Progress Report

Tasks Completed

So far we have mainly completed three tasks:

1. We successfully set up a Twitter developer app that can interact with Twitter API using Tweepy. Now we can pull tweets by id, date, tag, and geographic location. By overwriting the Tweepy stream client, we can filter and sample real-time tweets related to COVID-19. To better analyze the sentiment of the retrieved tweets, we also implement some text preprocessing, such as handling special characters, single characters, and tag removal.
2. Try three different models for sentiment analysis using the SEM Eval 2016 dataset : (a) Use CountVectorizer to create a bag of word model and train a log regression model for sentiment analysis. (b) Combine the Lexicon-based sentiment analyzer TextBlob with Sentiment Intensity Analyzer from the NLTK library to generate sentiment labels and subjectivity scores. (c) Adapted transformer with FASTAI. Using the transformers library from Huggingface (https://github.com/huggingface/transformers), cooperating with the learning method from FASTAI, we are able to generate a model with an accuracy of 0.61. We did not spend too much time tuning the parameter since the learner class from FASTAI can help us improve the model in a single one-cycle fitting method.
3. Implement simple data visualization such as generating word cloud for tweets with the same sentiment and calculating term frequency for different sentiments.

Tasks Pending

We currently still have three unfinished tasks:

1. Collect enough training tweets related to COVID-19 and label all of them manually for testing purposes.
2. Add more data analysis and visualization about the sentiment of public feedback for different COVID-19 vaccines.
3. Develop a Front-end application that displays all the data visualization and allows a quick sentiment analysis for tweets from users.

Challenges

1. We only have limited access to Twitter API and cannot access the search-all endpoint of Tweepy to pull tweets given any time intervals. Specifically, we need the Elevated Access of Twitter APIs to retrieve tweet contents. This significantly reduces our data sources. Moreover, the maximum number of tweets retrieved using tweet id is 100 per request. Due to the limited number of requests to Twitter API every 15 minutes, we are not able to handle a large number of user requests about sentiment trends given a time period.
2. Currently, we are not able to train the model based on a large dataset that satisfied all our requirements because it costs too much time. Finding a proper

learning rate and fitting the model would require an hour to train on a dataset containing only 600 entries. At the same time, we also need a seed function to generate random tests that would help us improve our model. Currently, the Top 3 scores based on general tweet sentiment analysis have an accuracy between 72.83 and 74.71. We want to achieve an accuracy somewhere around 70 without putting too much time and effort into it since we also wanted to develop a graphic user interface where users can visualize the data and have an intuitive experience with our results.