**Change of Covid Sentiment Overtime: An Analysis of Prominent Words Present In Covid Tweets - Checkpoint Report**

Viswesh Uppalapati, Skylar Wang

4th February 2022

**Introduction and Context**

People are highly opinionated beings and the power of social media to influence public opinion is an important area of study. We often have strong opinions of some topics and malleable ones of others, however, there is always room for change. Many external experiences, events, or factors can have an influence on opinion. One such factor is the intake of information from the different social media platforms that exist today. The existence of opinionated messages about any topic all over the internet and the ease of its accessibility can result in a shift in public opinion over time.

The Covid-19 pandemic has been a daily occurrence in recent years of our life. A global pandemic is often commonly talked about in current events and social media. Although people may have polarizing opinions on the topic, there is always a shift in sentiment over time based on the things that are happening around the world. In this study, we will analyze the Covid-19 related tweets from the beginning of the pandemic to the present day to characterize each month by the most prominent words that are present in the sample of tweets from that month. In doing so, we expect to create an interactive visual representation of the change in prominent words that characterized each month. In addition, any major shift in prominent words could be linked to that of an external event around that time in current events that will be reported along with the visualization. In doing so, we can better understand the shift in public opinion towards the pandemic over time.

**Methods and Data**

**Dataset and Data Collection**

The dataset we used and sampled from was shared by Rabindra Lamsal on IEEE DataPort. IEEE DataPort is a globally accessible data platform that allows storing and sharing datasets. The dataset includes CSV files that contain ID and sentiment scores of the tweets related to the COVID-19 pandemic. Lamsal collected coronavirus-related tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. The dataset dates from October 01, 2019, to the present. Due to the re-design of the dataset on March 20, 2020 in compliance with Twitter’s content redistribution policy and our goal of doing monthly analysis, we decided to limit our samples of tweets from May 1, 2020 to February 1, 2022. From this interval in time, we sampled 100,000 from each day. The dataset contains a total of 1,926,515,560 tweets (up until February 4, 2022) and we sampled a total of 2,013,000 tweets. The tweets are collected globally, but include only those tweeted in english.

The tweet IDs need to be hydrated in order to get the raw data of the tweet, so we used twarc. Twarc is a command line tool and Python library for collecting and archiving Twitter JSON data via the Twitter API. We take each of the sampled tweet ids from the dataset and hydrate those tweet ids in order to get the content of the tweets.

The sentiment scores contained in the dataset after March 20, 2020 were computed using the TextBlob’s Sentiment Analysis module [3]. TextBlob’s sentiment analysis model computes the sentiment polarity as a continuous value where the sentiment scores are defined in the range [-1,+1]. A tweet is considered to have a Positive sentiment if its score falls between (0,+1], a Negative sentiment if its score is in the range [-1,0), and Neutral sentiment if its score is 0. The sentiment is considered stronger if the score’s absolute value is closer to 1.

**Moving Forward**

With the dataset of tweets gathered via the database of tweet ids and the Twitter API, we intend to perform text analysis on various aspects of the dataset. The approximate 90000 tweets of each month during the considered timeframe will be used for tf-idf analysis where each tweet with text content will be considered as a document in order to extract the most relevant and important words of the collection of tweets within each month. In addition, the tweets of each month can undergo unigram, bigram, and trigram frequency analysis where the most important respective n-grams are extracted from each tweet. These words will be used in the visualization to show the most relevant and used words during the pandemic over that month of data. In addition, any major shifts and changes in these words between each month will be recorded. Most importantly, we will use the Twitter-roBERTa-base sentiment analysis tool recommended by our mentor to perform sentiment analysis on the tweets collected from each month. Lastly, the sentiment scores found from the database will be used along with these extracted relevant words in order to classify the tweets to the specific month of the pandemic.

**References**

1. <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment?text=I+like+you.+I+love+you>
2. <https://huggingface.co/>
3. <https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset>