COVID-19: Social Media Sentiment Analysis on Reopening

Mohammed Emtiaz Ahmed, Md Rafiqul Islam Rabin, Farah Naz Chowdhury

Department of Computer Science, University of Houston

mahmed24@uh.edu, mrabin@central.uh.edu, fchowdh2@central.uh.edu

Abstract

The novel coronavirus (COVID-19) pandemic is the most talked topic in social media platforms in 2020. People are using social media such as Twitter to express their opinion and share information on a number of issues related to the COVID-19 in this stay at home order. In this paper, we investigate the sentiment and emotion of peoples in the U.S.A on the subject of *reopening*. We choose the social media platform Twitter for our analysis and study the Tweets to discover the sentimental perspective, emotional perspective, and triggering words towards the reopening. During this COVID-19 pandemic, researchers have made some analysis on various social media dataset regarding *lockdown* and *stay at home*. However, in our analysis, we are particularly interested to analyse public sentiment on *reopening*. Our major finding is that when all states resorted to lockdown in March, people showed dominant emotion of fear, but as reopening starts people have less fear. While this may be true, due to this reopening phase daily positive cases are rising compared to the lockdown situation. Overall, people have a less negative sentiment towards the situation of reopening.

1 Introduction

COVID-19, a new strain of coronavirus, originated in Wuhan, China in December 2019 (Huang et al., 2020). On 11th March 2020, the World Health Organization announced the COVID-19 outbreak as a pandemic. This highly infectious COVID-19 has crossed boundaries in a speed anybody could have imagined turning the world upside down. COVID-19 has already claimed more than 0.3 million lives, with around 6 million people infected in more than 200 countries or territories [1] up to May 31, 2020.

---

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/

It has shaken the global economic and social structure so badly, that resuming normal life is a thing to envision. As a measure to control the situation, a number of countries have resorted to complete lockdown (Mitjà et al., 2020). When the whole world is COVID-19 stricken, the lockdown has been the only solution to get a hold (Ghosal et al., 2020). Since then much of the debate has been going on how the countries should ease the lockdown program. As lockdown is not the ultimate solution, easing the restrictions is the only way to minimize economic loss and keep the society functional. As the whole world is still not free from the deadly clutch of corona and continuing to claim human lives, reopen decisions from the government is a huge step to take. Public sentiment towards this reopen decision can help policymakers and general people to better understand the situation through classified information.

This COVID-19 outbreak has not only caused economic breakdown (Fernandes, 2020) across the world but also left a great impact on human minds. To capture human emotions, social media websites work as one of the best possible sources (De Choudhury and Counts, 2012). It is the fastest way for people to express themselves and thus the news feed is flooded with data reflecting thoughts dominating the peoples minds at this time. Apart from sharing their thoughts, people use these media to disseminate the information. During this lockdown, people have taken social networks to express their feelings and thus find a way to calm themselves down.

Emotions and sentiments are the driving force that people are sharing in social media. If we analyze the social media posts then we can find the insights of those posts along with emotions and sentiments. From this kind of analysis, we may find what people like, what they want, and what are the main concerns of them. Twitter has a large number

of daily active users all around the world where people share their thoughts and information regarding any topic of recent concern through this medium. It is a searchable archive of human thought. The affluence of publicly available data shared through Twitter has encouraged researchers to determine the sentiments on almost everything, including sentiments towards any product, service. Researchers and practitioners are already giving useful information on issues related to this COVID-19 outbreak. We have mainly contributed to the issue of the reopening phase giving an insight into people's reactions. Coordination of responses from online and from the real-world is sure to reveal promising results to address the current crisis. Our major finding is that during this reopen phase fear among people is less dominant compared to the previous times when the states made lockdown decisions. Another finding is that as the reopen phase starts, the number of new cases is also increasing. People are sharing their thoughts and opinions to combat this situation.

Contributions. This paper makes the following contributions.

- First, we generate the word cloud and N-gram representation of tweets to see the insights of twitter users during the reopening phase (more details in section 5.1).
- Second, we conduct a shallow analysis to study the sentiment and the emotion in reopening related tweets (more details in section 5.2 and section 5.3).
- Third, we analyze the actual effects of lockdown and reopening using real data (more details in section 5.4).
- Finally, we have made our code publicly available for an extensible work in similar areas.

The structure of this paper is organized as follows. Section 2 presents a literature review and some considerations of the previous work. Section 3 contains the information of dataset. Section 4 describes the experimental setup to analysis the reopening tweet. Section 5 provides an analysis of the results and findings. Section 6 discusses some of the challenges and threats to validity. Finally, Section 7 concludes the paper with future works.

2 Related Work

Researchers all around the world are analyzing Twitter data to discover people's reactions to the corona virus related issues such as lockdown, safety measures. (Barkur et al., 2020) analyzed Twitter data and found that people of India were positive towards the lockdown decision of their government to flatten the curve. (Dubey, 2020) made a sentiment analysis on the tweets related to COVID19 in some or the other way from March 11 to March 31 on twelve different countries, USA, Italy, Spain, Germany, China, France, UK, Switzerland, Belgium, Netherland, and Australia. They found that while most of the countries are taking a positive attitude towards this situation, there is also negative emotion such as fear, sadness present among people. Countries specially France, Switzerland, Netherlands, and the United States of America have shown distrust and anger more frequently compared to other countries. (Rajput et al., 2020) Made a statistical analysis of Twitter data posted during this Coronavirus outbreak. The number of tweet ids tweeting about coronavirus rose rapidly making several peaks during February and March. An empirical analysis of words in the messages showed that the most frequent words were Coronavirus, Covid19, and Wuhan. The immense number of tweet messages within a period of 2-3 months and the frequency of these words clearly show the threats the global population is exposed to. (Abd-Alrazaq et al., 2020) Aimed to identify the main topics posted by Twitter users related to the COVID-19 pandemic between February 2, 2020, and March 15, 2020. Users on Twitter discussed 12 main topics related to COVID-19 that were grouped into four main themes, 1) the origin of the virus; 2) its sources; 3) its impact on people, countries, and the economy that were represented by six topics: deaths, fear, and stress, travel bans and warnings, economic losses, panic, increased racism, and 4) the last theme was ways of mitigating the risk of infection. (Cinelli et al., 2020) ran a comparative analysis of five different social media platforms Twitter, Instagram, YouTube, Reddit, and Gab about information diffusion about COVID-19. They wanted to investigate how information is spread during any critical moment. Their analysis suggests that information spread depends on the interaction pattern of users engaged with the topic rather than how reliable the information is. (Samuel and Ali) Identified public sentiment using coronavirus specific tweets and provided insights on how fear sentiment evolved over time specially when COVID-19 hit the peak levels in the United States. As they classified tweets using sentiment

analysis, the fear sentiment, which was the most dominant emotion across the entire tweets and by the end of march its seriousness became clearly evident as the fear curve showed a steep rise.

3 Dataset

We used data from multiple sources. The primary source of this analysis is the social platform Twitter. We used the Python library `Tweepy` [2] (Roesslein, 2009) and Twitter developers API [3] for collecting the tweet data from Twitter. We used a unique Twitter dataset, which is specifically collected for this study using a date range from May 3, 2020, to May 15, 2020. Tweets were extracted using the hashtags: #covid19, #covid, #corona, #coronaviras, #corona-virus, #covid19-virus, and #sarscov2. We collected a total of 5,703,590 tweets from Twitter. In addition, we collected the COVID-19 time-series dataset from the GitHub repository CSSEGISandData/COVID-19 (Dong et al., 2020), maintained by the amazing team at Johns Hopkins University Center for Systems Science and Engineering (CSSE). It contains country and state-wise daily new cases, recovered and death data of COVID-19.

4 Methodology

In this section, we will describe the data preprocessing, and the experimental setup for sentiment and emotion analysis.

4.1 Data Preprocessing

The data we used from the Twitter post between May 3-15, 2020 is noisy and unstructured text in nature. Therefore, we applied the following preprocessing steps to clean the raw tweet.

- URL and Email: Removing URLs and emails from tweets using regular expression.
- Mention and Tagging: Removing mentions (@ symbol and word after @ symbol) and hashtags (only # symbol) from tweets using regular expression.
- Noisy Words: Removing irrelevant words (i.e. RT, &) after manual inspection, and stripping non-ASCII characters from the tweet.
- Newline and Whitespace: Removing newlines and extra whitespaces from tweets.

- US-reopen: In section 5.1 - 5.3, we first filtered tweets by locations (States of US) and then selected tweets having word '*reopen*'.
- Stopwords and Gazetteers: In section 5.1 for most frequent words, we removed words from tweets that are found in the Python NLTK `stopwords` and `gazetteers` corpus [4]. We also split the tweets into a list of words using the Python NLTK TweetTokenizer, and represent different forms of words to a single word using the Python NLTK WordNetLemmatizer (Loper and Bird, 2002).
- Multi-line Tweet: In section 5.2 for emotion analysis, we converted the multi-line tweet into a single-line tweet in order to use the corresponding APIs.

4.2 Sentiment Analysis

We used the Python library `TextBlob` (Loria, 2018) for finding the sentiments from tweets. In `TextBlob`, sentiments of tweets are analyzed in two perspectives: (1) Polarity and (2) Subjectivity. Based on tweets, this library returns the polarity score and subjectivity score. Polarity score is float value within the range [-1 to 1] where 0 indicates Neutral sentiments, the positive score represents Positive sentiments and the negative score represents Negative sentiments. Subjectivity score is also a float value within the range [0 to 1] where 0 indicates that it is a Fact and other values indicate the public opinions.

4.3 Emotion Analysis

We used the *Python* implementation of IBM Watson Tone Analyzer service [5] to detect emotional tones in tweets. The `ToneAnalyses` API uses the general-purpose endpoint to analyze the emotional tones of a tweet and reports the following seven tones: analytical, tentative, confident, joy, fear, sadness, and anger. The emotion detection process is summarized as following steps:

- Create a Tone Analyzer service in IBM Cloud for a specific region and API plan.
- Get the API key, the service endpoint URL, and the version of API.
- Authenticate to the API using IBM Cloud Identity and Access Management (IAM).

---

[2] https://github.com/tweepy/tweepy
[3] https://developer.twitter.com/en/docs
[4] https://www.nltk.org/api/nltk.corpus.html
[5] https://cloud.ibm.com/apidocs/tone-analyzer?code=python

Figure 1: **Word Cloud Representations.** These are the visual representations of words within tweets whose importance is visualized by way of their size. (a) word cloud representation of the overall dataset, (b) word cloud representation of reopen related tweets within the United States.

- Request up to 100 sentences (no more than 128 KB size in total) for the sentence-level tone analysis.
- Parse the JSON responses that provide the results for each sentence of the tweet.

Here for our analysis, the `ToneAnalyses` requires a tweet as a single sentence for detecting sentence-level emotional tones.

5 Results and Analysis

In this paper, we seek to answer the following research questions.

- RQ1: What are the most frequent words?
- RQ2: What are the trends of sentiment?
- RQ3: What are the trends of emotion?
- RQ4: What are the effects of reopening?

5.1 RQ1: Most Frequent Words

5.1.1 Word Cloud Representation

The tweets were organized into word clouds to analyze which words have been frequently used by Twitter users around the world. As word cloud visualization consists of the size and visual emphasis of words being weighted by their frequency of occurrence in the textual corpus, we can get insights from the most frequent words occurring. From figure 1a, we can see that along with the words 'COVID', 'corona', or 'virus', words like 'new', and 'cases' got a large number of mentions. No wonder the rising number of new cases everyday around the world contribute to the good number of mentions of these words. Our main goal is to see the effect of reopening in our analysis that's why we also investigate the most frequent words

used in reopen related tweets. Figure 1b represents only the reopen related most frequent words within the United States. We can see that 'businesses', 'schools', 'gyms', 'economy', and 'protesters' are the most frequently used words. Few tweets are 'Protesting to reopen the economy', 'reopen the economy against shelter in place which is a false choice', 'As schools make plans to reopen after #COVID19 shutdowns, how should we proceed?', 'Any rush to reopen without adequate testing and tracing will cause a resurgence.', 'As we reopen shuttered offices and buildings, there is a danger of an outbreak of Legionnaire disease'. From these few tweets, we can see that few people are protesting to reopen, but most of the people are concerned about the reopening timing, aftermath of reopening, and thinking about the precaution measures needed after reopening.

5.1.2 N-gram Representation

We searched for the most frequent words that appeared in twitter posts of our collected data. For this section, we only considered actual tweets and ignored all kinds of re-tweets from our dataset. After applying the preprocessing steps mentioned in section 4.1, we also removed hashtags and non-alphabetic characters from tweets. We split the rest of the tweet into chunks of n consecutive words. Next, we merged the chunks of all tweets as a flat list, where $n \in [1, 2, 3]$. After that, we count the frequency of each unique chunk and sorted those by frequency. The top 15 most frequent chunks, known as n-gram, are shown in Table 1.

From Table 1, we can see that the 'business', 'economy', 'back', and 'need' clearly indicate the

4

1-gram	2-gram	3-gram
state	testing site	drivethru testing site
business	social distancing	moderate discussion business
plan	white house	discussion business leader
economy	state begin	keep calling transparency
people	small business	calling transparency investigation
today	look like	transparency investigation fort
back	public health	stock fluctuated economic
case	stay home	fluctuated economic plan
testing	back work	economic plan mixed
country	wear mask	plan mixed earnings
need	hair salon	mixed earnings company
phase	next week	training class course
county	task force	class course offer
restaurant	business begin	course offer full
week	contact tracing	offer full service

Table 1: Top-15 N-grams (US-reopen).

economic breakdown and necessity to overcome this situation through reopen. The 'plan', 'case', 'testing', and 'phase' surely emphasize the safety measures that need to be followed while reopening as a control measure. The 'state' and 'people' surely stands for how all states beginning to reopen in some way.

From 2-grams, the 'social distancing', 'stay home', 'wear mask', and 'contact tracing' imply users try to raise awareness through their posts. The 'testing site', 'public health', and 'task force' imply the safety measures people and policymakers should abide by to curb the spread during the reopen situation. The 'state begin', 'small business', 'back work', and 'business begin' indicate as people are going to resume their work after states are reopening.

From 3-grams, the economic breakdown, stockmarket fluctuation, business plan, and training course are the topic that draws our attention. This surely indicates there were numerous posts about how the loss due to this pandemic can be overcome through the strategy of the people who are in charge of the decisions.

5.2 RQ2: Emotion Analysis

From figure 2a, we can see that the highest percentage of emotional tone is 'Analytical'(34.7%). The second highest tone was 'Joy'(17.35%). The next few tones are 'Tentative', 'Sadness', and 'Confident', respectively. 'Anger' and 'Fear' had the lowest percentage in our data collection.

The 'Analytical' tone implies that people shared information and their constructive thoughts through their posts. 'Joy' implies that people showed a positive attitude towards the situation. (Samuel and Ali)

showed that towards the end of March the 'Fear' emotion was the most dominant among all emotions and hit a peak. But after the announcement of the lockdown easement program people had less fear.

From figure 2b, we can see that, during May 13-15, 'Anger', 'Sadness' and 'Analytical' had high peaks. The rest emotional tones 'Confident', 'Joy', 'Tentative', and 'Fear' had a steady curve during the time period of our data collection.

5.3 RQ3: Sentiment Analysis

Sentiment analysis identies words contextual polarity and subjectivity. We used Python library `TextBlob` (Loria, 2018) for sentiment analysis. The details of `TextBlob` is available in section 4. Fig. 3 represents the details of twitter sentiments. Fig. 3a corresponds to the barplot of sentiment polarities of tweets and Fig. 3b corresponds to the daily counts of polarities. It can be seen that the majority of tweets have a neutral sentiment (43.66%) followed by positive (39.89%) sentiments. Also, the daily trend of positive and negative tweets follows the same pattern. So, people are taking the reopen decision positively. Fig. 3c indicates that people are mostly posting their own opinion and Fig. 3d shows an interesting point that day by day people are posting their own opinion rather than facts.

5.4 RQ4: Effect of Reopen

We saw that many peoples are protesting for reopening states and countries. We also saw it in Fig. 1b. We investigated the reopening effect in the United States. Fig. 4(a) and (b), shows that after reopening the new COVID19 cases are increasing. On the other hand, Fig. 4(c) and (d), shows that lockdown helps to reduce the number of new COVID19 cases. Lockdown also helps to reduce the new cases in other countries (Lau et al., 2020). It would be better if the lockdown continues for few more days or if the states or country decide to reopen, then they can take proper precautions before the reopen and people need to strictly maintain the instructions from health organizations.

6 Threats to Validity

The followings are the main challenges and threats to the validity of our study.

Limited Data. We only conducted a shallow analysis of the twitter post from May 3, 2020 to May
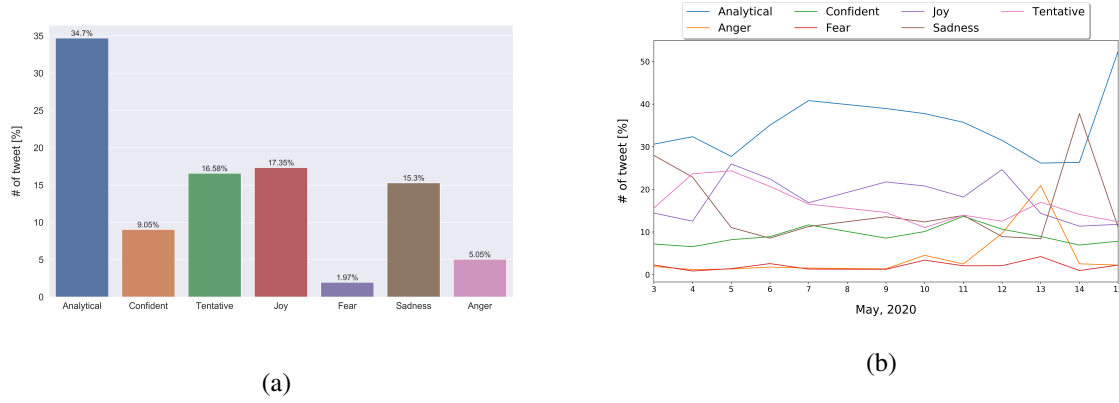
(a)

(b)

Figure 2: **Emotional Distribution.** (a) tweet percentages of different kinds of emotional tones (b) daily tweet percentages of different kinds of emotional tones.
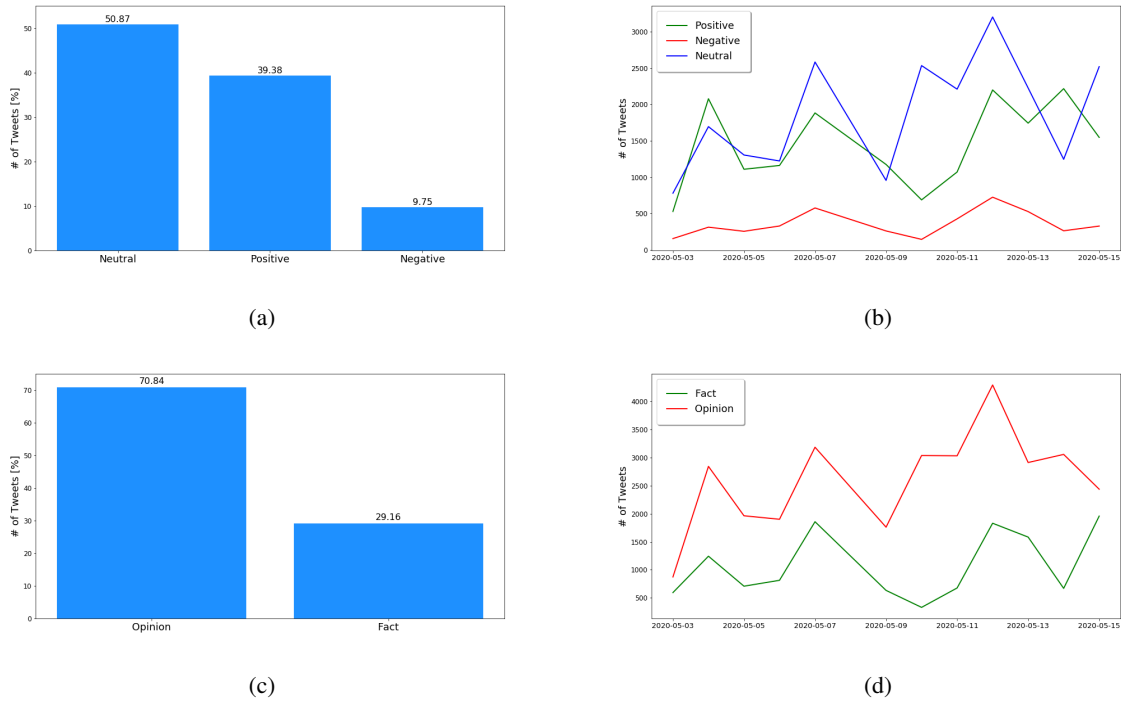

(a)

(b)

(c)

(d)

Figure 3: **Sentiment Distribution.** (a) tweet percentages of different kinds of sentiments (b) daily tweet counts of different kinds of sentiments (c) distribution of subjectivity types and (d) daily tweet counts of different kinds of subjectivity.

15, 2020. Therefore, our results may not reflect the actual effect. We leave the large scale evaluation as future work.

Noisy Data. The dataset contains posts from social media (i.e. Twitter) that are noisy and unstructured. Having mis-spelling and unknown words are also common in the post. Additionally, the dataset contains other languages but written using the English alphabet that can negatively impact the analysis. There are many posts whose length is very small (i.e. around 3 words). Therefore, dataset cleaning is challenging and important as it has a direct impact on the analysis.

Internal Validity. We used some open source toolchains in this paper and some issues may exist in those APIs. However, to reduce the issue in our implementation, other collaborators reviewed the code and manually inspected some results.

7 Conclusion

In this work, we try to get insights into public reaction as the reopen phase starts. There has been
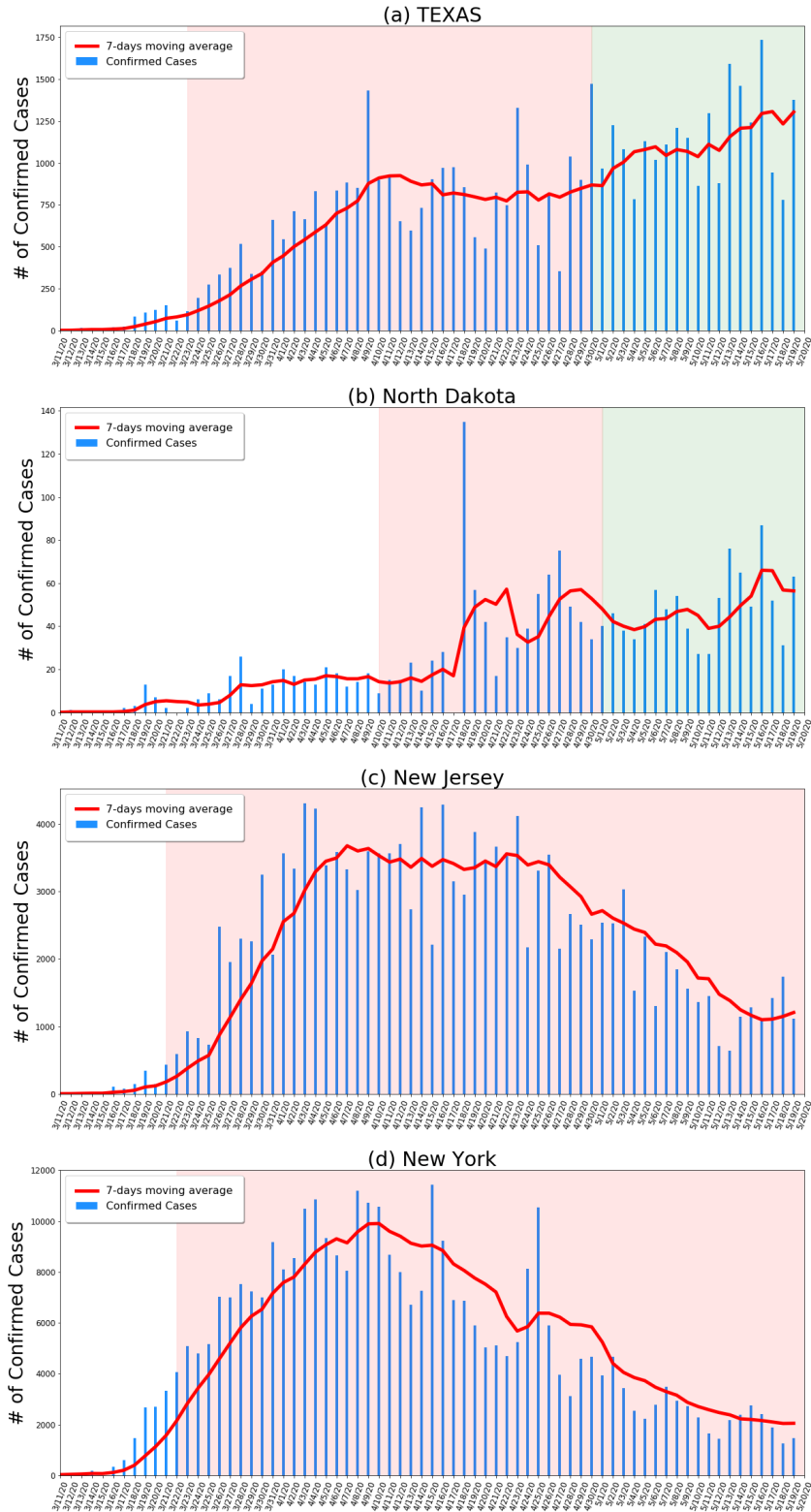
Figure 4: **Effect of reopen.** This figure represents the daily confirmed cases from 4 different states of the United States. We used two kinds of shaded regions to indicate the lockdown and reopen times: red shaded area represents the lockdown period, and the green shaded area represents the reopen period.

some analysis from social media data about how people are reacting in the lockdown time. We make an effort to understand whether there is a change in public sentiment from the lockdown phase to the reopen phase. We have made our analysis on Twitter data on reopening related issues during this

COVID-19 outbreak. From our analysis, we have found that even though people are showing a positive attitude to reopen their areas to make the economy functional, they are also urging the authority in concern to make plans for avoiding the highly predictable second wave. From real data, we have seen that the new cases are increasing as the reopen phase starts. Given this situation, the coordination of responses from online and from the real world is sure to reveal promising results to address the current crisis. We want to put stress on the fact that along with traditional public health surveillance, infoveillance studies can play a vital role.

Source Code. We have released the source code for public dissemination and can be found at https://github.com/emtiaz-ahmed/COVID19-Twitter-Reopen.

References

Alaa Abd-Alrazaq, Dari Alhuwail, Mowafa Househ, Mounir Hamdi, and Zubair Shah. 2020. Top concerns of tweeters during the covid-19 pandemic: Infoveillance study. *J Med Internet Res*, 22(4):e19016.

Gopalkrishna Barkur, Vibha, and Giridhar B. Kamath. 2020. Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india. *Asian Journal of Psychiatry*, 51.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic.

Munmun De Choudhury and Scott Counts. 2012. The nature of emotional expression in social media: measurement, inference and utility. *Human Computer Interaction Consortium (HCIC)*.

Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Akash Dutt Dubey. 2020. Twitter sentiment analysis during covid-19 outbreak.

Nuno Fernandes. 2020. Economic effects of coronavirus outbreak (covid-19) on the world economy. *Available at SSRN 3557504*.

Samit Ghosal, Rahul Bhattacharyya, and Milan Majumder. 2020. Impact of complete lockdown on total infection and death rates: A hierarchical cluster analysis. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.

Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506.

Hien Lau, Veria Khosrawipour, Piotr Kocbach, Agata Mikolajczyk, Justyna Schubert, Jacek Bania, and Tanja Khosrawipour. 2020. The positive impact of lockdown in wuhan on containing the covid-19 outbreak in china. *Journal of travel medicine*, 27(3):taaa037.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2.

Oriol Mitjà, Àlex Arenas, Xavier Rodó, Aurelio Tobias, Joe Brew, and José M Benlloch. 2020. Experts request to the spanish government: move spain towards complete lockdown. *The Lancet*, 395(10231):1193–1194.

Nikhil Kumar Rajput, Bhavya Ahuja Grover, and Vipin Kumar Rathi. 2020. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic.

Joshua Roesslein. 2009. tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5.

J. Samuel and M.M.; Esawi E.; Samuel Y. Ali, G.G.M.N.; Rahman. Covid-19 public sentiment insights and machine learning for tweets classification.