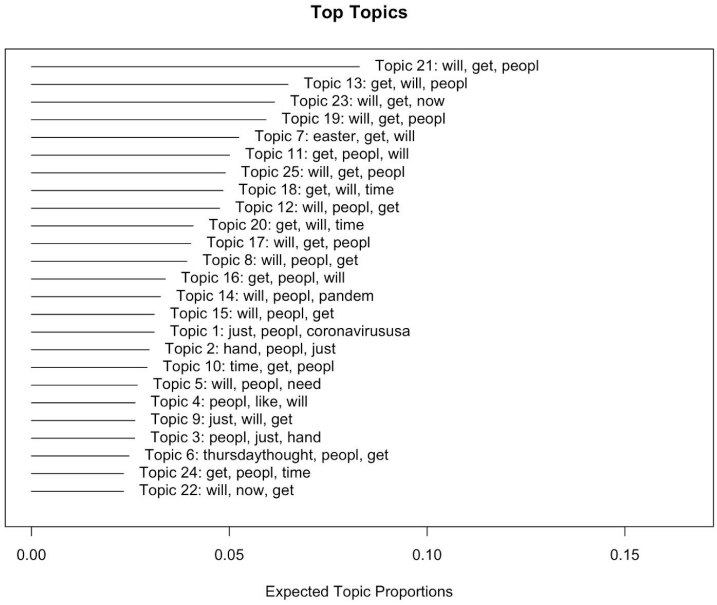

Figure 12. Topic Proportions from STM

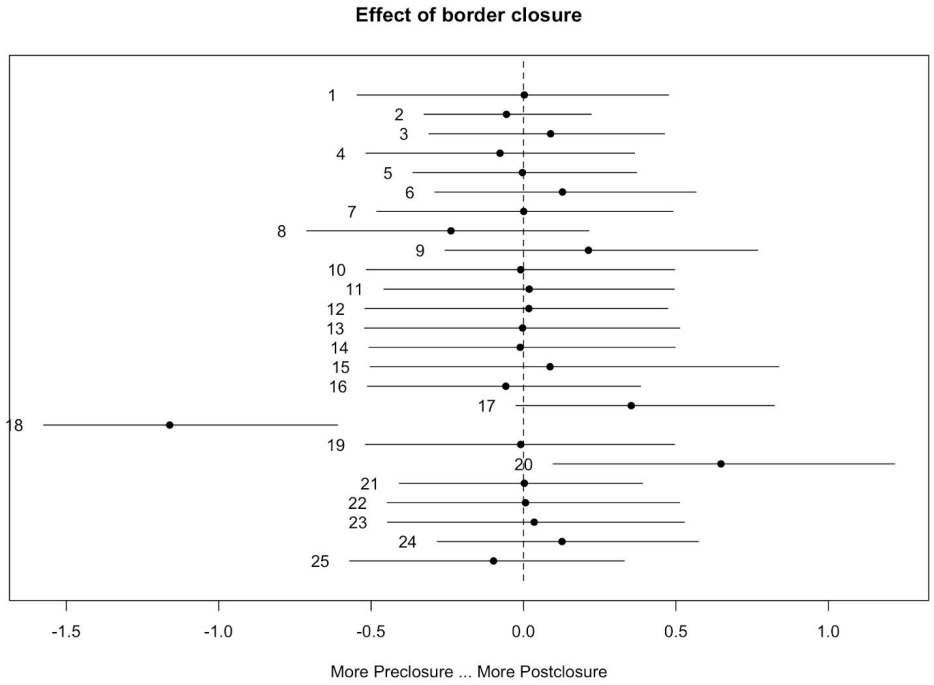

Figure 13. Topical Prevalence Contrast

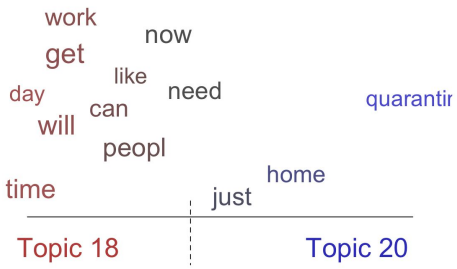

Figure 14. Topical Contrast Between Topics 18 and 20

```
Topic 18 Top Words:
         Highest Prob: time, get, quarantin, peopl, now, will, friend
         FREX: quaratinelif, stayathom, con, voi, siamo, italian, cari
         Lift: rand, quaratinelif, westandwithitali, voi, siamo, quarantineact, cari
         Score: quaratinelif, voi, siamo, cari, westandwithitali, stayathom, covidiot
Topic 20 Top Words:
         Highest Prob: wwgwga, get, will, peopl, just, now, time
         FREX: wwgwga, thank, help, new, worker, american, lockdown
         Lift: wwgwga, prove, eventu, model, modern, gas, dish
         Score: wwgwga, stayhom, stayathom, quarantinelif, socialdistanc, shelterinplac, lockdown
```

Figure 15. Words for Topic 18 and 20

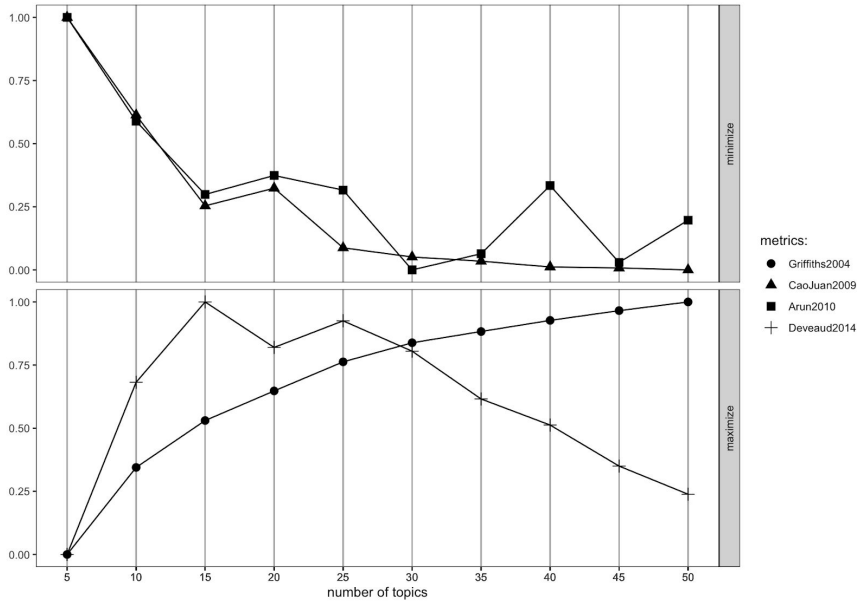

Figure 16. Find Optimal Number of Topics

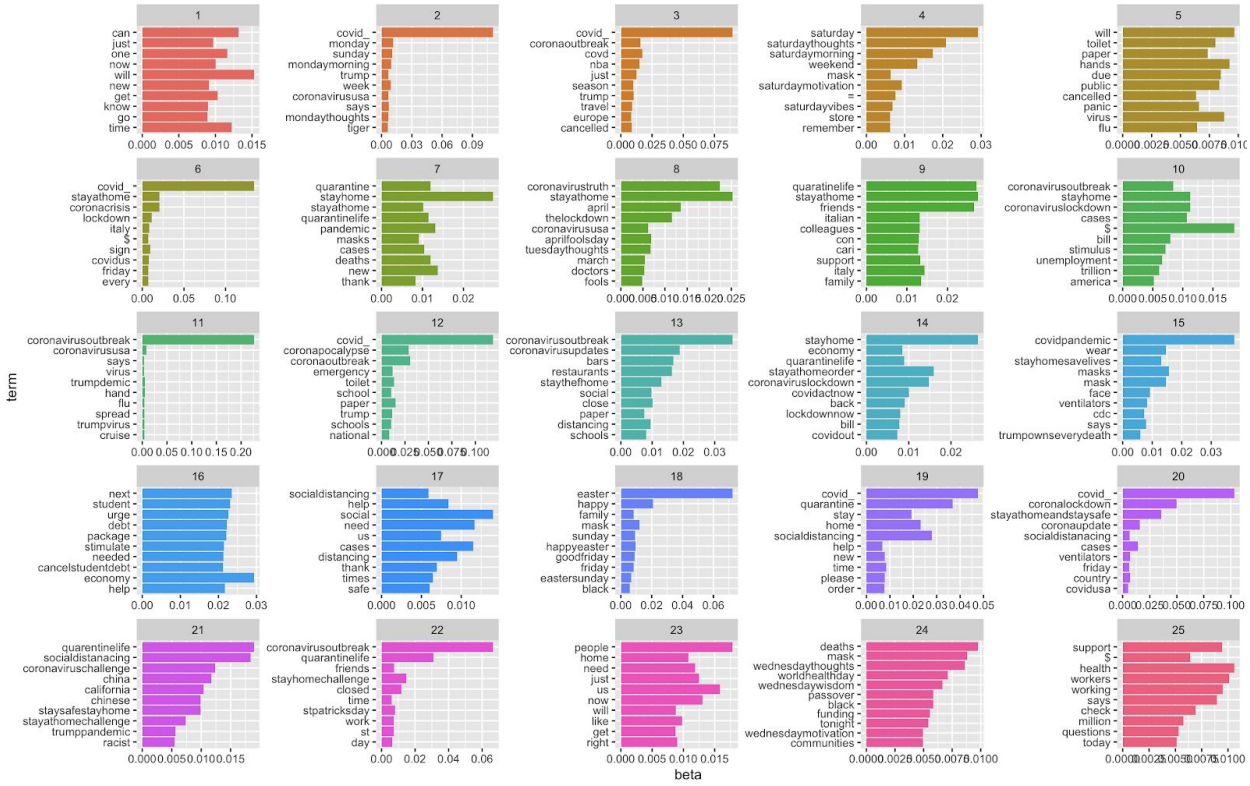

Figure 17. Top 10 words per Topic

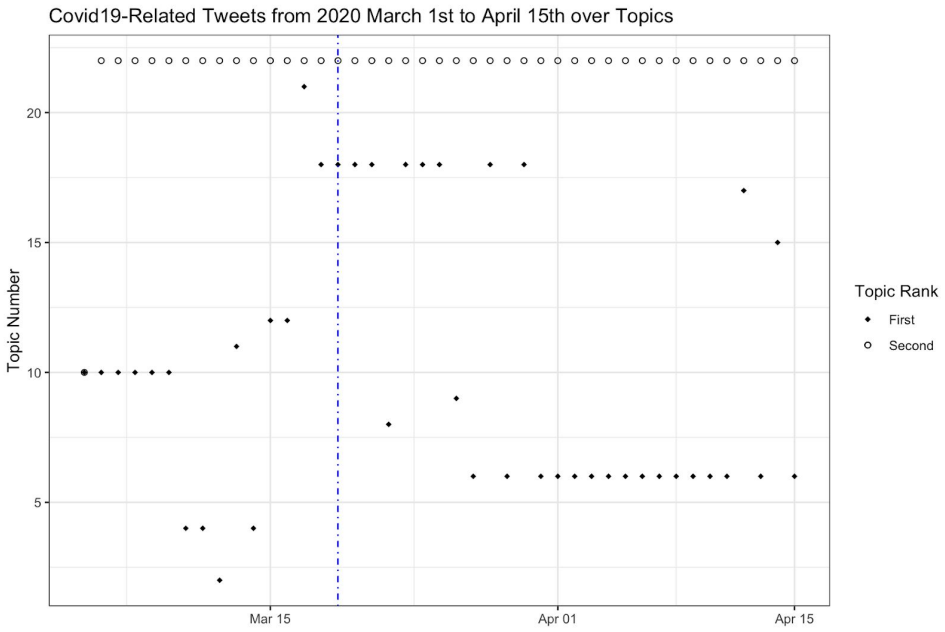

Figure 18. Topic Ranks Per Day

Appendix B (results tables):

	Positiveness	Negativeness
Pre-Travel ban	47%	53%
Post Travel ban	52%	48%

Table 1: sentiment polarity comparison

Model	Naive Bayes	SVM Linear	SVM Radial	Random Forest
Accuracy	0.69	0.66	0.71	0.72
Recall	0.88	0.81	0.99	0.98
Precision	0.73	0.74	0.71	0.74
F1 Score	0.80	0.77	0.82	0.85

Table 2: supervised classification results

Position	Pre-closure	Post-closure
	Breakbtw	Get
	Balletboy	Peopl
	Jesusisal	Time
	wearamaskinpubl	Need
	Liftpenalti	Work
	Snlathom	Home
	Keepamericaclos	Quarantin
	studentdebtstimulus	Test
	prayfortheworldwegotthiswa	Help
	Crossfitt	Stay
	Hoppi	Today
	Weightlossdiet	Know
	firecoachtrump	Case
	Fernandacalfat	Socialdistanc
	Newtorkc	Keep
	Hehasrisen	Health
	Breakingloc	Close
	Breakinglocalnew	State
	Saveusp	Everyon
	savefauci	famili

Table 3: Top 20 Word Weights for Two Positions

	Topic 11	Topic 19	Topic 7
Preclosure	0.094621253	0.02979096	0.00491803
Postclosure	0.004382071	0.09405549	0.11024051

Table 4. The mean contribution of each topic over each period