Final Project

8 th August 2020

Jason Zheng, Kenny Dove, Sam Tsai, Veronica Valencia, Zainab Busari

OVERVIEW:

In this project, we will use Twitter data to retrieve people's tweets regarding to COVID19 in the United States. During the rapidly evolving pandemic, the United States has the highest number of COVID19 infections and number of deaths. Social media data provides real time access to the impact of COVID19 on the nation's health system, economic activities, and a new social norm called "social distancing". The overarching goal of this project is to access twitter data through an API, analyze it with sentiment analysis, and classify them using machine learning techniques.

GOALS:

1. To apply our Twitter developer account and have access to twitter data using Tweepy API.
2. To use sensitivity analyses to understand the general impressions that Americans have of COVID-19: positive, negative, or neutral.(clustering and sentimental analysis)
3. To classify tweets into several groups, e.g. economic issues, barriers to access health care systems, social distancing, school opening, wearing masks, vaccines etc.
4. Create a bar chart or heatmap summarizing the location by state of twitters.

Techniques:

First, we will apply for a Twitter developer account to have access to Twitter data, .which allows us to retrieve the data using an API called "Tweepy". The free Twitter account will allow us to download up to 7 days of most recent tweets. We will restrict our search within 7 days and in the United States only. Simply Because there are 500 million tweets sent each day, which equates to 6,000 tweets every second, we will only take a sample of the total tweets related to COVID-19. The specific keywords to identify in our data is as follows:

{"corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic", "#pandemic" "#2019ncov", "2019ncov", "quarantine", "#quarantine", "flatten the

curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines"}

The next step is to use “TextBlob” to conduct a sentiment analysis and rate each tweet as positive, negative, or neutral. However, this method won’t get us too far. There are a lot of dimensions people are discussing, for examples, some may commenting on the economic issues, such as reopening the economy, job loss, unemployment benefits, etc; others may concern about their health and whether it is safe to go to the hospitals to get care; while some others are thinking about vaccines debating on when the vaccine will be out and also if they are willing to take the vaccine immediately.

Therefore, we will conduct a machine learning process to classify tweets into different groups. We will perform a supervised learning and specify each domain and flag tweets into different groups: economic issues, health issues, or others.