Sentiments Analysis of Covid-19/Covid-19 Vaccine Tweets

Stuart Finley-10147762, Eissa Khan-20082302, Ivy You-20164288,
Lee Zhang-10179204

Abstract—Since 2019– the start of a global pandemic, the public has engaged more frequently in using numerous social media platforms to express their opinions towards the pandemic and vaccines. Twitter is one of the social media that has been widely used. Due to the ease of crawling tweet data, tweets offer an efficient way of analyzing the public's attitudes towards certain things. Developing a program for sentiment analysis is an approach used to measure people's perceptions towards pandemic and vaccine all over the world. This paper reports on the design of a sentiment analysis pipeline, using a tweets dataset from Kaggle. Predictive models classify twitter users' sentiments into positive and negative, which is represented in a line chart.

Index Terms—Group 8, Health Sciences, Sentiment analysis, coronavirus, covid-vaccines, Application project.

1 Introduction

The goal of our project is to perform sentiment analysis on twitter data. This dataset was taken from Kaggle and is comprised of tweets with the hashtag #CovidVaccine. Covid has irreversibly changed the lives of millions around the world, and continues to induce anxiety, fear and emotional hardships daily. Twitter is a widely used open platform for discourse and communication, and understanding the general sentiment on twitter towards covid might help give insight to the understanding of public opinion towards curbing the spread and increasing vaccine rates. Furthermore, understanding of keywords related to Covid that have negative sentiment and that have positive sentiment could be of value to government officials and health care professionals in determining the best way to communicate with the public.

We set out to answer the following questions:

1. Are sentiments towards covid vaccines related to vaccination rates?
2. How has sentiment towards covid vaccine & the pandemic changed since 2020?
3. Which country has a generally positive sentiment towards covid vaccines and which ones has negative sentiments?

Through selecting the best-performing model, Multi-Layer Perceptron(MLP) classifier, for sentiment prediction and using data visualization techniques using the matplotlib python package, we found that positive sentiments are positively correlated with increases in number of people vaccinated, there have been more negative sentiments towards covid and covid vaccines than positive ones, and that every country has more negative sentiments than positive ones.

2 Related Work

Alam et al.(2021) used deep-learning approaches to unravel twitter users' sentiments towards the vaccination process. They used a similar Kaggle dataset which was obtained using Twitter API, with tweets collected from July 2021 to December 2021. The tweet text data were preprocessed, tokenized and labeled. The Natural Language Processing(NLP) tool Valence Aware Dictionary for sEntiment Reason(VADER) was used for the main sentiment analysis task and found 33.96% positive, 17.55% negative, and 48.49% neutral responses in all the tweets. Long short-term memory(LSTM) and bidirectional LSTM(Bi-LSTM) were used to assess the performance of the predictive models. LSTM assessed an accuracy of 90.59% with Bi-LSTM assessed an accuracy of 90.83%. In addition, the authors utilized a visualization tool WordCloud to visualize the words most commonly associated with each of the three polarities–positive, negative, or neutral.

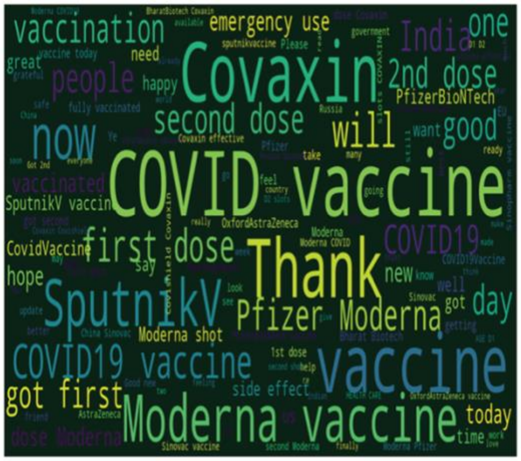


Figure 1. Words associated with positive sentiments

3 Methodology

3.11 Feature Engineering-feature exploration

First, we need to extract feature words from a dataset that has been nicely labeled. In this case, we used the movie dataset from NLTK to find positive and negative words frequency. It works because the words that represent positive and negative sentiments are similar in different texts. To preprocess this dataset, we tokenized the words and removed the punctuation, stop-words, as well as the common name and locations. Then, we calculated the frequency of each word in positive and negative comments, and removed words that are both classified as positive and negative. After these, we extract top 200 frequency positive and negative words from the NLTK movie dataset. We also tried choosing top 50, and 100 words, and using 200 words will have the best performance in the later prediction.

3.12 Feature Engineering-Extracting Features

In this part, we are going to use two methods to extract features. One is using multi-dimensions to value each word, and another is only using the words we explored before. For the first method. We calculated the probabilities of being compound sentences, positive/negative sentences, and the number of positive and negative words in comments. Then, use these five dimensions to value each sentence, and give them a score. For method 2, we only used 400 words that were explored before, and will put them directly into model training.

3.2 Model Building and Selection

In order to choose the best performing model on the sentiment analysis task, we selected 8 models to train using a movie dataset built in the NLTK package. The 8 models we selected include

–Bernoulli Naive Bayes(NB)

–Complement Naive Bayes

–K-Neighbor Classifier

–Decision Tree Classifier

–Random Forest Classifier

–Logistic Regression

–Multi-layer Perceptron(MLP) Classifier

–AdaBoost Classifier

These models classified movie comments' sentiments as either positive or negative. Their performance will be mentioned in the Experiments & Results section.

3.3 Visualization

We used the Matplotlib python package to visualize the changing counts of negative & positive sentiments with respect to the timeline from January 2020 to March 2022. The sentiments are both in response to the coronavirus pandemic, the main vaccines such as Pfizer-BioNTech and Moderna vaccines, as well as the vaccination campaign. The visualization was done separately for method 1 and 2(see figures 3&4 in the results section).

Additional graphs were made counting the number of positive and negative sentiments with respect to the user_location feature(see figures 5&6).

4 Dataset

The dataset is a Kaggle dataset comprised of tweets from 2020 to present that have the #CovidVaccine hashtag created by KASH. The dataset has 380,042 records and consists of 13 different attributes. Of the 13 attributes, four were used in the sentiment analysis and are the user location, the user followers, the data as well as the actual tweet text.

With any text analysis, there is a requirement for preprocessing. First, the text content was tokenized. Tokenization is the process of splitting the content of the tweet into smaller units, which are essentially elements of an array. Next, stop words are removed. The reason for removing stop words is that they are words that don't add information to the sentence. So for the purpose of sentiment analysis, these words provide no value.

It's also important to understand the quality of the data in the dataset. The date column had different formats for the date, so the date needed to be converted to a common format. This process is quite straightforward and involves utilizing pandas pd.to_datetime() function. Having the dates in the same format is especially important when comparing sentiment to vaccination rates as datasets can be merged on the date. There were also values within the dataset that were 'not a number' (NAN). The attributes with the most NAN values are user location, user description and the hashtags. The number of user locations that are NAN are roughly 21% of the entries. As this constitutes a large proportion of the data entries , we will preserve those rows with no user locations in our data analysis.

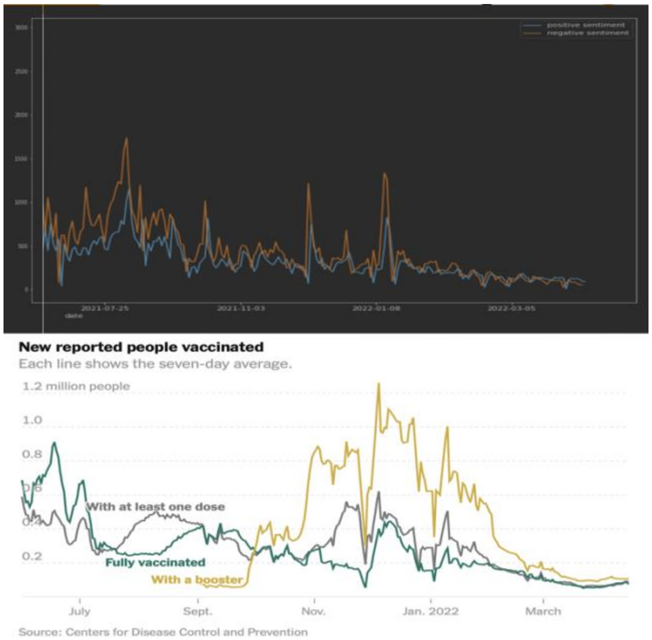
Stemming involves reducing words to their root form, for example, stemming eats, eaten, and eating results in eat. Stemming is a key component of Natural Language Processing(NLP) and functions to reduce the size of the overall data and thereby reducing model runtime and complexity.

5 Experiments and Results

Model	Accuracy of A (top 4)	Model	Accuracy of B (top 4)
Logistic Regression	75.50%	Multinomial NB	79.50%
MLP CLassifier	80.00%	Complement NB	79.50%
AdaBoost Classifier	75.25%	Bernoulli NB	78.50%
RandomForest Classifier	74.25%	MLP Classifier	75.25%

Table 1.
Model Performance.

- *Member 1, 2, 3 &4 are at Queen's University*
  E-mails: 14sf@queensu.ca, 17eak2@queensu.ca, 18yy90@queensu.ca, 14lz22@queensu.ca,

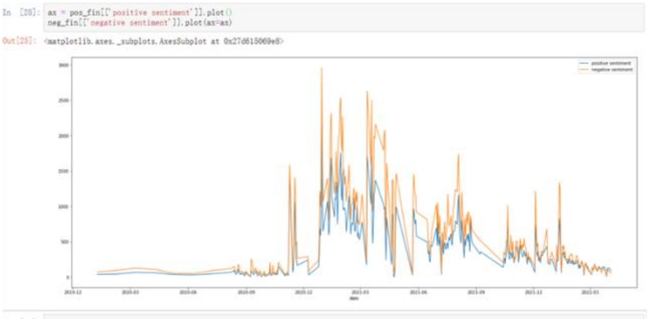


Source: Centers for Disease Control and Prevention

Figure 2: Comparison of the sentiment analysis to the number of vaccinated people in the US.a

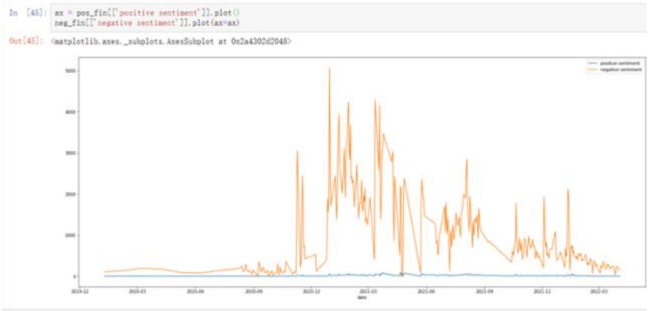
Answer to RQ1: Are sentiments towards covid vaccines related to vaccination rates?

As illustrated in figure 2; our respective results from our models showcase positive and negative sentiment users experienced (y-axis) with noted dates (x-axis) in relation to covid vaccines and vaccinated rates. Assessing the noted results, we are able to establish a distinguished relationship between more users holding negative sentiment towards covid-19 vaccines and thus hindering the outcome of not being vaccinated.



```
In [25]: ax = pos_fin[['positive sentiment']].plot()
         neg_fin[['negative sentiment']].plot(ax=ax)
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x27d615069e8>
```

Figure 3: Graph of the sentiment from the content of the tweets from December 2019 until March 2022 using method 1.



```
In [45]: ax = pos_fin[['positive sentiment']].plot()
         neg_fin[['negative sentiment']].plot(ax=ax)
Out[45]: <matplotlib.axes._subplots.AxesSubplot at 0x2a4302d2048>
```

Figure 4: Graph of the sentiment from the content of the tweets from December 2019 until March 2022 using method 2.

Answer to RQ2: How has sentiment towards covid vaccine & the pandemic changed since 2020?

As illustrated in the above figures 3 & 4; we were able to establish a fair conclusion from our assessed results that negative sentiment towards covid-19 vaccine has overpowered positive sentiment. With both sentiments beginning out to be fairly neutral in early 2020 and the negative sentiment showcasing itself to be very high during the end of 2020 till end of 2021. We can also notice both negative and positive sentiments towards the Covid-19 both once again meet neutrality levels when approaching recent data of March, 2022.

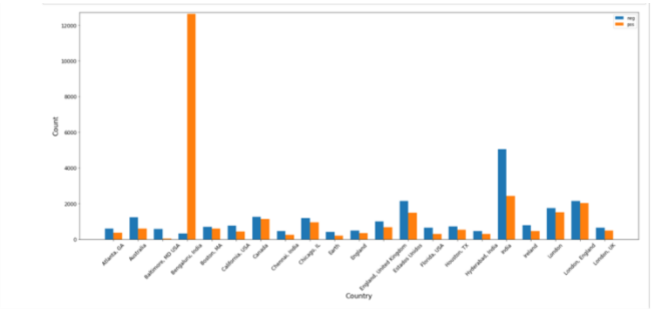


Figure 5: Bar plot of the number of positive tweets and negative tweets by location. Contains the first set of locations

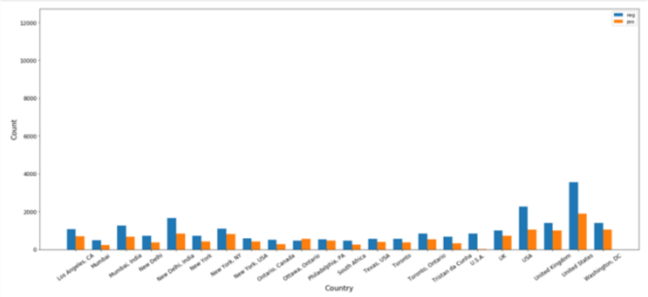


Figure 6: Bar plot of the number of positive tweets and negative tweets by location. Additional locations

Answer to RQ3:Which country has a generally positive sentiment towards covid vaccines and which ones has negative sentiments?

As illustrated in Figure 5 & 6; after determining the results from our models, we were able to conclude that India had the highest positive sentiment towards Covid-19 vaccine and the United States of America having the highest negative sentiment. Our conclusion includes taking an overall average of all the cities and provinces that belong to either India or the USA.

6 Group Member Contributios

Stuart: Provided any required support for coding, prepared slide decks, assisted in building the presentations as well as writings of final reports. Audio recordings for a large portion of the slides.

Eissa: Completed the entire data preprocessing phase, doded and deployed the entire NLTK package with initial models to train and provided graphical and numerical results of selected models.

Ivy: Adjusted codings, provided graphs of results, Methodology and discussion.

Lee: Motivation, Baseline, Methodology, Visualization, Answers to RQs slides, audio recordings for above listed slides in presentation. Related work, Introduction, Methodology, Conclusion and future work in Final report.

7 Replication Package

Zipfile included with submission

8 Conclusion and Future Work

References

[1] Alam,K.N.,Khan,M.S., Dhruba, A.R.Khan,M.M., Al-Amri, J.F., Masud, M., & Rawashdeh, M.(2021). Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data. *Computational and Mathematical Methods in Medicine, 2021.*

[2] Centers for Disease Control and Prevention. COVID Data Tracker. Atlanta, GA: US Department of Health and Human Services, CDC; 2022, April 19. http://covid.cdc.gov/covid-data-tracker

[3] KASH.(2022).*Covid Vaccine Tweet*[Dataset]. Kaggle. https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets