

CoronaVis: A Real-time COVID-19 Tweets Analyzer

Md. Yasin Kabir
Department of Computer Science
Missouri University of
Science and Technology, USA
mkabir@mst.edu

Sanjay Madria
Department of Computer Science
Missouri University of
Science and Technology, USA
madrias@mst.edu

Abstract—The goal of CoronaVis is to use tweets as the information shared by the people to visualize topic modeling, study subjectivity and to model the human emotions during the COVID-19 pandemic. The main objective is to explore the psychology and behavior of the societies at large which can assist in managing the economic and social crisis during the ongoing pandemic as well as the after-effects of it. The novel coronavirus (COVID-19) pandemic forced people to stay at home to reduce the spread of the virus by maintaining the social distancing. However, social media is keeping people connected both locally and globally. People are sharing information (e.g. personal opinions, some facts, news, status, etc.) on social media platforms which can be helpful to understand the various public behavior such as emotions, sentiments, and mobility during the ongoing pandemic. In this paper, we describe the CoronaVis Twitter dataset (focused on the United States) that we have been collecting from early March 2020. The dataset is available to the research community at <https://github.com/mykabar/COVID19>. We would like to share this data with the hope that it will enable the community to find out more useful insights and create different applications and models to fight with COVID-19 pandemic, and any future pandemics as well.

Index Terms—COVID-19, Coronavirus, Pandemic, Twitter;

I. INTRODUCTION

At the time of writing this document, there were more than 2.5 million confirmed coronavirus cases all over the world, and in USA more than 1M people are infected¹. The number of infected people, active cases, and fatality keep rising every day. The first confirmed case of novel coronavirus disease was reported in Wuhan, China. However, over the last couple of months, the virus spread explosively all over the world. At the time of writing this document, the United States has the maximum number of coronavirus cases and fatalities.

Every country is taking preventive measurements to fight against the COVID-19 pandemic. "Social Distancing" or Stay at home became the most widely used directive all over the world. Social distancing is forcing people to stay at home. As a result, this is impacting the public event, business, educations, and almost every other activity of the human life. People are losing their jobs, and getting infected from corona and thus, stress is rising at the personal and at the community levels. Studies of behavioral economics show that emotions (Joy, Anger, Worry, Disgust, Fear, etc.) can deeply affect

individual behavior and decision-making. Social networks have the hidden potential to reveal valuable insights on human emotions at the personal and community level. Recent works [1]–[4] show that twitter data, and human emotions analysis can be useful for predicting crimes, stock market, election vote, disaster management, and more. To find out the useful insights from public opinions and shared posts in social media, and to model the public emotions, we have started collecting tweets from 5th March 2020. We have collected and processed over 100 million tweets related to Coronavirus (focused on USA) which is about 700GB data in size. We understand that processing this huge amount of data in real-time requires a substantial amount of time and resources. Hence, we decided to share the processed dataset with the community so that they can skip the raw data processing part and dive into data analytics and modeling.

II. DATA DESCRIPTION

A. Data Collection

We are continuously collecting the data since March 5, 2020 and will keep fetching the tweets using Twitter Streaming API² and Tweepy³. We have collected around 700GB of raw data until April 24, 2020 and saved this data as JSON files. However, we dynamically process this data in real-time for the CoronaVis⁴ application. We processed several features from the tweets such as (Tweet ID, Tweet Text, User Location if available, User Type), etc. We will keep collecting the data and update the data repository (<https://github.com/mykabar/COVID19>) once in every week. Future versions of this paper will record the updated information and analytics on the collected data.

B. Data Description

The processed data is saved and updated in the git repository using the fetched data. Every single file contains data of the date that is specified as the name of that file. All the data file contains 6 different attributes (tweet_id, created_at, loc, text, user_id, verified). The data contains tweets only with

¹<https://coronavirus.jhu.edu/map.html>

²<https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data>

³<https://www.tweepy.org/>

⁴<https://mykabar.github.io/coronavis/>

the location information as we focused our analysis only on the US data. To keep the privacy, in both our application and in shared data, we introduced some data anonymization. We included tweet ID with the data, hence the researcher can re-fetch the original tweet if that tweet is still public. However, we respect if a user wants to remove tweets, and to ensure that we processed the tweet text and user name so that that can not be directly searched and linked to a particular user.

TABLE I
DATA ATTRIBUTES

Feature	Description
tweet_id	Unique ID of a tweet.
created_at	Creation time of a tweet.
loc	State level user location.
text	Processed tweet text. All the text are in small letters, non-English characters and few stop words are removed.
user_id	Pseudo user id. The exact user name is transformed to a anonymous id to preserve the privacy of the user.
verified	Denotes whether the tweet post is verified or not (1 or 0).

There are some gaps in the collected datasets due to API and connectivity issues. However, we are filling up those gapes using other publicly available data repositories. We will also perform more data analysis and update this paper with the more updated data and insights.

C. Data Analytics

Figures 1 to 4 represent some example of data analytics from the processed dataset. Figure 1 represents a word cloud consists of the top 20 frequent bi-grams in the tweets.

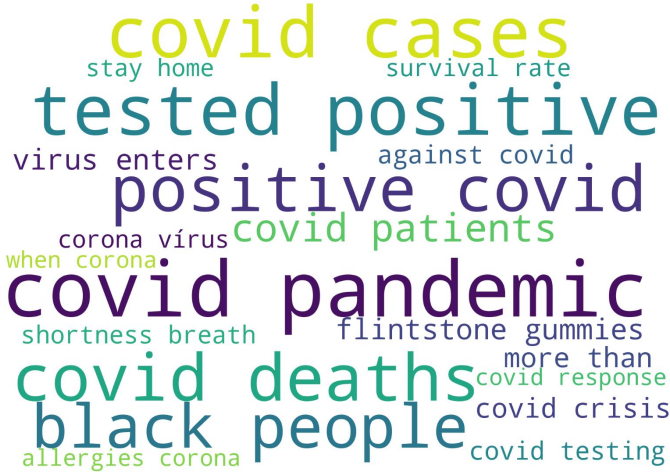


Fig. 1. Top 20 frequent bigrams

Polarity over time for different states is depicted in figure 2. In figure 2 (a) represents a clickable geo-map where a user can select a state and observe polarity (sentiment) of that state over a specific time period. The word cloud (b) provides an idea of what has driven the sentiments in that state, and (c) figure 2 shows the polarity over time. In figure 2, the polarity

analysis for New York (NY) states is presented. The CoronaVis application also allows observing the subjectivity analysis of any state in a similar fashion.

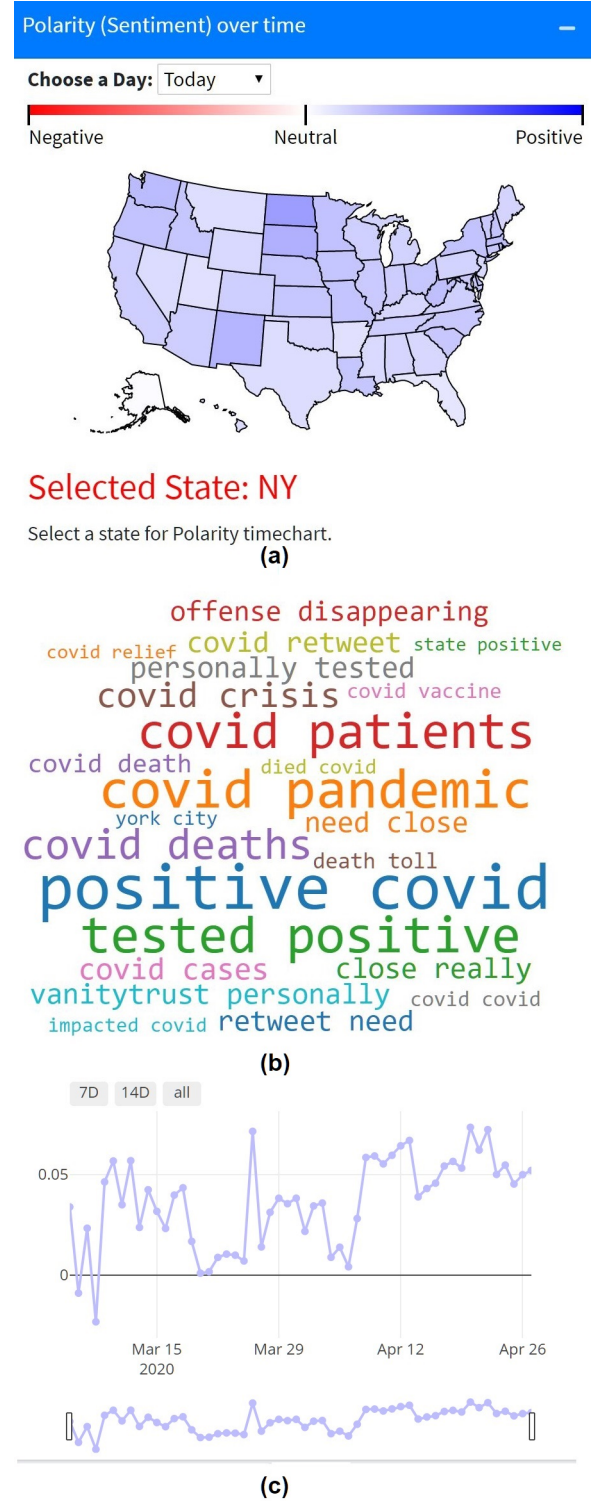


Fig. 2. Polarity over time

The number of the users moving between two or more states and the number of coronavirus cases in those states are

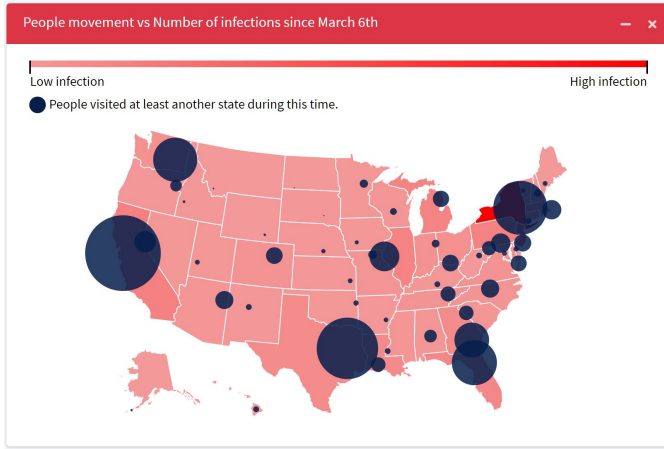


Fig. 3. Tweet frequency per day

depicted in Figure 3. We observe that the topic trends became different for different states in USA. To observe how the topic trends are evolving, and what challenges the community are discussing, we have created two different graphs. Figure 4 shows the most frequent topics, and the selected featured topics for New York (NY) state. From the figure, we can observe that the topic **test** became highly frequent on April 20th, and we can also observe that the topics such as **die**, **death** also became popular on the same date.

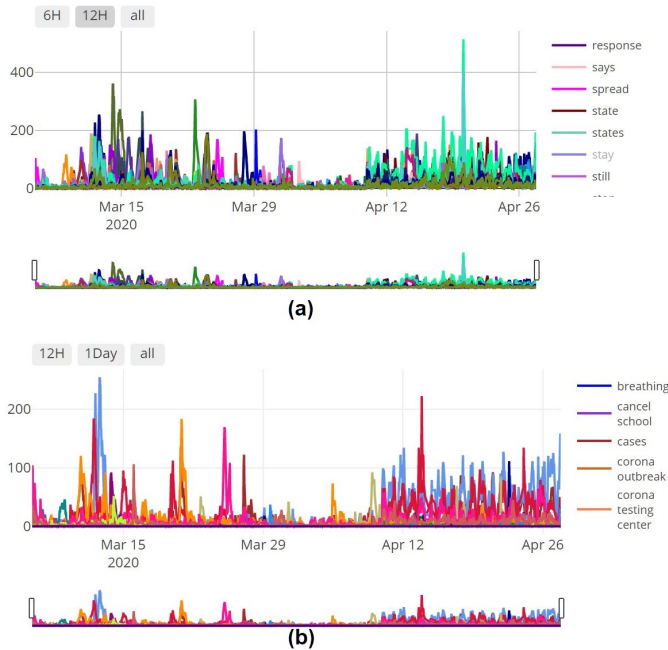


Fig. 4. Polarity over time

The live application (CoronaVis) is accessible at <https://mykabir.github.io/coronavis/>. We are working continuously to add more real-time data visualization and analytics on the website.

III. DATA ACCESS AND USE POLICY

The data is accessible from the Github repository <https://github.com/mykabir/COVID19>. The repository contains two different folders. One is the **data** folder containing the tweeter data and another is a **src** folder which contains some basic code presenting the way to read the data and some basic data analytics. The **data** folder contains the data in CSV file format for each day from 5th March 2020 to till date and named by the particular date with the format YEAR-MONTH-DATE. If you have any suggestions or concern, please send an email at mkabir@mst.edu or madrias@mst.edu.

A. Use Policy

This dataset is released in compliance with Twitter's Developer Terms & Conditions⁵. The data repository will be continuously updated every week. The data repository, containing codes, and CoronaVis⁶, 2020 W2C lab, Missouri University of Science and Technology, all rights reserved, can be used for educational, academic, and government research purposes with proper citation (Please cite this paper). Any commercial use of any materials is strictly prohibited. Taking and sharing a screenshot is allowed with appropriate citation.

REFERENCES

- [1] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using twitter sentiment and weather," in *2015 Systems and Information Engineering Design Symposium*. IEEE, 2015, pp. 63–68.
- [2] M. S. Gerber, "Predicting crime using twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, 2014.
- [3] P. Grover, A. K. Kar, Y. K. Dwivedi, and M. Janssen, "Polarization and acculturation in us election 2016 outcomes—can twitter analytics predict changes in voting preferences," *Technological Forecasting and Social Change*, vol. 145, pp. 438–460, 2019.
- [4] M. Y. Kabir and S. Madria, "A deep learning approach for tweet classification and rescue scheduling for effective disaster management," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 269–278.

⁵<https://developer.twitter.com/en/developer-terms/agreement-and-policy>

⁶<https://mykabir.github.io/coronavis>