Sentiment Analysis of Tweets during different Phases of COVID-19 Pandemic

Darsh Shah
Data Science
University of Colorado Boulder
Boulder, Colorado, USA
darsh.shah@colorado.edu

Vishwam Thakkar
Data Science
University of Colorado Boulder
Boulder, Colorado, USA
vishwam.thakkar@colorado.edu

Rujula Nadipi
Data Science
University of Colorado Boulder
Boulder, Colorado, USA
Rujula.nadipi@colorado.edu

Xingyu Chen
Data Science
University of Colorado Boulder
Boulder, Colorado, USA
xich9921@colorado.edu

ABSTRACT

The COVID19 has continued to spread all over the world. It affects not only public health and global economics but also mental health and mood. While the impact and treatment of the COVID19 has been widely studied, relatively fewer analyses on the sentimental reaction of the population have been available. In this article, we scrape COVID19 related tweets on Twitter, and examine the tweets from Mar 15, 2020, to Mar 15, 2021, in the U.S. Applying several ML methods, we evaluate the sentiment intensity scores and visualize the information over different periods of the pandemic. For example, how is their mood variety related to the vaccine, lockdown, masks, daily confirmed case.

KEYWORDS

COVID-19, Sentiment Analysis, Tweets, Twitter, Pandemic.

1 Related Work

There are some reports done on sentiment analysis on tweets from the COVID-19 Pandemic timeline during recent times. In this section, we will discuss the most related studies and compare them to the current study.

Authors from [1] analyzed the overall emotions of COVID-19 related Tweets in the timeline March 2020 to September 2020. They used the SENWAVE dataset [4] as the training dataset which has hand-labelled sentiments. It features 11 different sentiments labelled by a group of 50 experts for 10,000 tweets worldwide during COVID-19 pandemic in 2020. This project will also try to use the SENWAVE [4] dataset as well to predict the sentiments in the Tweet but compared to the report from authors in [1], the project will be focusing on a larger time frame which is from March 2020 to March 2021 which will be further divided into smaller intervals based on changes in the COVID-19 cases and vaccination trends. This is important as there can be a change in sentiment polarity after the introduction of vaccines as well as with ease in the lockdown restrictions. Also, the report in question analyses the data from India and 2 of its most hit states in contrast to this project which will be focusing on the tweets from the United States which is more actively engaged in social media. The team is inspired from the report on using the BERT or LSTM models in order to achieve accurate results as they are tried and tested.

Authors from [2] have done their analysis on smaller time intervals, which is a 2-day time interval between 22nd March 2020 to 31st March 2020. The report has also done some basic level sentiment checks like positive, negative and neutral rather than focusing on the larger group of emotions. This project will be taking the inspiration from the report on dividing the time frame into smaller intervals but rather than focusing on a week of time frame, the project will be focusing more on a year of time frame divided strategically in smaller intervals based on different phases during the COVID-19 pandemic. Also, the authors have focused on the time frame during the rise of the coronavirus as opposed to the goal of this project which is to focus more on different sentiments throughout the COVID-19 pandemic.

Authors from [3] have done the basic sentiment prediction that is if a tweet is with a positive sentiment or a negative sentiment. The report focuses hugely on combining multiple models like NBSVM, CNN, DistilBERT, BiGRU and FastText and other methods to improve the accuracy in cleaning and pre-processing the tweets and hence bettering the accuracy of the predictions of the sentiments. This project would like to draw inspiration from the authors in testing out multiple models and methods to better the accuracy of the analysis. The team's focus is different from the authors of the report as we would be working and focusing more on sentimental analysis of the tweets and moreover dividing the tweets in broader emotional categories.

2 Proposed Work and Methods

The project's end goal is to determine the changes in the sentiments of the population of the US during different phases of the COVID-19 pandemic using the twitter posts from US. In contrast to the usual sentimental analysis which only predicts the positive, negative and neutral sentiments, the project will try to include a wider range of emotions like Anger, Disgust, Fear, Sadness, Surprise, Joy, Trust, etc. The project will focus on a time frame of March 2020 to March 2021 and further divide it into smaller time intervals in accordance to the changing phases of the COVID-19 in order to better understand the changing emotions of the population of the US based on twitter posts.

This approach will help in analyzing the changes in emotions of people from pre, during and post lockdowns, changes in emotions after vaccine introductions, changes in emotions from ease of restrictions and more. The project will also help determine the changes in trends of social media use during the lockdown and after it.

The project will be doing a text mining of tweets from the twitter API. The data extracted has multiple attributes, some of which will be used in the project, like:

Attribute	**Type**	**Description**
Created_at	String	UTC time when this tweet was created.
Text	String	The content matter of the tweet.
Source	String	Utility used to post the Tweet.
Truncated	Boolean	Indicates whether the value of the text parameter was truncated.
User	User Object	The user who posted this Tweet.
Coordinates	Coordinates	Represents the geographic location of this Tweet
Place	Place	Indicates that the tweet is associated a Place.
Retweeted_status	Tweet	Users can amplify the broadcast of Tweets authored by other users by retweeting.
Retweet_count	Int	Number of times this tweet has been retweeted.
Retweeted	Boolean	Indicates whether the tweet has been retweeted or not.
Lang	String	Indicates a BCP 47 language identifier.

The team will be working in python and utilizing its vast range of libraries in order to achieve the desired outcome. The data will be divided into multiple data tables according to the strategic time interval allocations. The team will be testing multiple models inspired by [3] to do a text classification of the tweets to better the accuracy of the results rather than using the raw textual data of the tweets itself. The textual data will need some cleaning as it can contain emoticons and other abbreviations which can be helpful in recognizing the emotions and hence will be required to be converted into useful meaning using python libraries like "emoji" which helps interpret the emoticon and puts it in simple textual language. Moreover, the team will be using pandas, NumPy and matplotlib to deal with multiple tabular data, do exploratory data analysis and to plot different visualizations respectively. The team will be either using the SENWAVE dataset or use a library "EmotionPredictor" to predict the emotions. By using the SENWAVE dataset we should be able to accurately predict the emotions by training the final decided model on the SENWAVE dataset before implementing it to define emotions of the final dataset with COVID 19 tweets.

The project also aims to analyze the trending topics on twitter regarding COVID-19 during the mentioned time frame. Analyzing the trending topics on twitter on COVID-19 will also help us in understanding the sentiments of people towards those trending topics.

3 Evaluation

Model Accuracy: Overall accuracy of the model in predicting the sentiments.

The project will be focusing on the accuracy of the model on predicting the sentiments. We plan to possibly test multiple models and their combinations and select the highest accuracy model in order to best predict the sentiments. Accuracy is an important part of the evaluation of the results as we will be comparing the sentiments of different phases of the COVID-19 pandemic. To get the best comparison, the accuracy of the model in predicting the sentiments will play an important role.

Model evaluation: Can model predict the correct emotions.

Along with evaluating the Model accuracy, we also need to confirm if it accurately defines the correct emotions. Since the model might be based on text classification in order to predict the correct emotion as defined in the training dataset, it is possible that the model might not take into consideration the "sarcasm" in the tweets and instead of predicting based on the context of the tweet, it might predict based on the words used in the tweets. This will predict wrong emotions for certain tweets and hence flaw the final analysis.

Topic targeted audience: Hottest topic during certain period; time plots, trending topics related to COVID with time series plot.

The aim of the project is to predict the trending sentiments during different phases of the COVID-19 pandemic. In order to achieve this, there needs to be a comparison between the trending topics during a certain phase and generate a causality relation of sentiments with the trending topic of that phase. This will help evaluate the differences between phases of the pandemic and at the same time help us better classify the sentiments of the users as well.

TextCloud: The larger the text, the more frequency of the word. TextCloud before vaccine and TextCloud after vaccination. The TextCloud will be used to get a better general view of the trending sentiments before and after the vaccination announcement. The goal of this analysis is to evaluate the changing sentiments of people from overall negative to positive or maybe the other way around and check if there is even a major difference in the outlook of people from the two phases defined by a breakthrough of humankind in fighting the COVID-19 pandemic.

Correlation: How two emotional attributes are related to each other? The correlation between different emotions of a tweet will help us evaluate the overall sentiment of a particular tweet. It will help us define if two emotions usually occur together. Let's say for example there is a high correlation between "Anxious" and "depressed" that would help us define the possibility of two emotions co-existing in a tweet and better evaluate the overall sentiment of the tweets.

Negative vs Positive emotion: This evaluation will help us distinguish between the two major categories of sentiments. That is "Positive" and "Negative". Understanding the distribution of emotions in a tweet and how much each emotion contributes to the overall negative or positive sentiment will help us evaluate which kind of emotion plays the biggest role in defining a certain sentiment of people in a certain time frame.

4 Milestones

These are our milestones breakdown for our project:

At the beginning, we need to configure the platforms to assist the construction of the fundamental frame. We need to set up the API in the Twitter Developer Website so we can get access token to mine the tweets from Twitter. Next, we want to work on the variables in the data set and clean up the data set. For example, we do not care about user id and username, so we select the column such as likes, full text of tweet, user location, etc. Continuously, as we know, a complete program is not just in technical parts like code reviewing, so improving and enhancing the report understandability to general audience is also considerable. Therefore, making the report more understandable to our audience is our next task. Finally, wrapping up everything we have to then get ready for Demo and Presentation. Evaluating our model for our text mining and propose possible extensions for the future.

Milestone 1	1. Set up the API in Twitter Developer Website. Create the object and repository for transactions. 2. Finish the environment for our project. 3. Implement codes to scrape tweets and prepare some documentation.
Milestone 2	1. Modify the data structure 2. Start to work on the plot function 3. Start to work on the model function
Milestone 3	1. Add new tables to the data structure 2. More features added to the dataset. 3. Publish the GitHub page of documentation.
Milestone 4	1. Compare the model for each feature. 2. Draw conclusions. 3. Analyze the accuracy of the model.
Milestone 5	1. Testing all the functions. 2. Application Instructions (GitHub) 3. Final report. 4. Presentation.

REFERENCES

[1] Chandra R, Krishna A (2021) COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS ONE 16(8): e0255615. https://doi.org/10.1371/journal.pone.0255615

[2] Ahmed MS, Aurpa TT, Anwar MM (2021) Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic. PLoS ONE 16(8): e0253300. https://doi.org/10.1371/journal.pone.0253300

[3] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Somayeh Asadi, U. Rajendra Acharrya, A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets, Knowledge-Based Systems, Volume 228, 2021, 107242, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2021.107242.

[4] Yang Q, Alamro H, Albaradei S, Salhi A, Lv X, Ma C, et al. SenWave: Monitoring the Global Sentiments under the COVID-19 Pandemic. arXiv preprint arXiv:200610842. 2020, https://arxiv.org/pdf/2006.10842.pdf