Project Report

Twitter Sentiment Analysis of Tweets about COVID-19 vaccine

Course: Natural Language Processing (COMP 8780)

Instructor: Professor Vasile Rus

Submitted By :

Sameer Neupane

UID: U00696079

Date: 27/4/2021

Introduction

In the field of natural language processing, sentiment analysis has played a crucial role in scientific as well as commercial areas such as recommending systems, data mining, etc. Social media apps such as Twitter, Youtube, Facebook, Reddit contain lots of data from their users which can be used for sentiment analysis in order to comprehend the opinions of people on various topics such as elections, sports, and even COVID 19 in recent days.

Among the various platforms, Twitter has one of the broad adoption and active participation from its users. As of 2021, Twitter has 192 million daily active users and nearly 500 million tweets are sent out every day. The US alone has nearly 70 million active users in a month. Hence Twitter has attracted lots of sentiment analysis research to gain more insights on issues related to public health, business, major events, political decisions, etc.

This current project explores the Twitter Sentiment Analysis (TSA) of the general opinion of people on Twitter regarding the COVID-19 vaccines. As of recently *(1)*, the total number of cases of COVID-19 around the world is more than 124 million and in the US alone, it's nearing 30 million cases. Regarding deaths, worldwide it's about 3 million and in the US, it's more than half a million.

Fortunately, scientists around the world were able to accelerate the development of vaccines and now we have 3 different FDA-approved vaccines *(3)* for emergency use in the US namely Pfizer-BioNTech, Moderna, and Janssen. The wide distribution and inoculation of vaccines to the general population are imperative to get the level of herd immunity needed for us to go back to pre COVID-19 days. However, an article in the Economist *(4)* notes that the increase in vaccine hesitancy in certain sectors of the general population is putting the progress against COVID-19 at risk. Hence TSA can be an avenue to analyze the overall perception of people regarding the COVID-19 vaccines.

Various works *(2,5)* have explored the TSA using keywords such as coronavirus, COVID-19, etc to analyze tweets regarding people's opinions of the pandemic. However, this work involves the COVID-19 vaccines rather than the pandemic itself. This project explores polarity and subjectivity along with the emotional analysis of the tweets of people tweeting about the COVID-19 vaccines along with the three approved vaccines.

Related Works

Various research studies have used Twitter Sentiment Analysis especially to find the user behaviors, user perceptions, etc on various topics or events around the globe. Vast opinion sharing of people on Twitter has increased motivation for researchers to analyze the sentiments of the tweets for applications such as product advertising, movie recommendations, political advertisement, etc. Researchers also have used the TSA to explain and predict product sales, outcomes of elections, and stock market movements. TSA has also been effectively used in various applications such as users' opinions on a product band, presidential debate performances, election results, supreme court decisions, etc. It has also been used to study people's opinions on various natural calamities and global events such as COVID-19. Kaur *(5)* analyzes the sentiments regarding COVID-19. Prabhakar Kaila et al. *(10)* showed that the collected data from Twitter was suitable and worthy to be applied to the experiments, regarding the COVID-2019 outbreak. Rajput, N. K, and et al. *(11)* in their work collected data related to the tweets published during January 2020 and also investigated the tweets corresponding to two main aspects: first comprehending the word occurrence pattern and accordingly the second sentiment recognition. Medford et al. constructed a list of hashtags related to COVID-19 to search for relevant tweets during a two-week interval from January 14th to 28th, 2020.

TSA research mainly has two motivations. The first is to focus on the application of TSA such as to gain insights into various business or social issues, predict key indicators, etc. The second is to advance the state of art research techniques for TSA research so as to increase the accuracy and gain clearer insights regarding the issues of interest.

Textblob is a popular method to do sentiment analysis of tweets. It uses the Naive Bayes model for classification and it is trained on NLTK (Natural Language ToolKit) to detect the valence of the tweets. BB uses the Bayes theorem to predict the sentiment probability of a tweet.

Another work involving Twitter datasets is human emotions detection. Plutchik has defined 8 different human emotions namely *Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust.* NRC Word-Emotion Association Lexicon has 10,170 lexical items which consist of the most frequently used words such as nouns, verbs, adverbs along with 640 words from WordNet Affect Lexicon and 8,132 from General Inquirer. These are used to detect the above-mentioned emotions from the tweets.

Data and Methodology

In this section, the data collection as well as implementation of the sentiment analysis is discussed.

Dataset:

In this work, we utilized the Tweepy python library to extract the tweets from Twitter API. Tweepy allows users to extract tweets based on keywords, hashtags, trends etc. In this work, we have extracted tweets using 4 different keywords. The keywords used and the number of tweets extracted for each keyword are shown below in the table 1.

Tweets with keyword	**Number of tweets**
Covid-19 vaccine	12852
Pfizer-Biontech	2542
Moderna	13096
Janssen	2806
Total Tweets	31296

Table 1 : Twitter Dataset

Methodology:

In this section , the implementation of the sentiment analysis is explained step by step. The flow diagram of the implementation is shown in figure 1.

1. The first step is to establish the connection to Twitter API. For that, we need to have a twitter Dev account and it provides the required access keys in order to gain authentication to connect to Twitter API using the tweepy library.

2. Once the connection is established, Tweepy allows users to extract tweets based on keywords. In this project, four different keywords were used to fetch the tweets. In addition to “COVID-19 vaccine”, keywords with individual vaccine names were also used. The rationale behind using vaccine names was to see if there is any difference in sentiment across the various vaccines.

3. Once the tweets are retrieved, then following preprocessing steps as suggested by the paper *(13)* are followed to sanitize the tweets.

a. First all the tweets are converted to lowercase.

b. URLs, user mentions and hashtags are removed from the tweets as they don't contain any sentiment.

   For example: a tweet *"RT @kottke United States to resume use of Johnson &amp; Johnson COVID-19 vaccine. https://t.co/oViqqGI7Te'" is transformed to "United States to resume use of Johnson & Johnson COVID-19 vaccine."*

c. Stop words are removed from the tweets. The set of stopwords was obtained from nltk corpus.

   For example: a tweet *"what is the point if the science tells us the vaccines are percent effective so if you have a vaccine quite" is transformed to "point science tells us vaccines percent effective vaccine quite"*

d. Punctuations as well as numbers are also removed from the tweets.

e. As also suggested by the paper *(14),* sentiment analysis performance was improved by expanding acronyms and replacing negations. Hence, the acronyms and slangs in the tweets are expanded using the acronym dictionary Internet Slang Dict (17). We have also used a user-compiled contractions dictionary *(15)* to replace the negation.

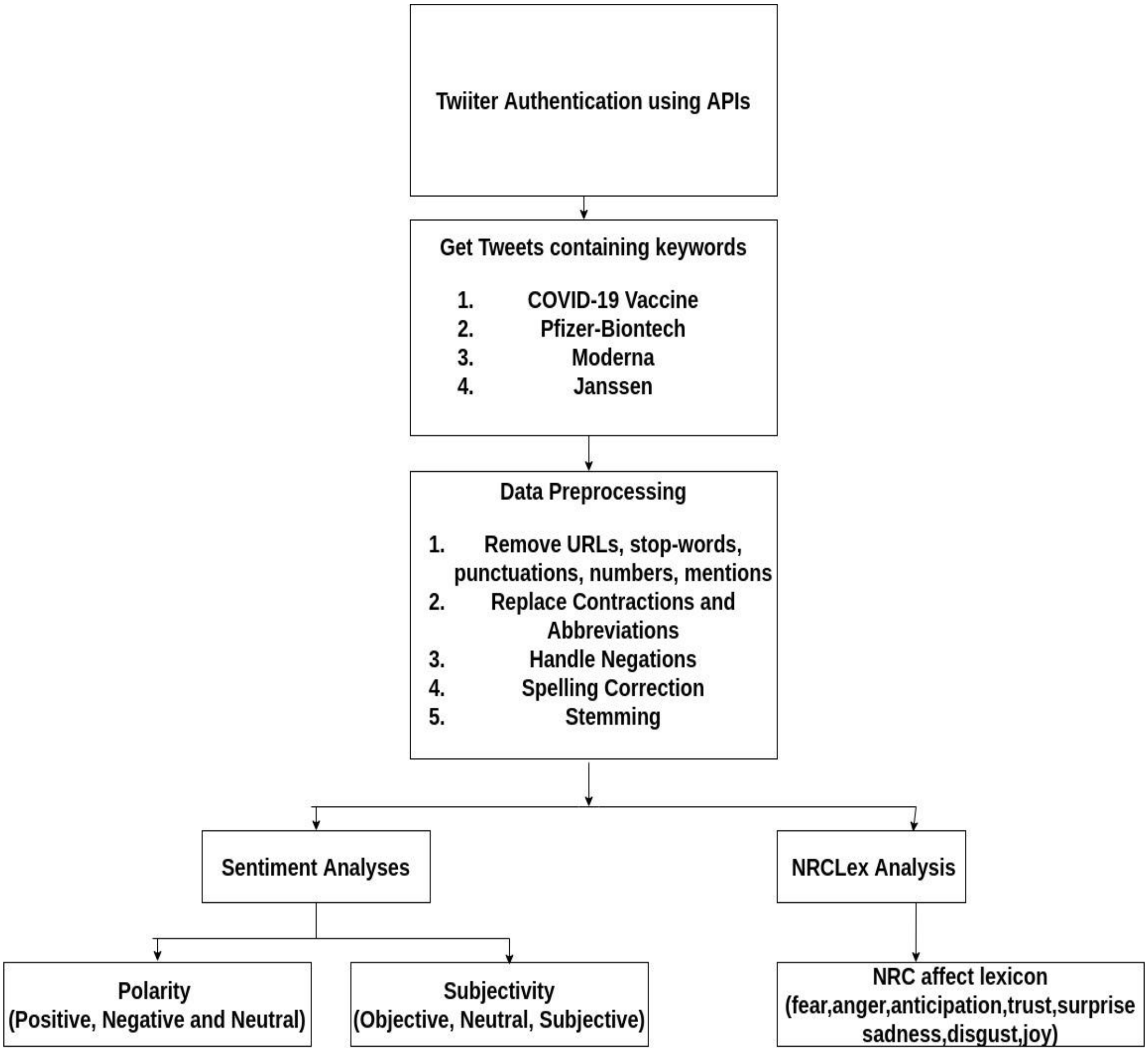
   For example: a tweet *"nyc isn't doing well with vaccinations" is transformed to "New York City is not doing well with vaccinations".*

f. Tweets may contain many spelling errors. In this work, we have use pyspellchecker module in python which is based on Norvig's spelling corrector *(18)*.

   First we check whether the word is present in the english vocab words corpus obtained from nltk. If the words are not present in the corpus, then the spelling checker tries to find the correction for the word.

   Eg: distanci -->distance

g. Finally we made use of Porter Stemmer to stem the words in the tweets. Previous works have suggested that stemming the words yields better results.



Twiiter Authentication using APIs
Get Tweets containing keywords
1. COVID-19 Vaccine
2. Pfizer-Biontech
3. Moderna
4. Janssen
Data Preprocessing
1. Remove URLs, stop-words, punctuations, numbers, mentions
2. Replace Contractions and Abbreviations
3. Handle Negations
4. Spelling Correction
5. Stemming
Sentiment Analyses
NRCLex Analysis
Polarity (Positive, Negative and Neutral)
Subjectivity (Objective, Neutral, Subjective)
NRC affect lexicon (fear,anger,anticipation,trust,surprise sadness,disgust,joy)

Fig 1 : Flowchart of Twitter Sentiment Analysis

4. Textblob is a python library which provides the interface to perform sentiment analysis. According to Textblob documentation *(20)*, It uses the Naive Bayes model for classification and it is trained on NLTK (Natural Language ToolKit) to detect the valence of the tweets. BB uses the Bayes theorem to predict the sentiment probability of a tweet.It returns output for polarity as well as subjectivity of each tweet.

   Polarity values returned are between [-1,1], where -1 corresponds to a negative sentiment and 1 denotes a positive sentiment.

   Subjectivity values lie between [0,1] where -1 corresponds to objectivity and 1 denotes subjectivity. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

In this work, we performed sentiment analysis in the whole tweet corpus as well as tweets extracted using the individual vaccine names.

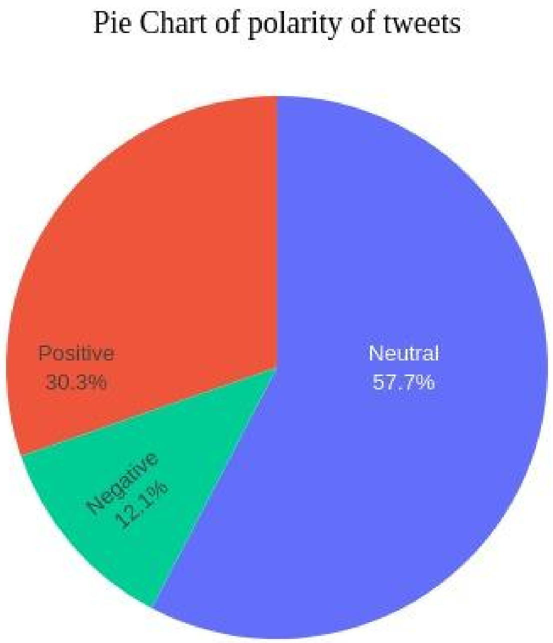
5. NRCLex *(19)* is a python library that measures the emotional effect of a tweet. It has a dictionary which contains approximately 27,000 words, and is based on the National Research Council Canada (NRC) affect lexicon and the NLTK library's WordNet synonym sets. Following emotions are measured by NRCLex:
   - fear
   - anger
   - anticipation
   - trust
   - surprise
   - positive
   - negative
   - sadness
   - disgust
   - joy

Results

In this section, the results of the sentiment (polarity and subjectivity) analyses and NRC emotion analysis is presented.

Sentiment Polarity

Fig 2 shows the polarity of tweets for all the tweets including all the four keywords. Overall, about 30 percent of the tweets showed positive sentiment and around 12 percent of them showed negative polarity. However, overwhelmingly more than 58 percent of the tweets exhibited neutral sentiment. Fig 4 shows the bar chart of the polarity across all the tweets and also tweets involving each keyword. Tweets containing individual vaccines had similar polarities among them. In both the figures, the unit of measurement is percentage.



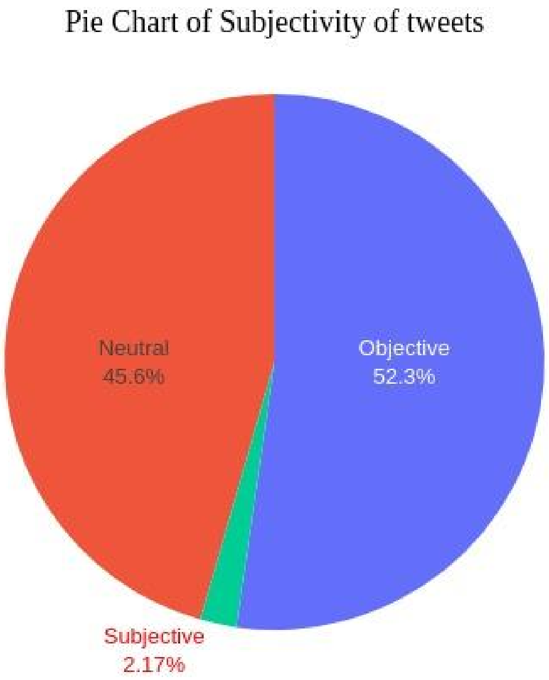


Fig 2. Polarity of Tweets

Fig 3. Subjectivity of Tweets

Subjectivity

In terms of the perspective of people who tweeted, we can see from figure 3 that most of them were objective which was about 52.3 %. It appears that only about 3% of people were being subjective. Rest of the tweets did not display either subjective or objective characteristics. Fig 5 shows the bar chart of the subjectivity across all the tweets and also tweets involving each keyword.

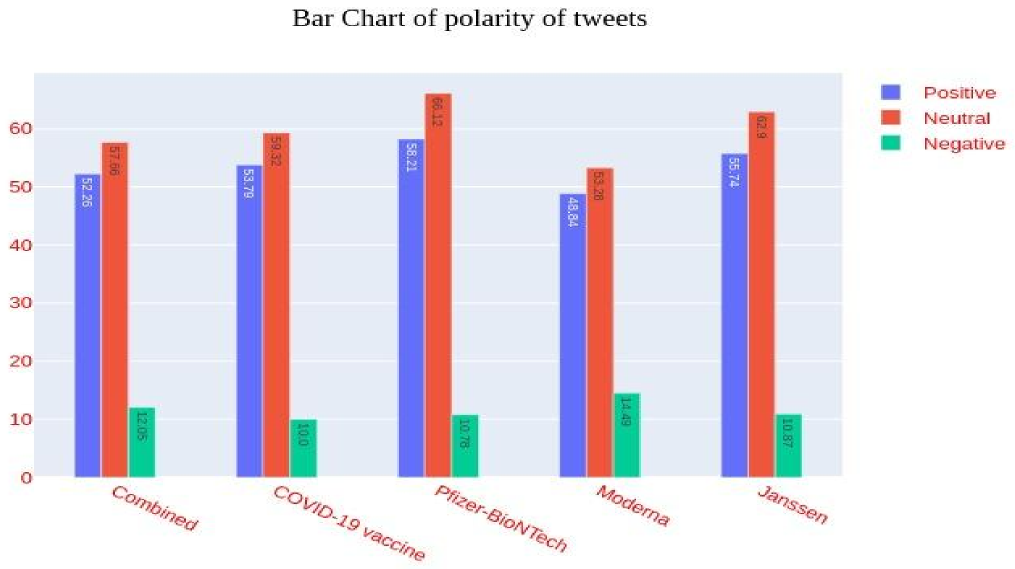


Fig. 4 BarChart comparison of Polarity across tweets of different keywords

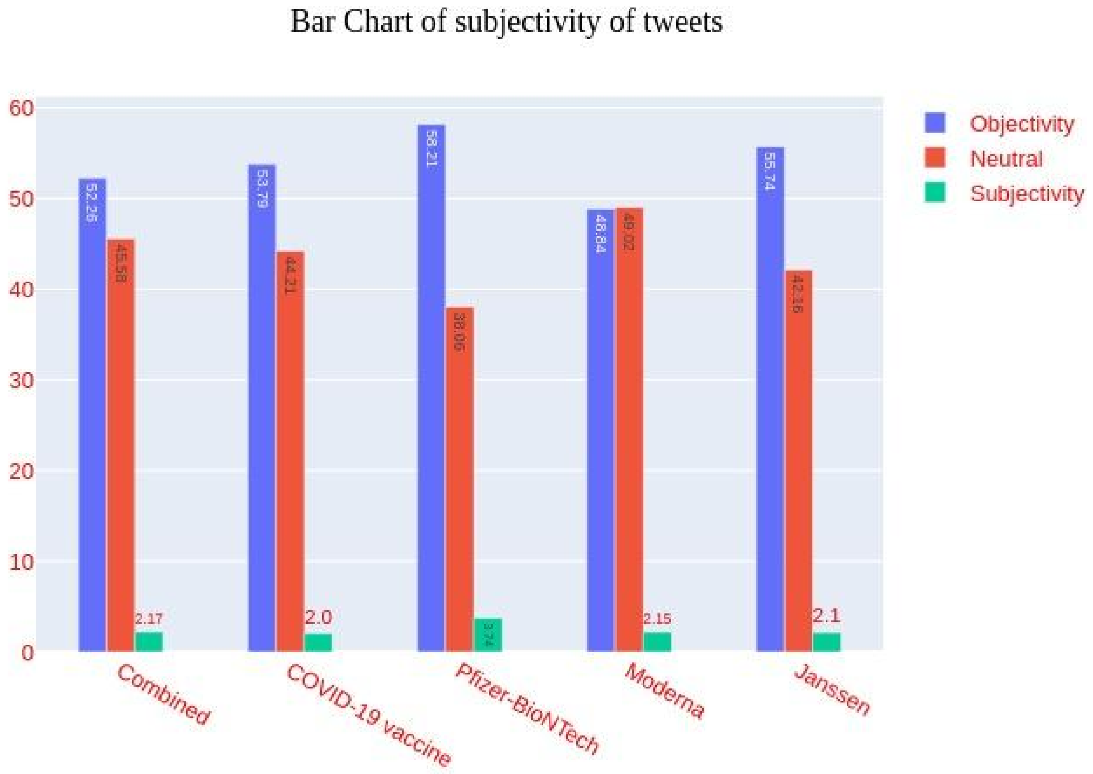


Fig. 5 BarChart comparison of Subjectivity across tweets of different keywords

Word Cloud , Most Frequent Words and Distribution

Then we organized the tweets in word clouds to see what words were most frequently across tweets of different polarities. Fig 7,8,9 and 10 show the word clouds created for the tweets which showed combined, positive, neutral and negative polarities respectively. Fig 6 shows the top 20 most frequent words and their counts in the collected tweets. Vaccine and covid were two most frequently used words.

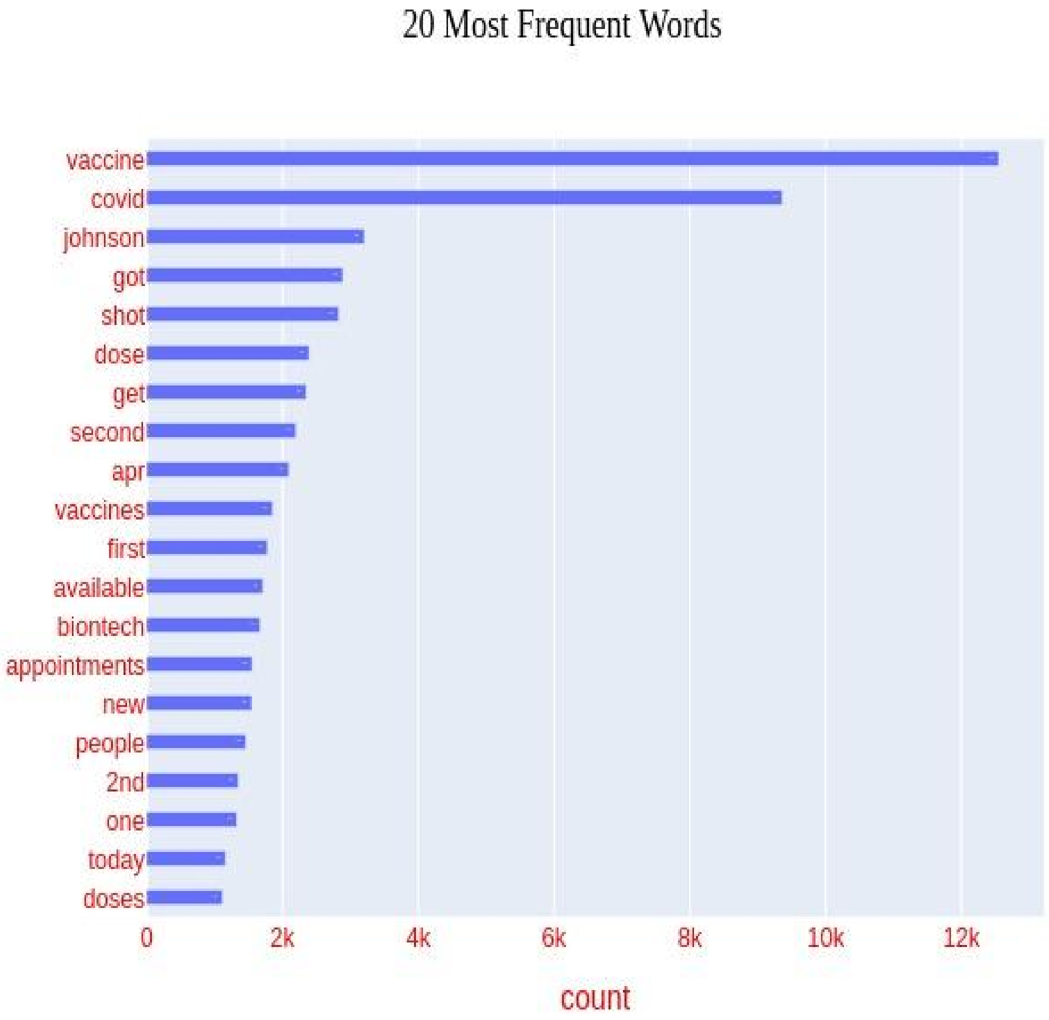


Fig 6. Top 20 Most Frequent Words

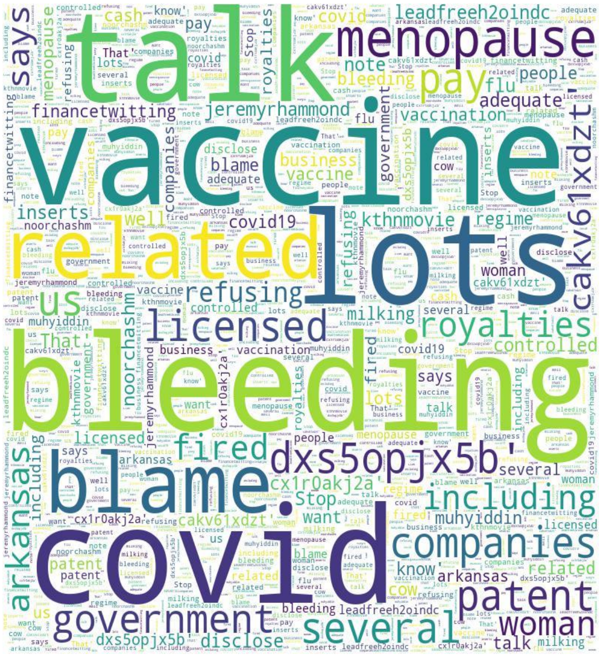


Fig 7. WordCloud-All Tweets

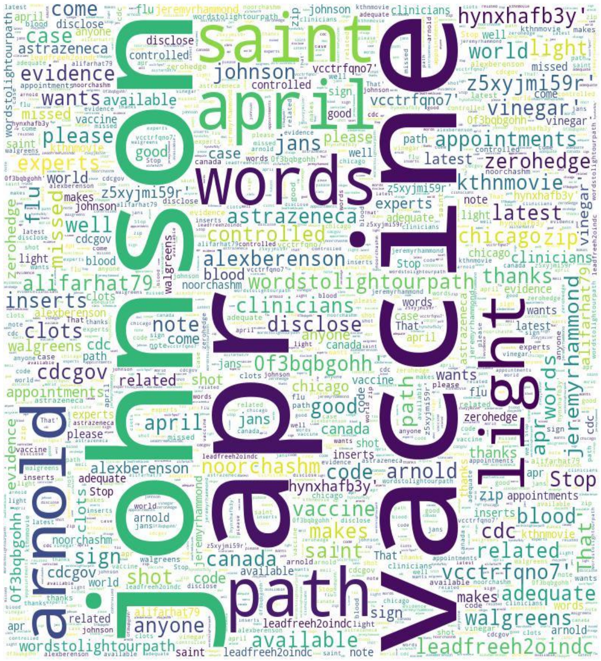


Fig 8. WordCloud-Positive

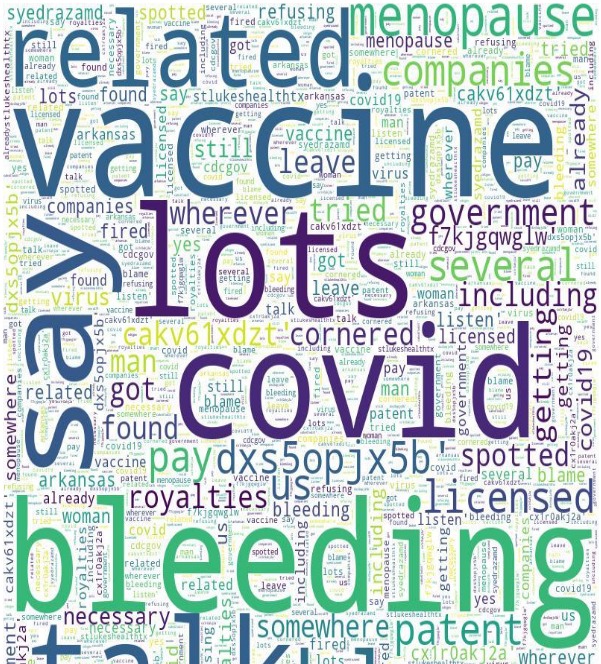


Fig 9. WordCloud-Neutral

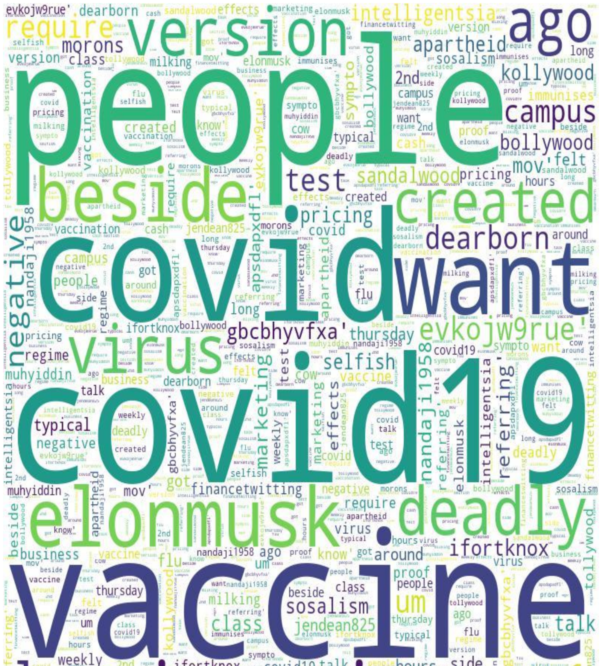


Fig 10. WordCloud-All Tweets

Fig 11 shows the histogram of polarity and subjectivity values of the tweets. Polarity values range from -1 to 1 and subjectivity values range from 0 to 1. To show both of them in the same distribution plot, both of those values are Min-Max normalised. As evident from the figure, we can see that most of the values for subjectivity lie in the lower range and in case of polarity, most of the values lie in the middle range.

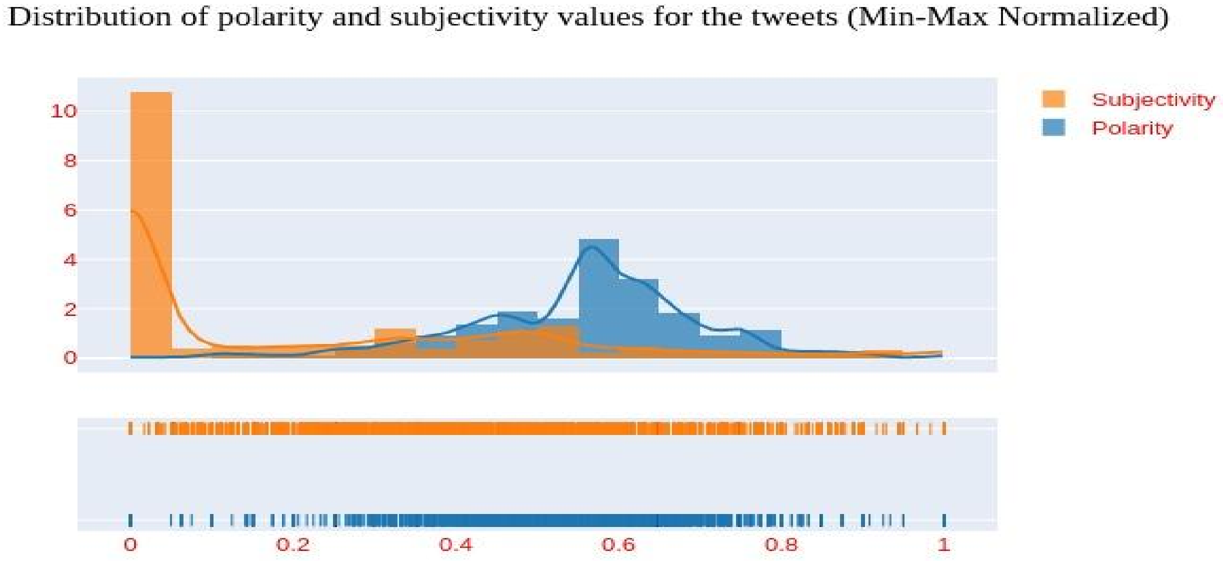


Fig 11. Distribution of polarity and subjectivity values (Min-Max Normalized)

Emotion Analysis

Fig 12. shows the percentage of emotions reactions associated with the extracted tweets. Interestingly, we can see that the two of the most associated emotions were trust and fear. This indicates that people are hopeful of the vaccines as well as some people still have some fear and anxiety regarding the vaccines. Anticipation in the third rank also shows that people are eager to have vaccines available to them. This might also help to explain why so many people are excited to have the vaccines available and also at the same time, many people have reservations about them getting vaccinated. The lowest emotion associated was disgust. Overall, we can see that people are optimistic about the COVID-19 vaccines.

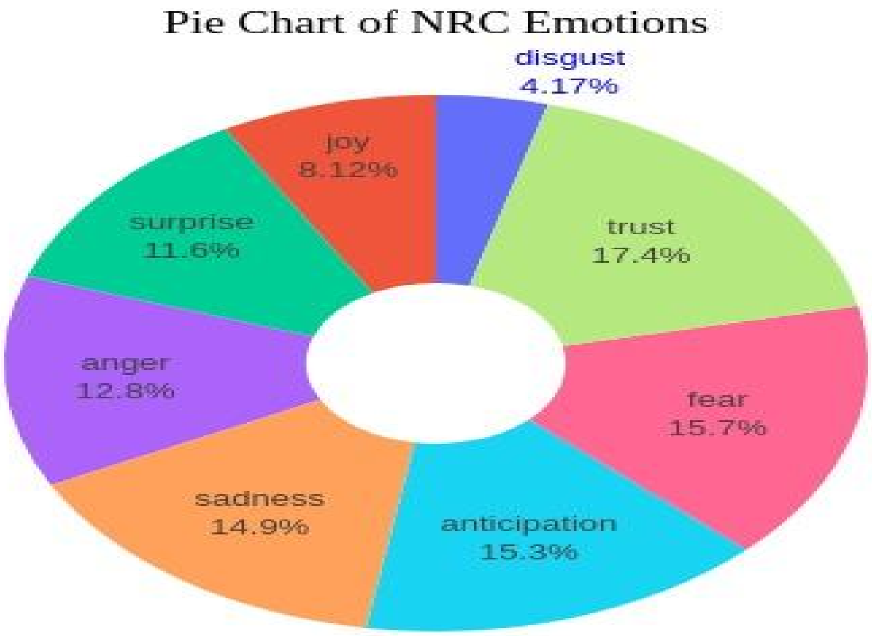


Fig 12. Pie Chart of Emotion Distribution

Conclusion

This work aimed to analyze the sentiment analysis across the tweets containing the keywords associated with the COVID-19 vaccines. Tweepy was used to collect the tweets and Textblob library was used to perform sentiment analysis. Such analysis especially regarding any new developments such as COVID-19 vaccines can enable governments or health agencies to raise awareness among the people so that it can have a positive impact on society. For further works, it would be interesting to advance this work by performing similar analysis across various countries and also tweets of different languages.

References:

1. https://coronavirus.jhu.edu/map.html
2. K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, “Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks”, *Kurdistan Journal of Applied Research*, vol. 5, no. 3, pp. 54-65, May 2020.
3. https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines
4. https://www.economist.com/briefing/2021/02/13/vaccine-hesitancy-is-putting-progress-against-covid-19-at-risk
5. C. Kaur and A. Sharma, "Twitter Sentiment Analysis on Coronavirus using Textblob," EasyChair2516-2314, 2020
6. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation (https://dl.acm.org/doi/abs/10.1145/3185045#sec-ref)
7. https://www.oberlo.com/blog/twitter-statistics
8. Plutchik R. The Nature of Emotions. American Scientist. 2001;89(4):344.
9. https://kjar.spu.edu.iq/index.php/kjar/article/view/512
10. D. Prabhakar Kaila, D. A. J. I. J. o. A. R. i. E. Prasad, and Technology,"Informational Flow on Twitter–Corona Virus Outbreak–Topic Modelling Approach," vol. 11, no. 3, 2020
11. N. K. Rajput, B. A. Grover, and V. K. J. a. p. a. Rathi, "Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic," 2020.
12. R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. J. m. Lehmann,"An" Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 outbreak," 2020.
13. https://stackoverflow.com/questions/19790188/expanding-english-language-contractions-in-python
14. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7862202
15. https://medium.com/coinmonks/remaking-of-shortened-sms-tweet-post-slangs-and-word-contraction-into-sentences-nlp-7bd1bbc6fcff
16. https://www.sciencedirect.com/science/article/pii/S0957417418303683
17. https://www.noslang.com/dictionary/
18. http://norvig.com/spell-correct.html
19. https://pypi.org/project/NRCLex/
20. https://textblob.readthedocs.io/en/dev/index.html