Research Proposal
Twitter Sentiment Analysis of COVID-19 related Tweets

Irisa Shrestha
Earlham College
Richmond, Indiana, United States
ishres19@earlham.edu

1 ABSTRACT

Nowadays, people use social media to express their feelings and thoughts regarding different topics. During the COVID-19 pandemic, people have been very active in social media, sharing information, personal experiences, and emotions. Twitter is one of the social media platforms where people are very active. Sentiment Analysis of Tweets can provide really interesting insights about the public's preferences and sentiments. In this research, we will determine whether the sentiment changes of people's COVID-19 related Tweets align with significant developments in the COVID-19 timeline.

2 INTRODUCTION

COVID-19 has affected millions of people all over the world. Everyday we hear the news of new COVID-19 cases deaths, travel restrictions and stay-at-home orders. This has had a negative impact on Americans' mental well-being [1]. During these hard times, people have used social media to express their feelings and share information. In my study, I will be using Twitter data and doing Twitter sentiment analysis to find out whether there are changes on how people feel after a major development in COVID-19 timeline. Are they feeling positively or negatively after the vaccine roll out? Are there increase in positive sentiments after the uplifts of lockdown? These questions can help governments and health authorities understand people's perception to different regulations and COVID-19 related developments.

Research Question: Does the Twitter users' sentiment changes align with major developments in COVID-19 timeline? Which sentiment analysis tool aligns more with the major developments in COVID-19 timeline?

To address this research question I will analyze Twitter data, COVID-19 related Tweets in the United States. I will use VADER (Valence Aware Dictionary and sEntiment Reasoner) [5] and Textblob [6] to do sentiment Analysis on these data. Then I will compare both of my Twitter sentiment analysis result to a COVID-19 timeline, which will include all the major developments during the pandemic. I would like to see if there are mass sentiment changes, during the time of a new development in COVID-19 pandemic and, which Sentiment Analysis tool's result was closer to the developments in the COVID-19 timeline.

3 RELATED WORK

Many researchers have used Twitter to predict the sentiments of people during the COVID-19 pandemic. Xue et al. used a Twitter dataset to identify latent topics related to COVID-19, the themes of the identified topics, how Twitter users are reacting to the pandemic emotionally, and how their sentiments are changing over time [9]. They used Latent Dirichlet allocation (LDA), a topic modelling algorithm, to identify patterns, themes and structures of the Tweets' text and identify the connect between those themes. For their sentiment analysis they used a machine learning model that classified each English Tweets into eight pairwise emotions in Plutchik's wheel of emotion. The emotions included joy-sadness, trust-disgust, fear-anger, and surprise-anticipation. Their research showed that from Jan 23, 2020 to March 07, 2020, fear was consistently the dominant emotion. However, after fear, Tweets about trust also slightly increased over time.

Doogan et al. also used a Twitter dataset, with 777,869 English-language Tweets about COVID-19 in six countries, and analyzed these to find the public perception towards Nonpharmaceutical interventions (NPIs) (wearing masks and social distancing) [4]. Valdez et al. used U.S Covid-19 related Tweets to find out what were the major themes in these Tweets [8]. According to their research, COVID-19 related Tweets focused on China in Febraury 2020. However, as COVID-19 started spreading in the United States, from March to April, the theme was was focused in US-centered topics like "lockdown" and "social distancing."

Valdez et al. also looked for patterns that emerged from longitudinal sentiment analysis of the COVID-19 Pandemic [8]. For this they applied the VADER sentiment analysis tool to the COVID-19 corpus (U.S Tweets related to COVID-19) and the user timeline data. The user timeline data contained the individual user's Twitter timelines (3200 most recent Tweets), and the Tweets did not have to be related to COVID-19. They used this dataset to measure the fluctuations in mood, behavior, and emotions of individual users included in the COVID-19 corpus and living in the 20 US cities with the most COVID-19 cases per 100,000 people. The COVID-19 corpus was used to evaluate the overall sentiment of COVID-19 related Tweets. With the COVID-19 corpus there were two significant change points. There was a significant increase with VADER sentiment after WHO declared COVID-19 as a pandemic and after Donald Trump declared a national emergency. Overall there was an increase in the percentage of positively scored Tweets over time. However, the user timeline data showed the exact opposite trend, VADER sentiment score decreased over time. There was also only one PELT-identified change. Pruned Exact Linear Time (PELT) algorithm attempts to identify change points in a given time series when strict conditions are satisfied. The change occurred when former NBA player, Kobe Bryant died (his death was not related to COVID-19). The VADER sentiment score decreased during this time. There was another short drop in sentiment right before WHO declared COVID-19 as a Pandemic. But overall people's sentiments

were lower than before the pandemic hit, and did not return to the levels before the pandemic.

4 DESIGN AND IMPLEMENTATION

In this research project I will use sentiment analysis tools in COVID-19 related Tweets data, to find changes in Twitter users' sentiment throughout the pandemic. I will compare the results of these tools with a major development in COVID-19 timeline to find out which result is more closely aligned with the timeline.

Firstly, I will create two COVID-19 development Timelines. I will use the timeline created by Derrick Bryson Taylor in his New York Times article [7]. One timeline will be restricted to COVID-19 development in U.S, the other timeline will have all the events that are included in Taylor's timeline. Secondly, I will use an open-source repository of COVID-19 related Tweets [2] to create the COVID-19 dataset. This repository contains multilingual COVID-19 corresponding Tweets IDs, which I will use to extract the contents from twitter's API. Thirdly, I will clean up my dataset by filtering out all the reTweets, non-English Tweets and Tweets that are not located in U.S. Fourthly, my cleaned up dataset will be used for sentiment analysis. I will be using two sentiment analysis tools, TextBlob and VADER. These results will be used to create two data visualizations. Finally, I will compare the two sentiment with both of my Timelines. For this I will create a line chart for both the sentiment analysis, then I will mark the major development dates, from the timeline, in the line chart. This way I will be able to see if the sentiments have changed right after the major development. The framework is shown in Figure 1.

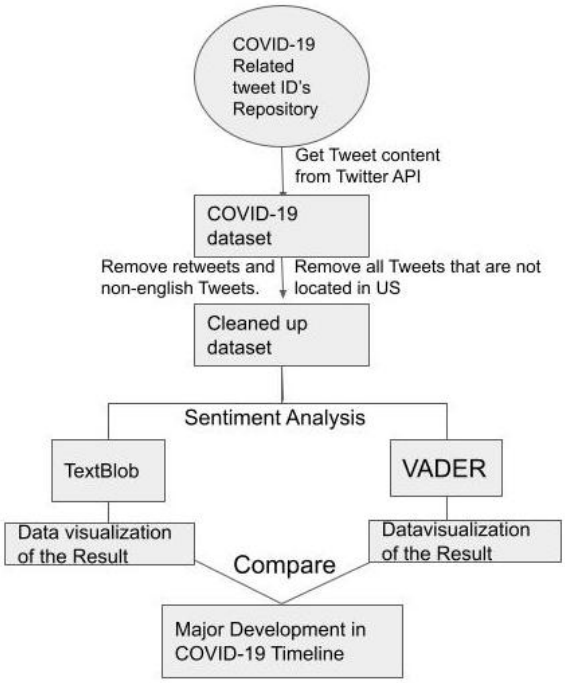


Figure 1: Framework of the Project

4.1 VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a Sentiment Analysis tool that is attuned to sentiments expressed in social media. It is a lexicon and rule based sentiment analysis tool [5]. The VADER lexicon contains 7500 common terms, each rated by 10 independent human raters [8]. For each text you use VADER sentiment analysis tool, it gives 4 scores, coumpound, pos, neu, and neg. The compound score is calculated by first adding the valence score of each word in the lexicon, then adjusted according to the rules, and finally normalized to be between -1 and +1. -1 being extreme negative and +1 being extreme positive. Pos, neu and neg scores are ratios for proportions of text that fall in positive, neutral and negative category. Sum of these scores will either be 1 or close to 1. The sentiment is positive when compound score is at least 0.05, neutral when compound score is more than -0.05 and less than 0.05, and negative when the compound score is at most -0.05. [3]. In my Twitter sentiment analysis I will calculate the average compound score of the Covid-19 related Tweets, and the number of positive, negative and neutral Tweets each day.

4.2 TextBlob

5 MAJOR RISKS

One of my major Risks is not being able to get the Twitter Dataset together, or having problems filtering out all the Non-U.S and Non-english Tweets. Another major risk is not being able to figure out an effiecient way to use Textblob or VADER for the huge Twitter dataset.

6 TIMELINE

First Week: 1. Create two Major COVID-19 developments timeline, one with only U.S related developments and the other with all U.S and international developments. 2. Get the COVID-19 Twitter dataset together by getting Tweets contents from the Twitter API, and using Tweets ID's from the COVID-19 Related Tweets ID's Repository [2].

Second Week: 1. Filter out all the non-U.S and non-English Tweets from the COVID-19 Twitter dataset. 2. Start experimenting with Textblob and VADER.

Third week: Experiment VADER and TextBlob with COVID-19 related Tweets corpus.

Fourth week: Work on the paper, and run the sentiment analysis on the entire COVID-19 corpus using VADER.

Fifth Week: Run the sentiment analysis on the entire COVID-19 corpus using TextBlob.

Sixth Week: Organize the VADER outputs, and start working on a code to calculate the average compound score, and the number of negative Tweets and positive Tweets each day.

Seventh Week: Continue working on the code and start organize the Texblob outputs.

Eighth Week: Start to write a code that calculates the average polarity and the number of negative Tweets and positive Tweets each day.

Ninth Week: Work on the paper, and continue with the code. Calculate the average subjectivity of positive Tweets and Negative Tweets.

Tenth Week: Create a line chart, with compound score on the Y-axis and the dates on the x-axis for the VADER sentiment analysis output. The number of positive and negative Tweets each day, will be labeled in the line chart.

Eleventh Week: Create a line chart, with polarity scale on the Y-axis and the dates on the x-axis for the Textblob sentiment analysis output. The number of positive and negative Tweets each day, and the subjectivity will be labeled in the line chart.

Twelfth Week: Add the timeline dates on the VADER output line chart.

Thirteenth Week: Add the timeline dates on the TextBlob output line chart.

Fourteenth Week: Work on the paper, analyse the charts and find the differences between TextBlob's results and VADER's results. Should be done with the software part by now.

Fifteenth Week: Work on the paper.

7 ACKNOWLEDGEMENT

REFERENCES

[1] John Auerbach and Benjamin F Miller. 2020. COVID-19 exposes the cracks in our already fragile mental health system.
[2] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: The first public coronavirus twitter dataset. (2020).
[3] Cjhutto. [n. d.]. cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. https://github.com/cjhutto/vaderSentiment#citation-information
[4] Caitlin Doogan, Wray Buntine, Henry Linger, and Samantha Brunt. 2020. Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data. *Journal of medical Internet research* 22, 9 (2020), e21419.
[5] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.
[6] Steven Loria. 2018. textblob Documentation. *Release 0.15* 2 (2018).
[7] Derrick Bryson Taylor. 2020. A Timeline of the Coronavirus Pandemic. https://www.nytimes.com/article/coronavirus-timeline.html
[8] Danny Valdez, Marijn Ten Thij, Krishna Bathina, Lauren A Rutter, and Johan Bollen. 2020. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data. *Journal of medical Internet research* 22, 12 (2020), e21418.
[9] Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. 2020. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PloS one* 15, 9 (2020), e0239441.