Understanding the Correlation between COVID-19 Related Tweets and National Policies through Sentiment Analysis

Yue Liu[1] and Senqi Zhang[2]

***Abstract*— During the COVID-19 pandemics, billions of people are impacted during this global crysis. However, in the United States, the unavailability of public data in early 2020 makes it difficult for smaller groups to study and analyze the spread of the virus as well as the mental health of people during isolation. This paper introduces a new way to monitor the pandemics as well as a new approach for the government to understand the public using social media.**

I. INTRODUCTION

The COVID-19 pandemic has rapidly impacted over 200 countries, areas, and territories. On December 8, 2019, several cases of COVID-19 were first reported in Wuhan, China [1]. Just one and a half months later, on January 23, Wuhan, a city of more than 11 million, was locked down to prevent the spread [2] In March, while China starts to see its turning point, COVID-19 is just starting to become a global threat. On March 13, President of the United States, Donald Trump declared a national emergency, and people are encouraged to practice social distancing [2]. As of May 4, in the United States, there are 1.21 Million reported cases and nearly 70 thousand deaths [3].

In a difficult time like this, we would like to know COVID-19 has impacted the sentiment of the public and how sentiment is correlated with National Policies. In light of the deteriorating situation in the United States, the discussions related to “COVID19” and “Donald Trump” on Twitter has increased drastically. Therefore, we choose to analyze the streaming data on Twitter between March 14, 2020, and April 14, 2020, for this study.

In this study, we calculate the average sentiment score for each day by implementing Google NLP API and study the fluctuation of the average sentiment score. Then we related interesting changes to important national policies that are announced on that day. Then to further assure the correlation between the sentiment and the national policies, we use LDA and WorldCloud to testify whether the popular topic on Twitter is indeed related to the national policies.

II. RELATED WORK

Since the beginning of the COVID-19 pandemic, researchers have worked diligently in studying various aspects of fighting COVID-19. Data mining research focuses on studying the impacts and implication, including protection against the virus[5], new treatment method[6][7], real-time tracking of the situation[8], the effectiveness of social distancing[9][10], and spreading of misinformation[12][13]. Research on people’s sentiment has been conducted on the Chinese largest social media Weibo[10]. In our study, we want to focus on how public policies affect people’s sentiment over time and provide insights into new ways of understanding and tracking pandemics.

[1]University of Rochester, Email: yliu165@u.rochester.edu
[2]University of Rochester, Email: szhang71@u.rochester.edu

III. METHODOLOGY

A. Data Collection

All of our data are collected using Python Package Tweepy [14]. From March 14 2020 to April 14 2020, the data is collected every day from 10 a.m to 12 p.m EST. We choose this specific time period because statistics shows Twitter have highest volume of user activity from 10 a.m. to 1p.m[17]. About 200 thousands to 300 thousands tweets in English are collected each day and a total of 7 million tweets are collected during this period. The keyword we user to filter the streaming data are “Coronavirus”, “COVID19”, and “Trump” and all retweets are filtered. Noted that “Trump” is chosen for keyword as President Trump’s usage of Twitter has attracted worldwide attention ever since the beginning of his presidency. It has become somewhat a part of the way Trump Administration connect with publics. We believe adding this keyword will help in analyzing the public sentiments towards the national policies.

B. Data Preprocessing

Streaming tweets contain a lot of information. The first thing we do is disable collecting retweets because retweets are a reoccurring sample, we want to treat every tweet we collect with the same weight in our sentiment analysis. Then we only extract the “text” field for our sentiment analysis. At first, we plan to use other fields such as “geo” and “location” to do a location-based sentiment analysis with more dimensions, but the reality is most people either do not have the location service enabled for Twitter or they can simply customize the location to wherever they like, thus disqualifying the validity of the data. After that, we have a clean JSON file with one dictionary each line that looks like “Text:”, “COVID-19”. However, some sentences still contain extra apostrophe and quotation marks inside the dictionary structure. This causes our algorithm to report bugs and having difficulty working with the data structure. Therefore, we removed any additional apostrophe and quotation marks inside the dictionary. At last, we eliminate the weblinks and Unicode in the text.

C. Topic Analysis

We used two topic analysis method for our study. World Cloud and LDA Analysis. Word Cloud provides a simple count-based human-readable representation of the important topics among the tweets we collected. While LDA gives us more quantitative prove with Intra-class distance and Inter-class distance we can analyze. Note that we intentionally eliminated the keywords "coronavirus" and "Trump" in order to bring more interesting topics to the table.

D. Sentiment Analysis

We use Google NLP API to carry out the sentiment analysis. The sentiment score for a tweet is ranged [-1,1], where -1 is most negative, 1 is most positive and 0 is neutral. Because analyzing tweets is time-consuming and very costly, we randomly choose 20,000 tweets for each day to calculate the mean daily sentiment score.

IV. Results

Starting with Figure 1, Word Cloud result from over 7 million tweets, we are excited to see policy-related-keyword including "White House", "Social Distancing", "Stay Home" and COVID19-related-keywords including "tested positive", "public health", and "COVID19 pandemic". A very basic topic analysis already provided an indication of the correlation between national policies and people's sentiment towards the pandemic.

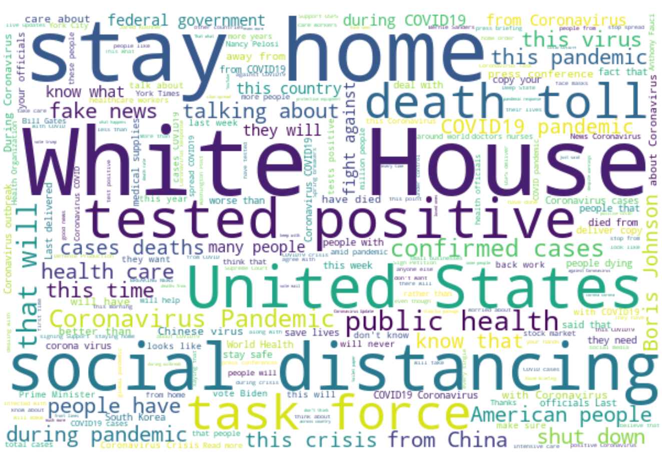

Fig. 1: Word Cloud Analysis of Tweets

Then we plot the daily average sentiment score against the date. The red line in the middle is the average sentiment score of all time. The average is negative, indicating a general negative sentiment during the pandemic. In addition, we can also see an overall downward trend indicating the sentiment is becoming more and more negative. From Figure 3, we also discovered some interesting days where a drastic change occurs, they are marked by blue and red circle respectively in Figure 2 where a blue circle indicates a positive change in the sentiment score and a red circle indicates a negative change in the sentiment score. Let's call the red circles negative peaks and blue circles positive peaks. We relate these dates and national policies announced by the Whitehouse press and

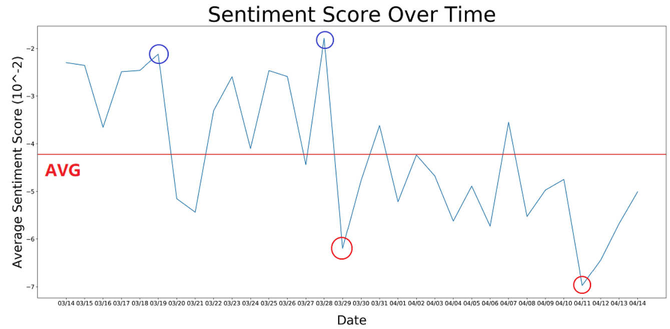

Fig. 2: Sentiment Score Over Time

found some obvious connection:

Negative Peaks:

Negative peaks represent a drastic negative change in public sentiment. As in Figure 1, March 29, and April 11 show a steep drop in the sentiment score. These are some of the headlines on the respective days.

March 29: "Trump announced that the White House would be extending its social distancing guidelines through April 30"[16] The interpretation of the negative peak on March 29 by intuition is that Americans love freedom and socializing but the social distancing guidelines prevent people from participating in any social events. Many annual events of celebration such as ULTRA 2020 is canceled in response to COVID-19. Therefore, people are naturally unhappy about this. By March 29, people have already practiced social distancing for more than a week, seeing the duration of the policy extended until the end of April must be very frustrating. On the other hand, by March 29, 29 of the 50 states in the United States have issued a stay at home order [17]. Workers are eager to go back to work to feed their families and themselves. Therefore, a steep increase in complaints contributed to the lowering of sentiment score.

Running tweets on March 29 through Word Cloud and LDA yield Figure 3 and Figure 4. In Word Cloud, "House Briefing" and "White House" stand out from the rest and they are likely words contained in a tweeter user's response towards the extended social distancing policy.

In LDA analysis, a total number of 5 topics are extracted from the tweets on March 29. In order to visualize it, each circle on the left indicates a topic, the larger the circle is, the more support it has. On the right is a ranking of the most relevant terms for each topic, as well as the weight indicated by the length of the bar. We made it into an interactive HTML file that will be included in the submission, you can hover the mouse above the circle to see the relevant terms and weight for each topic. The LDA model has a coherence value of 0.30 and perplexity of -13.32. The coherence value shows that the topics are not very coherent. Our explanation to this is people usually write in their own perspective on Social Media, subjectivity introduces variations in topics. The perplexity, on the other hand, shows that this model is excellent at predicting samples. From topic 1 we can see relevant terms such as "job" and "help" apart from terms related to "COVID-19" and "Trump", this can be understood by people are trying to go back to work to support their family and seeking help from the government.

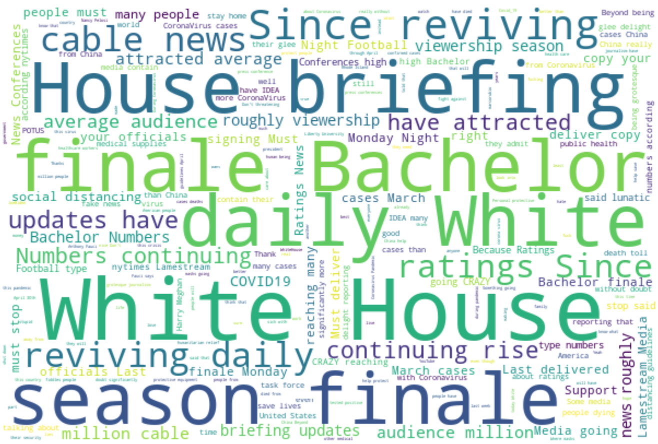

Fig. 3: Word Cloud Result for March 29

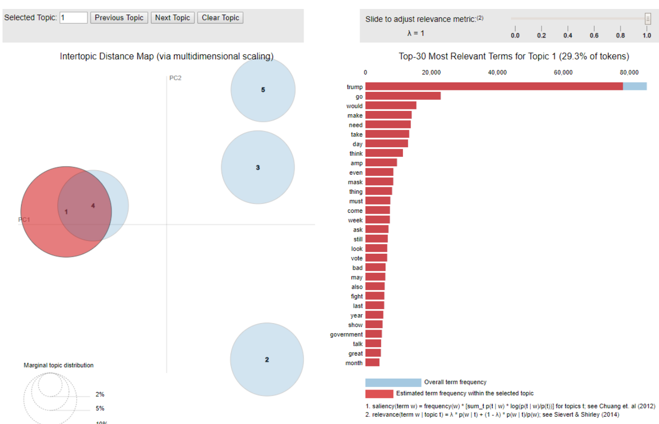

Fig. 4: LDA Result for March 29

April 11: "The United States becomes the country with the most coronavirus deaths globally."[16] The negative peaks on April 11 are very self-explanatory. The United States has become the country with the most number of confirmed cases. This is a very warning to anyone living in the United States as everyone's health is at great risk. The United States is the leading figure in innovation and technologies, one would expect the United States to set an exemplar of handling pandemics for the rest of the world. However, the reality is while Asian countries are starting to recover, United States infection number is increasing exponentially. This leads to doubts and frustration toward the government.

Running tweets on April 11 through Word Cloud and LDA yield Figure 5 and Figure 6. From Word Cloud, keywords such as "death toll," "Fake News", and "White House" catches our attention. Our interpretation is some people are blaming the mishandling of the pandemics to Trump Administration's downplay of the virus in early 2020. President Trump is known for addressing information and news that he does not agree with as "Fake News".

In LDA analysis, a total number of 5 topics are extracted from the tweets on April 11. The LDA model has a coherence value of 0.35 and perplexity of -8.07. From topic 1 we can see keywords such as "lie", "believe", "trump" and "tell". Combining these words, we conclude that most people on April 11 are complaining about how the government did not tell the truth from the beginning and now the pandemic is getting much worse.

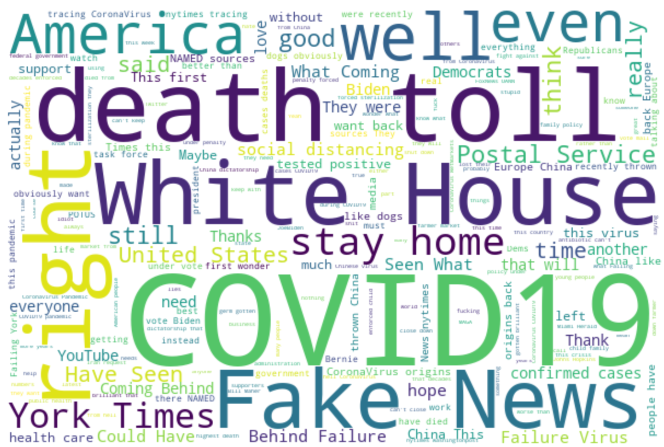

Fig. 5: Word Cloud Result for April 11

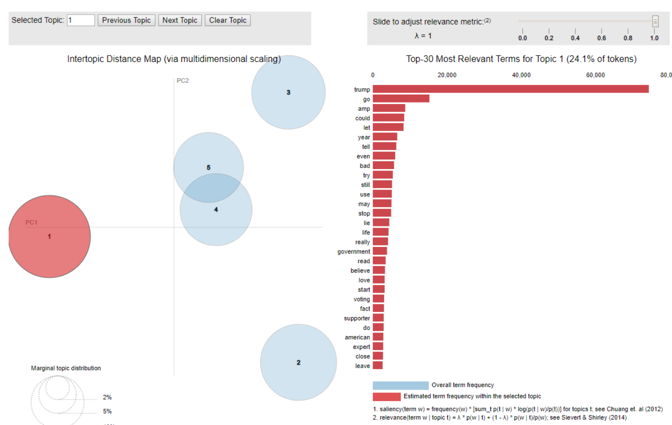

Fig. 6: LDA Result for April 11

Positive Peaks:

Positive peaks represent a drastic positive change in public sentiment. From Figure 1, March 19, and March 28 show a noticeable increase in the sentiment score. These are some of the headlines on the respective days.

March 19: "President Donald J. Trump is signing a legislative package that provides extensive assistance to Americans impacted by the coronavirus."[16] On March 19, the Trump government signed several legislative orders in helping people under the impact of COVID-19, especially orders that help supporting medium and small businesses as well as stimulus checks that every American can get. This relieved many people's lives in times of the worst as well as the average sentiment on that day.

Running tweets on March 19 through Word Cloud and LDA yield Figure 7 and Figure 8. There are keywords such as “million stock” and “Richard Burr”, which are related to the stock market. The causality is that the government signed legislative orders to help the broken stock market as well as supporting medium and small business to survive. Richard Burr is a scandal which senator Richard Burr’s brother sold of all his stock before the crash of the stock market, likely exploited non-public information.

In LDA analysis, a total number of 5 topics are extracted from the tweets on April 11. The LDA model has a coherence value of 0.33 and perplexity of - -8.00. From topic 2, we have terms including “need” “work”, and “help” which can be related to the government’s legislative orders in helping people during the pandemics.

Fig. 7: Word Cloud Result for March 19

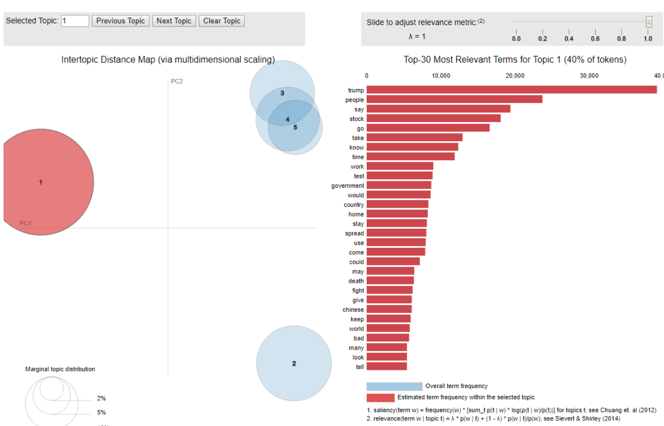

Fig. 8: LDA Result for March 19

March 28: “President Trump signed coronavirus relief bill at the White House”; “White House moves toward promoting face masks to fight virus” [16] On March 28, the positive change in sentiment score is correlated to the government’s further effort in relieving the financial regression caused by COVID-19. The government also promised increased medical supplies to the general public such as masks and encourage people to wear face masks to fight the virus.

Running tweets on March 28 through Word Cloud and LDA yield Figure 9 and Figure 10. “medical masks” and “masks available “catches our attention. The United States government was against personal usage of face mask at the beginning of the epidemic and changed its tune near the end of March. The government instead encourages people to always wear a mask when people are around. This likely caused a lot of confusion and debates over social media. Additionally, face masks are out of stock everywhere in the United States around the end of March. Government’s effort in making face masks more available makes people feel safer as they can wear a mask to protect themselves and others.

In LDA analysis, a total number of 5 topics are extracted from the tweets on April 11. The LDA model has a coherence value of 0.35 and perplexity of -8.19. From topic 2, we can make an inference from keywords “mask”, “supply”, “believe”, “much”, and “well” that the government is working hard at increasing the availability of masks to the public. And the public in response displayed an increase in sentiment score.

Fig. 9: Word Cloud Result for March 28

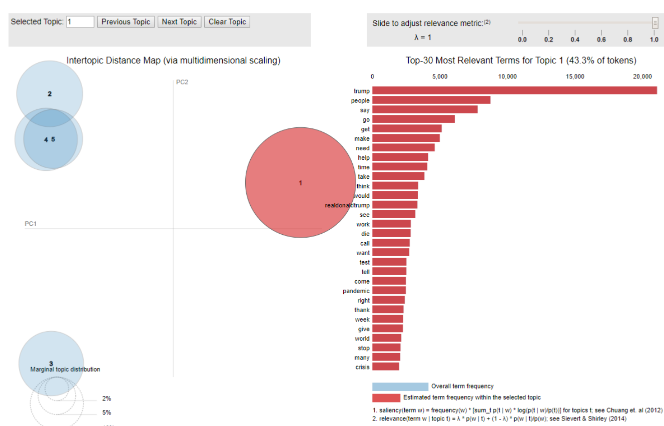

Fig. 10: LDA Result for March 28

Last but not the least, we tried to run sentiment analysis on more specific topics. We compared the sentiment score over time between all tweets and tweets containing “Trump” and Coronavirus” at the same time. Figure 11 shows the result. Tweets containing “Trump” and “Coronavirus” at the same time have a noticeable lower sentiment score across the board which indicates the public's dissatisfaction with the Trump Administration, or at least, Trump himself.

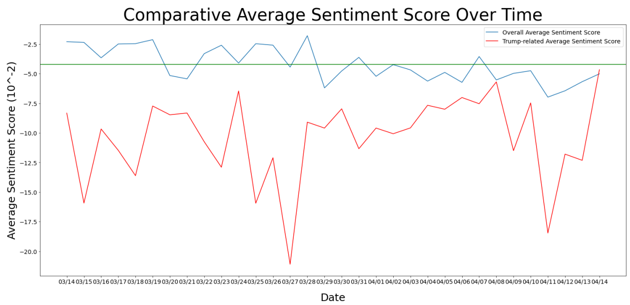

Fig. 11: Comparative Sentiment Score Graph

V. Conclusion and Future Work

We have presented a study focusing on discovering a correlation between COVID-19 related national policies and individual's sentiment behavior on social media during the ongoing COVID-19 pandemic. Utilizing Word Cloud and LDA analysis, we discovered an overlap between heated topic on twitter and national policies announced that day. By analyzing the fluctuation of sentiment scores over time, we further reassured the correlation between public sentiment and national policies. In addition, by comparing the sentiment score of tweets containing "Trump" and "Coronavirus" to all the tweets, we discovered a distinct dichotomy between the two curves. Results from these analyses show analyzing sentiment of users on social media on a large scale can be used as a metric to measure the popularity of a national policy among people. We believe we have introduced a new way of monitoring pandemics as well as a new approach for the government to connect and understand the public.

In the future, we want to try out different sentiment analysis tools such as VADER and see if we can have more accurate and credible results. We want to implement the location of the tweets to our algorithm so we can visualize the sentiment based on location using social media. In addition, we hope to develop a predictive algorithm based on data from social media that can be used for prediction of the spread of pandemics. Last but not the least, we will continue working on improving our visualization of LDA because it is not as intuitive as we expected in its current state.

References

[1] Timeline, the central events regarding Coronavirus in Wuhan., *https://www.nytimes.com/article/coronavirus-timeline.html*

[2] How the Coronavirus Pandemic Unfolded: A Timeline. , *https://www.sohu.com/a/373647917*$_1$14988

[3] Coronavirus Disease 2019 (COVID-19), Cases, Data, Surveillance., *https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/*

[4] Twitter by the Numbers: Stats, Demographics Fun Facts., *https://www.omnicoreagency.com/twitter-statistics/.*

[5] Zheng, Y., Ma, Y., Zhang, J. et al, *COVID-19 and the cardiovascular system*, Nat Rev Cardiol 17, 259–260 (2020).

[6] Dong, L., Hu, S., Gao, J, *Discovering drugs to treat coronavirus disease 2019 (COVID-19)*, Drug Discoveries Therapeutics 14,58-60 (2020), 259–260 (2020).

[7] Mehta, P., McAuley, D., Brown, M. et al, *COVID-19: consider cytokine storm syndromes and immunosuppression.* ,The Lancet 395,1033-1034 (2020).

[8] Dong, E., Du, H., Gardner, L, *An interactive web-based dashboard to track COVID-19 in real time*,The Lancet Infectious Disease 20,533-534 (2020)

[9] Anderson, M., Heesterbeek, H,*How will country-based mitigation measures influence the course of the COVID-19 epidemic*,The Lancet 395,931-934 (2020)

[10] Hellewell, J., Abbott S, Gimma,*Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts*,The Lancet Global Health 8, 488-496(2020).

[11] Zhao, Y., Xu, H,*Chinese Public Attention to COVID-19 Epidemic: Based on Social Media. medRxiv*

[12] Kouzy, R., Jaoude, J., Kraitem, A. et al,*Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter*,Cureus. 2020 Mar; 12(3): e7255

[13] Singh, L., Bansal, S., Bode L. et al,*A first look at COVID-19 information and misinformation sharing on Twitter*,arXiv:2003.13907 [cs.SI]

[14] Tweepy Documentation, *http://docs.tweepy.org/en/latest/*

[15] WordCloud Documentation, *https://amueller.github.io/*

[16] White House Briefing Statements, *https://www.whitehouse.gov/briefings-statements/*

[17] The Biggest Social Media Science Study: What 4.8 Million Tweets Say About the Best Time to Tweet, *https://buffer.com/resources/best-time-to-tweet-research*