Quest Journals
Journal of Software Engineering and Simulation
Volume 7 ~ Issue 9 (2021) pp: 12-24
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Public Insights on COVID19 Vaccination using Exploratory and Sentiment Analysis on tweets

Aryan Srivastava[1], Debarati Das[2]
(The Shri Ram School, Aravali)[1]
(Debarati Das, University of Minnesota, Twin Cities)[2]
Corresponding Author: Aryan Srivastava

ABSTRACT : *COVID-19, a major topic of discussion in every field, has shaken up the entire world. It is at times like these that social media blows up and the activity on these platforms drastically increases. In the past one year, there have been a lot of tweets related to Corona Virus and its Vaccines. This paper focuses towards exploring innumerable tweets regarding covid vaccines to summarize the public's opinion and discover key inferences from this data-set. This data-set pertains to the tweets related to the following vaccines:*

- *Pfizer/BioNTech*
- *Sinopharm*
- *Sinovac*
- *Moderna*
- *Oxford/AstraZeneca*
- *Covaxin*
- *Sputnik V*

Through the course of this paper, you will find word clouds, time series plots, histograms etc to show important features of this data-set like retweet counts with time, average text length, countries from where tweets have come, and many more. I have also carried out basic sentiment and emotional analysis on the tweets (using lexical features) to display the public's feedback about COVID 19 Vaccines. The objective of this paper is to eliminate the problem of twitter users, browsing through thousands of tweets to summarize the public's opinion about the covid vaccines to make their personal decisions about related situations.

KEYWORDS : *Covid 19 vaccine, sentiment, Twitter, emotion, tweets, vaccination, opinion*

Received 14 September, 2021; Revised: 27 September, 2021; Accepted 29 September, 2021 © The author(s) 2021. Published with open access at www.questjournals.org

I. INTRODUCTION

The novel coronavirus disease (COVID 19), which started on 31st December, 2019, is still in the air and has shaken up the entire world, attacking its health system and economy to its core. As of now there are approximately 200 million active covid cases and 4 million deaths, but in the past few months there have been some major developments in terms of vaccinations and handling the covid pandemic. For the same, many people have tweeted about the different covid vaccinations, namely Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. There are approximately 125 thousand tweets recorded till 8th August 2021 for these vaccinations. The challenge any person would face is how to summarize and form a general analysis of these tweets to summarize general feedback of the public to these vaccines. Through the course of this paper, we tackle this problem by using NLP (natural language processing), NRCLex (Lexicons), Machine Learning and data visualisations, analysing minute features of this data-set to carry out key inferences from the general public about these vaccines. The motivation behind the paper is to eliminate the problem of twitter users browsing through thousands of tweets to summarize the public's opinion about the covid vaccines to make their personal decisions about related situations.

II. MATERIALS AND METHODS

2.1 DATA-SET :

In this study, I have incorporated data from the" COVID-19 All Vaccines Tweets" data-set that pertains to the tweets related to COVID-19, different Coronavirus Vaccines, and lockdowns etc. Thisdata-set has been recorded and provided by Gabriel Preda on Kaggle who has been logging tweets from 20th December, 2021 till today. The data is available at https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets. This data-set consists of the following fields:

- Id (Numerical Id provided by twitter)
- User Name
- User Location (Location from where tweet has been sent)
- User Description (User's bio)
- User Followers (Number of followers on twitter)
- Users Friends (Number of friends on twitter)
- User favourites (Number of favourites on twitter for user)
- User Verified (Is the user verified?)
- Date (Date of tweet)
- Text (Text in tweet)
- Clean text (Pre-processed Text)
- Hashtags
- reTweets (Number of retweets)
- Favourites (Number of favourites for tweet)

For the purpose of this paper, we have used the following fields:

- User Name
- User Location (Location from where tweet has been sent)
- Date (Date of tweet)
- Text (Text in tweet)
- Clean text (Pre-processed Text)
- Hashtags
- reTweets (Number of retweets)

Additionally, the vaccines that this data-set focuses on are

- Pfizer/BioNTech
- Sinopharm
- Sinovac
- Moderna
- Oxford/AstraZeneca
- Covaxin
- Sputnik V

2.2 LIBRARIES:

- **Pandas:** For the purpose of this paper, this library has been used to convert the data-set into a dataframe, which makes data manipulation and analysis easier.
- **NLTK (Natural Language Toolkit):** This library is used for computational linguistics and provides easy-to-use interfaces to over 50 corpora and lexical resources. For the purpose of this paper, this library has been used for pre-processing of text (stemming, lemmatization, and removal of stop words), tokenization, sentiment analysis (NLTK Sentiment Intensity Analyser), and Emotional Analysis (Lexicon based).
- **Matplotlib:** This is a library used for making static, animated and interactive data visualizations. For the purpose of this paper, this library has been used for plotting time series graphs and histograms.
- **Wordcloud:** This is a library used for representing text data in which the size of each word indicates its frequency or importance.

2.3 METHODS AND DATA VISUALIZATION TECHNIQUES:

2.3.1 TOKENIZATION:

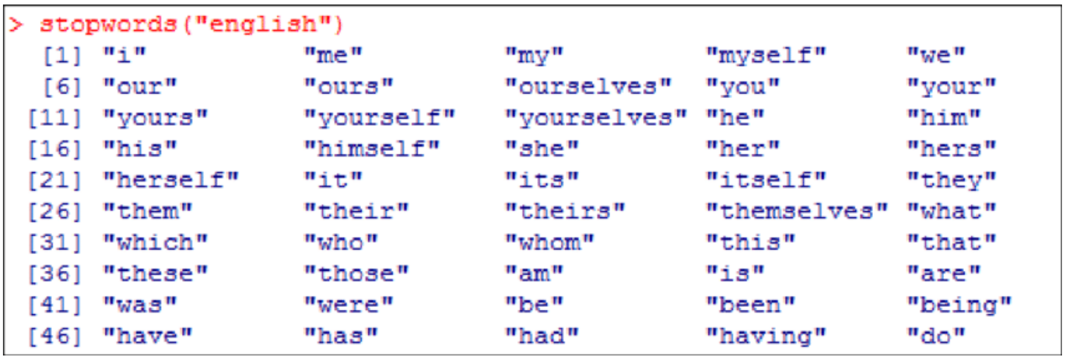
Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords. In this project I have used NLTK’s treebank tokenizer to execute word level tokenization on tweets.

2.3.2 PRE-PROCESSING:

Text Pre-processing is used to clean the text data and remove the unnecessary features like stopwords, hashtags, and emojis etc so that the text is ready to be fed to the model. For the purpose of this paper, we can categorize our pre-processing function into two parts:

STOPWORDS:

Stop words are a set of commonly used words in a language. Here is a list of generalized stopwords which have been used in this paper:

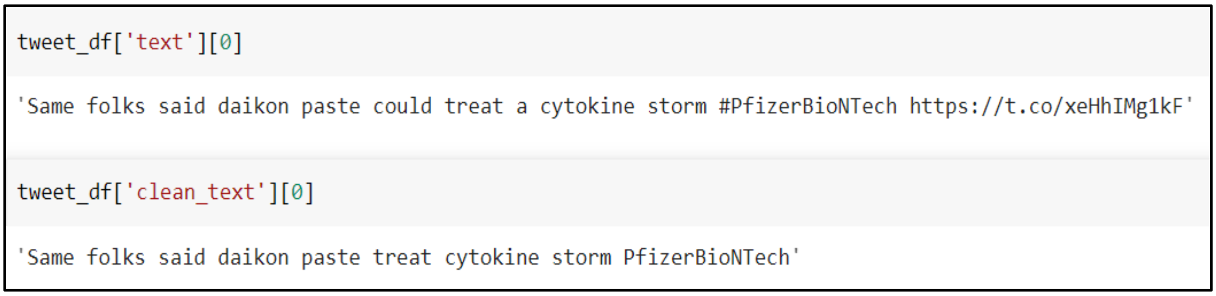


```
> stopwords("english")
 [1] "i"          "me"          "my"          "myself"      "we"
 [6] "our"        "ours"        "ourselves"   "you"         "your"
[11] "yours"      "yourself"    "yourselves"  "he"          "him"
[16] "his"        "himself"     "she"         "her"         "hers"
[21] "herself"    "it"          "its"         "itself"      "they"
[26] "them"       "their"       "theirs"      "themselves"  "what"
[31] "which"      "who"         "whom"        "this"        "that"
[36] "these"      "those"       "am"          "is"          "are"
[41] "was"        "were"        "be"          "been"        "being"
[46] "have"       "has"         "had"         "having"      "do"
```

I have also added a few stopwords like “amp”, “ed”, “https”, "00B8", "00BD”, and “00A0" that are unnecessary only to this data-set that are present in numerous tweets. In our pre-processing function we have removed these stopwords from each tweet to simplify our data that is to be fed to our sentiment analysis model.

LEMMATIZATION:

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form. In our pre-processing function we have used lemmatization to reduce the complexity of the data-set.For example, the words organize, organized, organization, organizer, and all other forms of the word organize will turn into the word organize after lemmatization. After applying all these techniques, this is how the text changes after pre-processing:

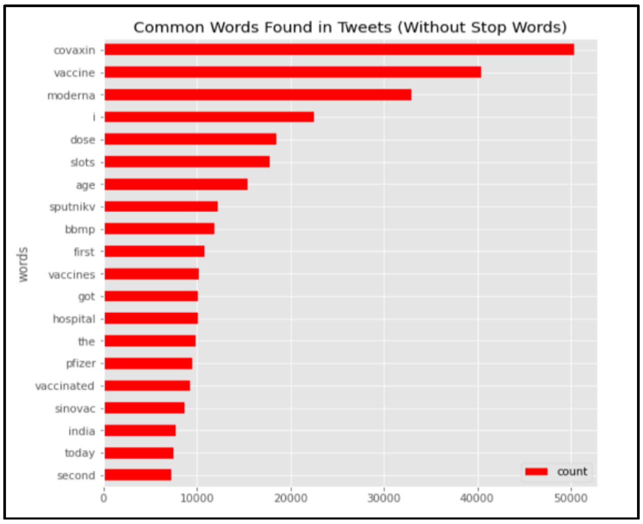


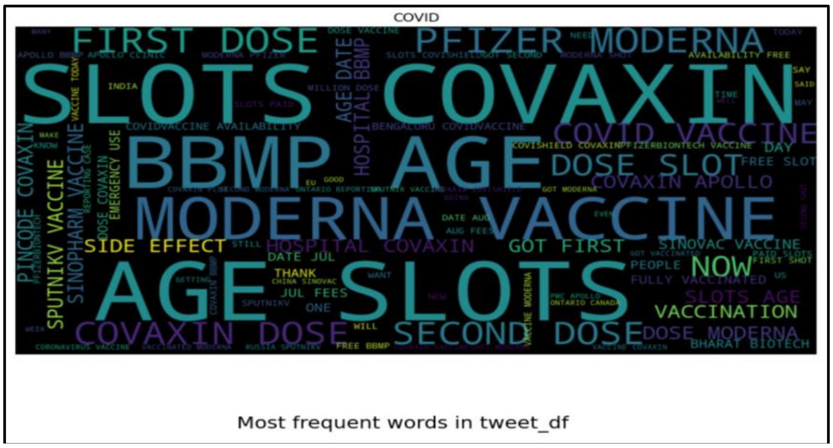
```
tweet_df['text'][0]

'Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF'

tweet_df['clean_text'][0]

'Same folks said daikon paste treat cytokine storm PfizerBioNTech'
```

2.3.3 WORDCLOUD:

A word cloud is a simple yet powerful visual representation object for text processing, which shows the most frequent word with bigger and bolder letters, and with different colors. The smaller the size of the word the lesser it's important. For the purpose of this paper, I have made wordclouds on the on the most frequent words from worldwide tweets, and tweets from India, China, US, and UK.

For example, when we look at covid vaccination related tweets worldwide, words like Covaxin, vaccine, Moderna, dose and slots etc have a high frequency in the data-set and therefore in our wordcloud they would have a larger size. In the figures below you would find the top 20 most frequent words in the worldwide covid vaccination tweet data-set and its corresponding wordcloud with top 100 most frequent words.



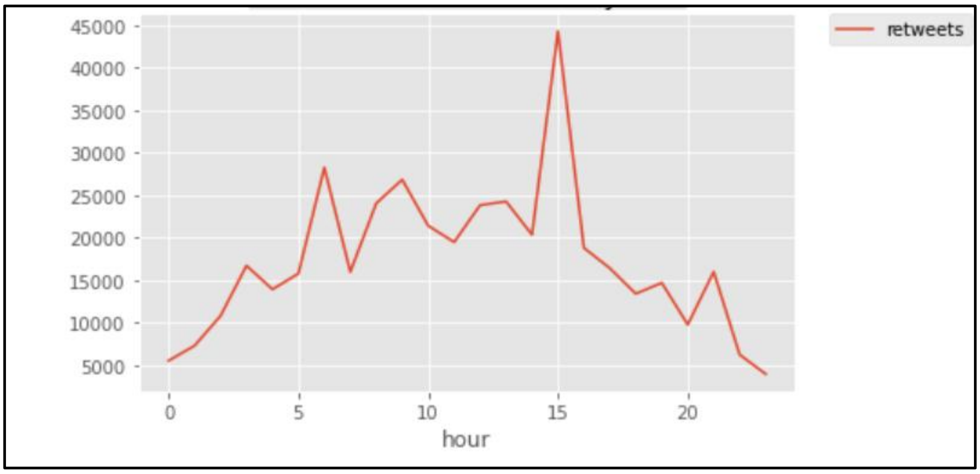


Most frequent words in tweet_df

2.3.4 TIME-SERIES PLOTTING:

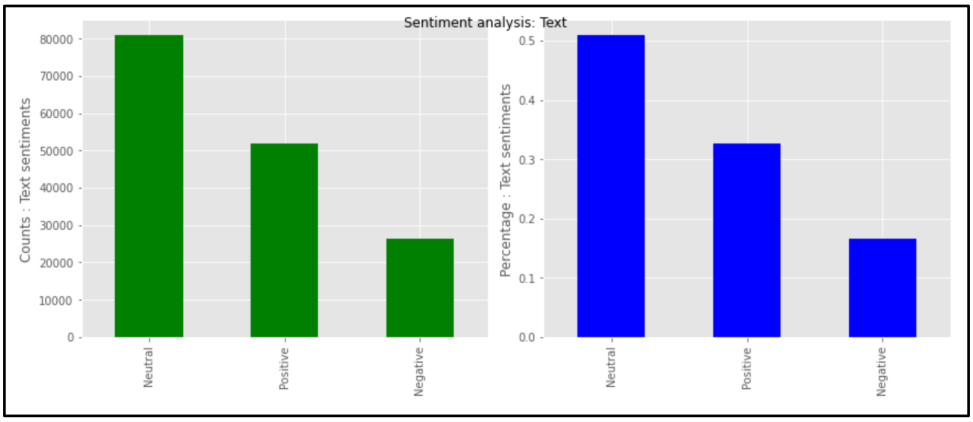
Time series graphs can be used to visualize trends in counts or numerical values over time. Because date and time information are continuous categorical data (expressed as a range of values), points are plotted along the x-axis and connected by a continuous line.

For the purpose of this paper, I have tracked features like number of retweets per hour, per day, and per minute, average length of tweets per hour, and sentiments and emotions of tweets per day. For example,



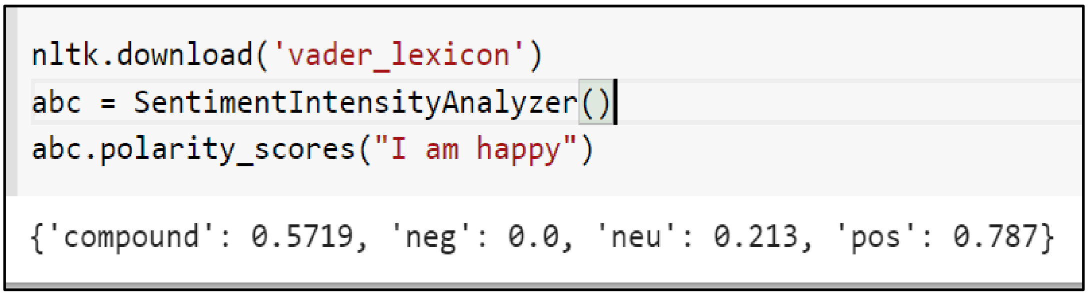
2.3.5 HISTOGRAMS:

In statistics, a histogram is representation of the distribution of numerical data, where the data are binned and the count for each bin is represented. For the purpose of this paper, we have used histograms to show word counts, distribution of tweets over the seven days of the week, distribution of tweets based on locations, and the percentage distribution of tweets based on sentiments and emotions. For example,



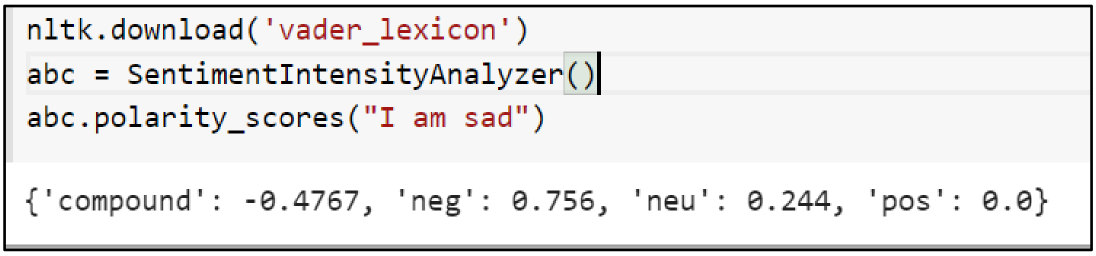
2.3.6 Sentiment Intensity Analyzer:

For the purpose of this paper, I have used the VADER (Valence Aware Dictionary and sentiment Reasoner) sentiment analysis tool provided by NLTK. Specifically, I used the polarity scores method that returns the compound (aggregate), positive and negative (extent of positivity/negativity) score of the text input, on a scale of -1 to +1, based on the lexical features (word based) of the input. In this project, I have only used the compound score to categorize each tweet under the categories positive, negative, and neutral. If the score is greater than 0, then it is positive, if less than 0, then negative and if it is equal to 0, then neutral.
For example, for the sentence "I am happy." we get the following scores,



```
nltk.download('vader_lexicon')
abc = SentimentIntensityAnalyzer()
abc.polarity_scores("I am happy")

{'compound': 0.5719, 'neg': 0.0, 'neu': 0.213, 'pos': 0.787}
```

And for the sentence" I am sad." we get,



```
nltk.download('vader_lexicon')
abc = SentimentIntensityAnalyzer()
abc.polarity_scores("I am sad")

{'compound': -0.4767, 'neg': 0.756, 'neu': 0.244, 'pos': 0.0}
```

Although, this method is not fully accurate as if the word mentioned in the input is beyond the scope of VADER's lexical dictionary, then our method would categorize it as neutral.

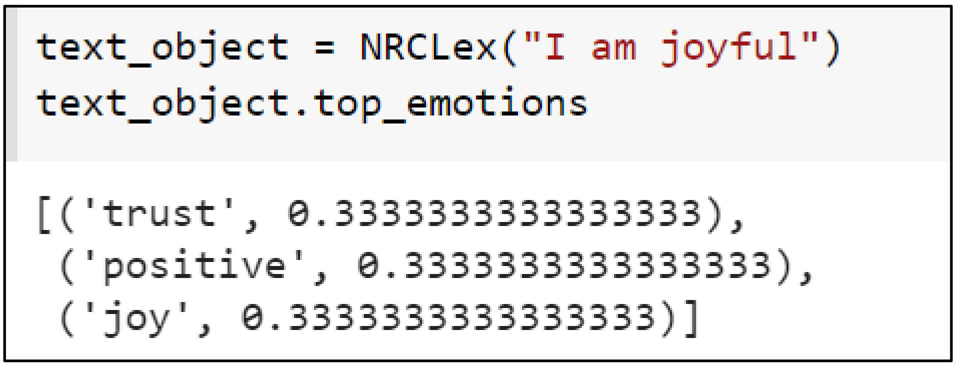
2.3.7 Emotion Recognition:

In this project, for emotion recognition, I have used the NRCLex library that allows us to measure the emotional effect from the body of the text based on lexical features. I have used the top emotions method of this library to recognize top emotions of the input. The emotions provided by the NRCLex library are as follows:

- Fear
- Anger
- Trust
- Surprise

- Sadness
- Disgust
- Joy

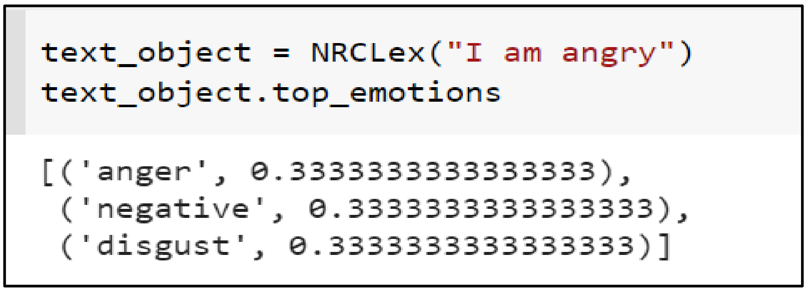
For example, for the sentence “I am joyful” we get the following emotions,



```
text_object = NRCLex("I am joyful")
text_object.top_emotions
```

```
[('trust', 0.3333333333333333),
 ('positive', 0.3333333333333333),
 ('joy', 0.3333333333333333)]
```

And for the sentence “I am angry” we get the following emotions,



```
text_object = NRCLex("I am angry")
text_object.top_emotions
```

```
[('anger', 0.3333333333333333),
 ('negative', 0.3333333333333333),
 ('disgust', 0.3333333333333333)]
```

However, even for this method there are limitations, as this algorithm would also not be able to detect emotions for words beyond the lexical dictionary used by NRCLEx.

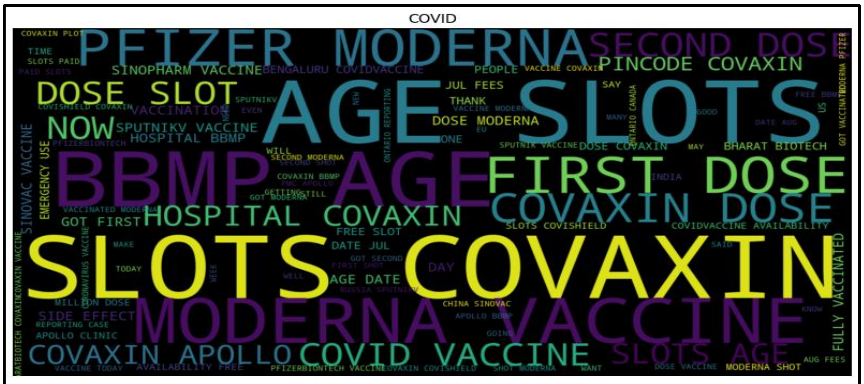
III.RESULTS AND FINDINGS

To carry out inferences from the vast number of tweets in our data-set, I have plotted the following graphs and figures for data visualization and interpretation -

I. Word Clouds
   a. Most frequent words in Worldwide tweets
   b. Most frequent words in Indian tweets
   c. Most frequent words in American tweets
   d. Most frequent words in tweets from UK
   e. Most frequent words in tweets from China
II. Time-series plot
   a. Number of retweets by hour
   b. Number of retweets by minute
   c. Number of retweets by day of year
   d. The average length of tweets by hour
III. Histograms
   a. Number and percentage of tweets by day of week
   b. Number and percentage of tweets by user location
IV. Sentiment Analysis
V. Emotion Recognition

3.1 Word Clouds:

3.1.1 Worldwide tweets:

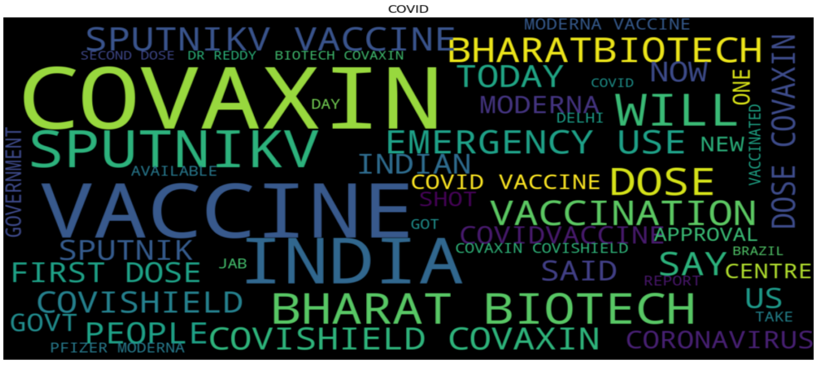


Most frequent words in tweet_df (worldwide)

From the word cloud above, we can make the following inferences,

- Vaccines like the Pfizer/BioNTech, Covaxin, Moderna, Sputnik V, and AstraZeneca are being talked about the most. A potential reason for this could be that these are the vaccines being used most by the public. According to CNBC TV, Pfizer/BioNTech, Covaxin, Moderna, Sputnik V, and AstraZeneca are a part of the top 8 vaccines used in the world. This verifies our analysis.
- Additionally, topics like vaccine slots, first dosage and second dosage, age, and hospitals are frequent topics of discussions, as these are some of the leading issues and doubts the users of the vaccine face before they proceed to take it.

3.1.2 Tweets from India:

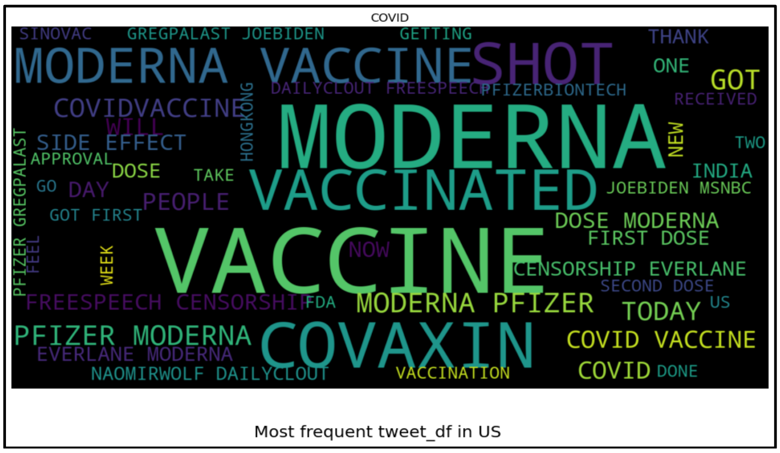


Most frequent tweet_df in India

From the word cloud above, the following inferences can be made:

- First of all, Covaxin and Covisheild were the two most used vaccines in India and therefore are one of the more prominent words in the word cloud. Additionally, vaccines like Moderna and Sputnik V have been used frequently and have been a topic of discussion in Covid Vaccines related tweets in India
- Also, Bharat Biotech and the word “Government " have been mentioned in this data-set quite frequently, probably because Bharat Biotech is a mass manufacturer of India’s prominent vaccine - Covaxin and the government tackles with the distribution and and supports the development of these vaccines.
- Additionally, topics like vaccine slots, first dosage and second dosage, age, and hospitals are frequent topics of discussions, as these are some of the leading issues and doubts the users of the vaccine face before they proceed to take it.

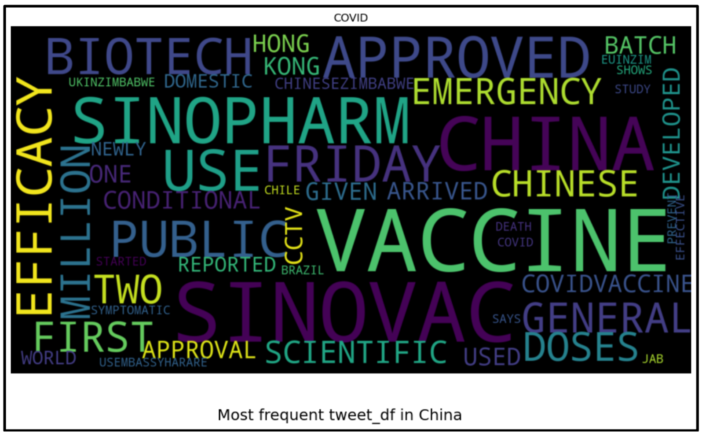
3.1.3. Tweets in USA:



Most frequent tweet_df in US

- This word cloud suggests that Pfizer and Moderna Vaccines were the most frequently used ones as they are very frequent in this data-set.
- Additionally, the president of the United States, Joe Biden has also been a topic of discussion as he has been directly involved in handling the covid pandemic.
- Also, topics like vaccine slots, first dosage and second dosage, age, and hospitals are frequent topics of discussions, as these are some of the leading issues and doubts the users of the vaccine face before they proceed to take it.

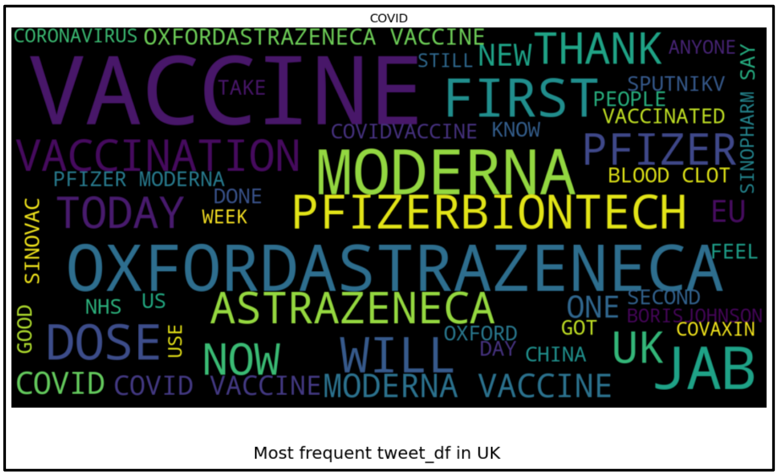
3.1.4 Tweets in China:



Most frequent tweet_df in China

- This word cloud suggests that the vaccines used most in China are Sinovac Vaccines and are mass manufactured by Sinopharm as the word “Sinovac” and “Sinopharm” are repetitive in the Chinese tweets.
- Additionally, topics like vaccine slots, first dosage and second dosage, age, and hospitals are frequent topics of discussions, as these are some of the leading issues and doubts the users of the vaccine face before they proceed to take it.

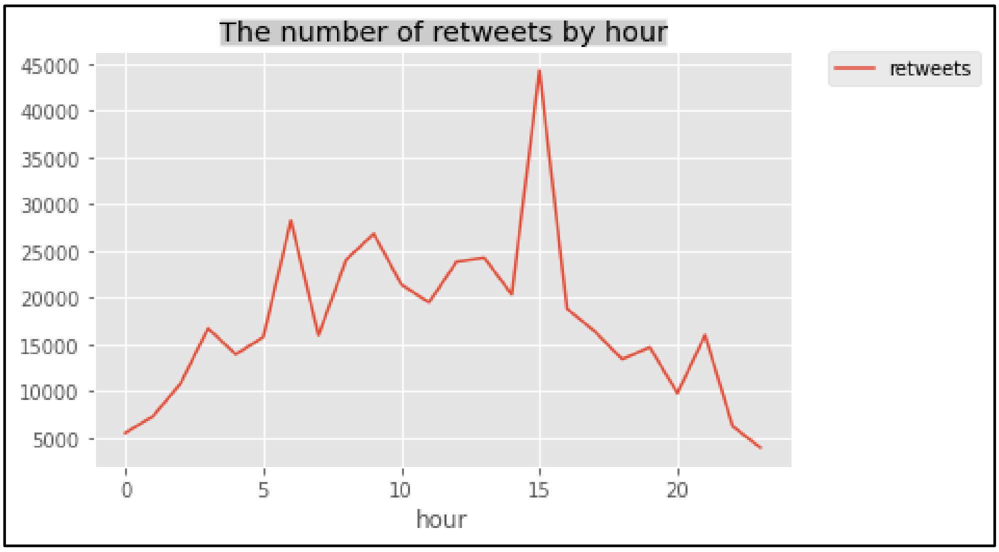
3.1.5 Tweets in UK:

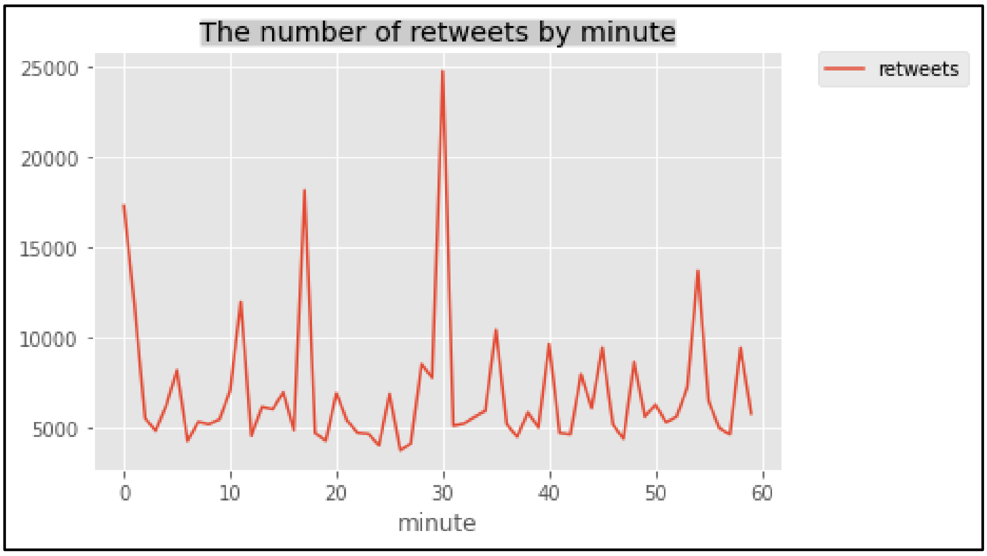


- This word cloud tells us that vaccines like AstraZeneca, Pfizer, and Moderna have been used the most as they are quite prevalent in this data-set.
- Additionally, a unique phrase “Blood Clots” is unique to this word cloud suggestive of a formation of blood clots after receiving the vaccine dosage in the UK.
- Also, topics like vaccine slots, first dosage and second dosage, age, and hospitals are frequent topics of discussions, as these are some of the leading issues and doubts the users of the vaccine face before they proceed to take it.

3.2 Time - Series Plots:

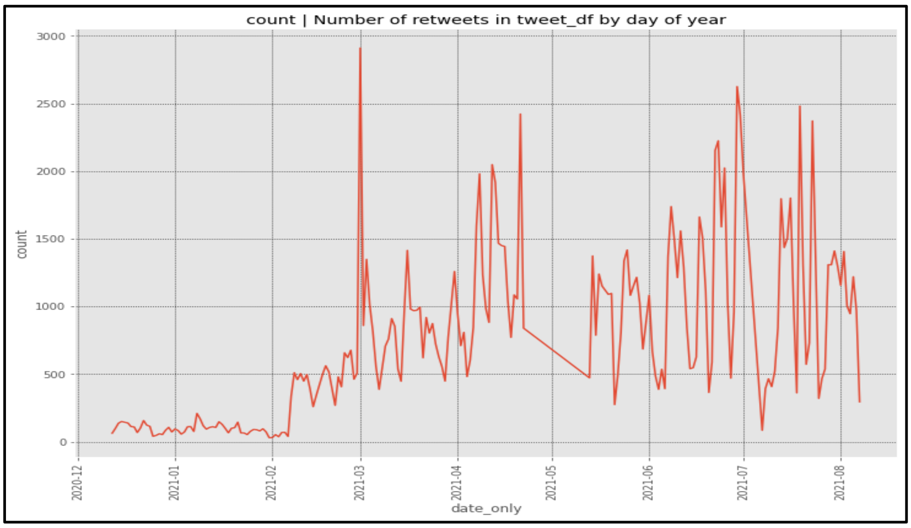
3.2.1 Numberof retweets by hour and minute:





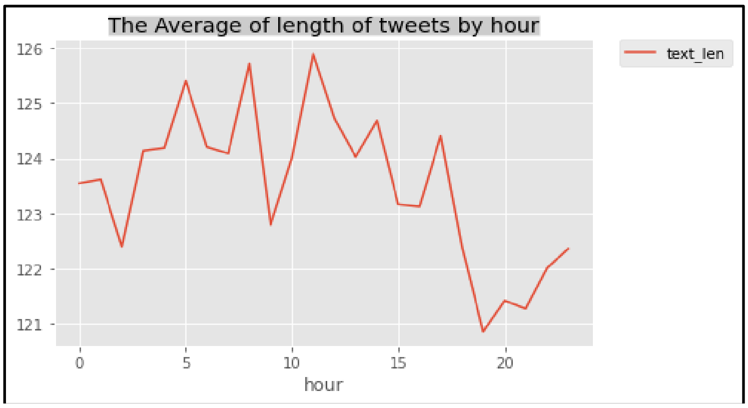
- The above plots show that the maximum number of retweets were observed at the 15th hour (3:00PM) of each day and at the 30th minute of each hour.
- First of all, the precise time of maximum activity on twitter about covid vaccines is around 3:30PM, suggested by these graphs.
- This may be because this time is close to the lunch time for most people where people browse their social media to tweet.

3.2.2 NUMBER OF RETWEETS BY DAY OF YEAR:



This graph suggests that there was maximum activity in March, 2021. This may be because of the maximum covid cases and deaths recorded for India (maximum activity by this country on twitter) in this month. Additionally, we have an increasing trend of activity starting in July, 2021 as during this period Covid Vaccinations have taken a stronger hold and have increased in number, becoming a major topic of discussion.

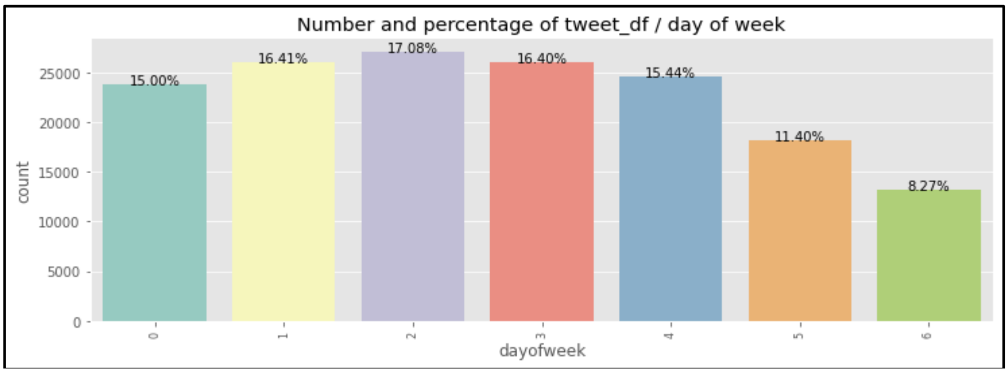
3.2.3 AVERAGE LENGTH OF TWEETS BY HOUR:



The above graph suggests that the average text length was 125/126 words during the time range of 5:00AM to 5:00PM after which it decreases.

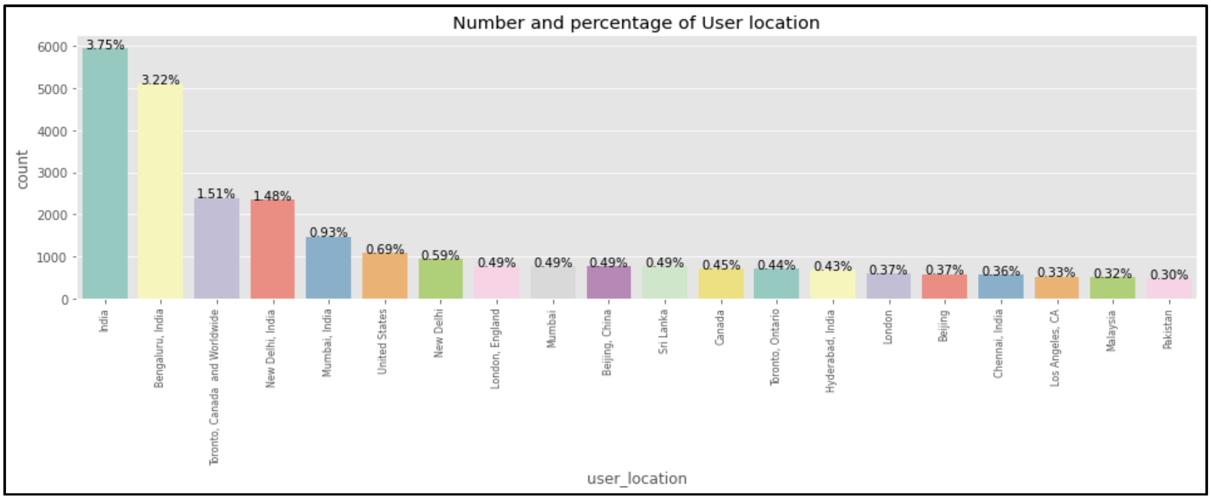
3.3 HISTOGRAMS:

3.3.1 NUMBER AND PERCENTAGE OF TWEETS BY DAY OF WEEK:



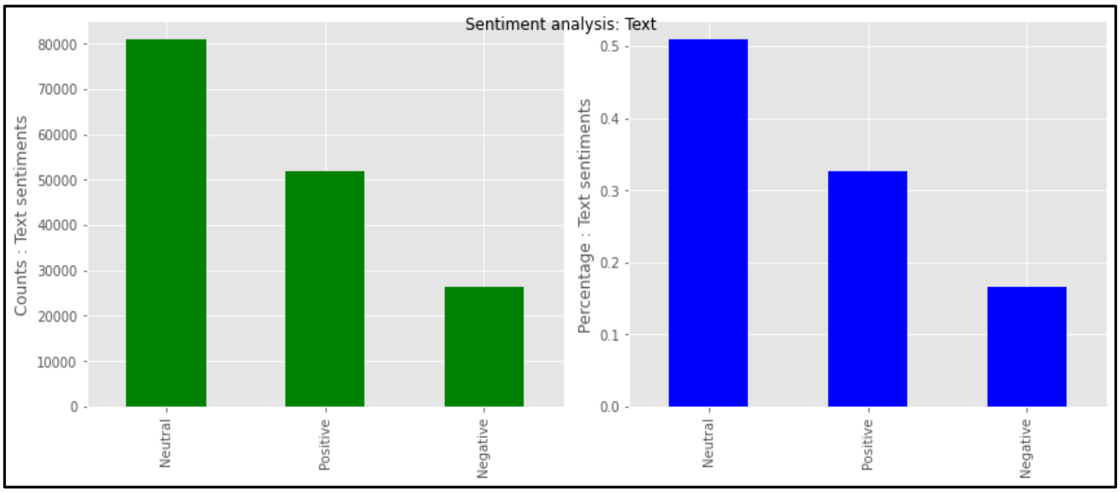
The above figure tells us that tweets are maximum on Wednesday, and comparatively more tweets on weekdays than on weekends. (0 denotes Monday and 6 denotes Sunday)

3.3.2 NUMBER AND PERCENTAGE OF TWEETS BY USER LOCATION:



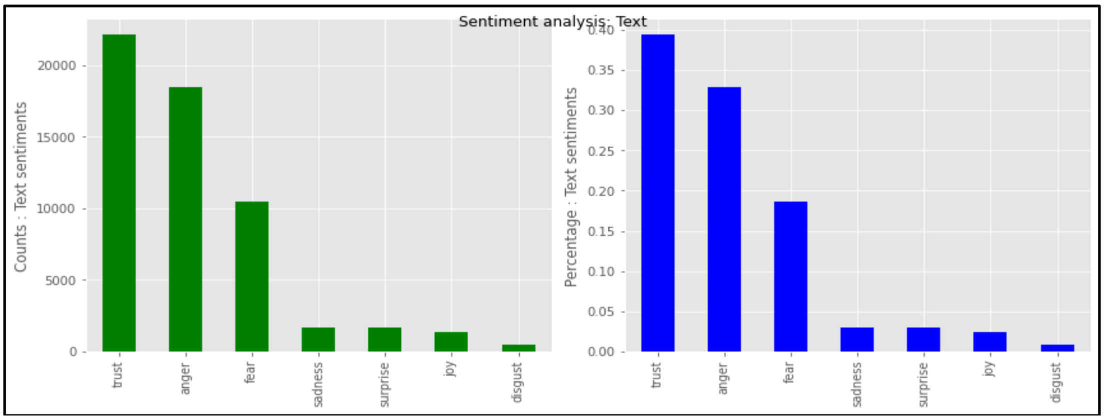
The plot above suggests that most of the tweets were concentrated in India, USA, UK, and China as these were the countries struck the most by Covid and they have a high vaccination progress.

3.4 SENTIMENT ANALYSIS:



From the above histogram, we can infer that most of the tweets in this data-set have a neutral or positive sentiment. It tells us that the public has had a generally moderate or positive feedback to the covid vaccinations.

3.5 EMOTIONAL RECOGNITION:



From the above histogram, we can infer that most of the tweets in this data-set have a trust and anger emotion. It tells us that the public trusts these vaccines but at the same time have feelings of anger and fear regarding the vaccinations and getting slots etc.

IV. CONCLUSION

In conclusion, we can say that we have successfully carried out Exploratory Analysis and baseline level sentiment and emotional analysis on the tweets of the COVID Vaccinations data-set from December, 2020 to 8th August, 2021. Throughout this paper, we have used algorithms of Natural Language Processing, and different types of data visualization techniques to analyze different features of this data-set and infer the feedback the public has for the most prevalent vaccines from the countries most affected by the novel coronavirus.

We have used sentiment and emotional analysis to take this research one step forward and focus specifically towards summarizing the public's opinion about these vaccines used.

To summarize the results, we can say that the vaccines have had a positive toll on the public and most people are either neutral or positive towards the different types of vaccines provided. People highly trust these vaccines, but there are negative emotions of anger and fear still in the community as still many people are affected by the fatal coronavirus and many of these vaccines have had negative side effects. Most of the tweets in this data-set are from India, USA, UK, and China as these were the countries most affected by the virus or most involved in the development of the coronavirus vaccines. We can additionally infer that the vaccines most used in today's time are the following,

- Pfizer/BioNTech
- Sinopharm

- Sinovac
- Moderna
- Oxford/AstraZeneca
- Covaxin
- Sputnik V

The tweets usually peak at the 15th hour (3:00PM) of the day, precisely at 3:30PM each day. The activity was the most in the month of march as this was the month in which coronavirus hit humanity the most and maxed out in cases and deaths. From July onwards, activity on twitter showed an increasing trend as this was the time when vaccine development and production gained a stronger hold.

Therefore, we can say that the public generally has positive feedback regarding the covid vaccinations and we are moving in the right direction when it comes to fighting this disease. It is important to stay safe in these uncertain times, so stay safe and isolated, get vaccinated and keep tweeting!

REFERENCES

[1]. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html

[2]. Python — sentiment analysis using vader. URL https://www.geeksforgeeks. org/python-sentiment-analysis-using-vader/

[3]. Mark M. Bailey. Nrclex. URL https://pypi.org/project/NRCLex/#description

[4]. Gopalkrishna Barkur, Vibha, and Giridhar B. Kamath. Sentiment analysis of nationwide lock down due to covid 19 outbreak: Evidence from india. Asian Journal of Psychiatry, 51: 102089, 2020. ISSN 1876-2018. doi: https://doi.org/10.1016/j.ajp.2020.102089. URL https: //www.sciencedirect.com/science/article/pii/S1876201820302008

[5]. Pawan Bhandarkar. Covid-19 eda: Man vs disease. 2020. URL https://www.kaggle. com/pawanbhandarkar/covid-19-eda-man-vs-disease

[6]. Venkateswarlu Bonta, Nandhini Kumaresh, and N. Janardhan. A comprehensive study on lexicon based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8: 1–6, 03 2019. doi: 10.51983/ajcst-2019.8.S2.2037

[7]. Nalini Chintalapudi, Gopi Battineni, and Francesco Amenta. Sentimental analysis of covid-19 tweets using deep learning models. Infectious Disease Reports, 13(2):329–339, 2021. ISSN 2036-7449. URL https://www.mdpi.com/2036-7449/13/2/32

[8]. Raj Gupta, Ajay Vishwanath, and Yinping Yang. Covid-19 twitter dataset with latent topics, sentiments and emotions attributes. 09 2020

[9]. Vishwanath Ajay Gupta, Raj and Yinping Yang. Covid-19 twitter dataset with latent topics, sentiments and emotions attributes. Inter-university Consortium for Political and Social Research [distributor], 2020. URL https://doi.org/10.3886/E120321V5

[10]. Shravan I.V. Sentiment analysis in python using nltk. OSFY - OpensourceForYou, 12 2016

[11]. https://matplotlib.org/

[12]. https://www.worldometers.info/coronavirus/

[13]. Gabriel Prada. Covid-19 all vaccines tweets. 2020. URL https://www.kaggle.com/ gpreda/all-covid19-vaccines-tweets

[14]. Jeff Raven. Covid-19 vaccine tracker. 2021. URL https://www.raps.org/ news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker

[15]. ankthon. Python — sentiment analysis using vader. URL https://www.geeksforgeeks. org/python-sentiment-analysis-using-vader/

[16]. Mark M. Bailey. Nrclex. URL https://pypi.org/project/NRCLex/#description

[17]. Gopalkrishna Barkur, Vibha, and Giridhar B. Kamath. Sentiment analysis of nationwide lock down due to covid 19 outbreak: Evidence from india. Asian Journal of Psychiatry, 51: 102089, 2020. ISSN 1876-2018. doi: https://doi.org/10.1016/j.ajp.2020.102089. URL https: //www.sciencedirect.com/science/article/pii/S1876201820302008

[18]. PAWAN BHANDARKAR. Covid-19 eda: Man vs disease. 2020. URL https://www.kaggle. com/pawanbhandarkar/covid-19-eda-man-vs-disease

[19]. Venkateswarlu Bonta, Nandhini Kumaresh, and N. Janardhan. A comprehensive study on lexicon based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8: 1–6, 03 2019. doi: 10.51983/ajcst-2019.8.S2.2037

[20]. Nalini Chintalapudi, Gopi Battineni, and Francesco Amenta. Sentimental analysis of covid-19 tweets using deep learning models. Infectious Disease Reports, 13(2):329–339, 2021. ISSN 2036-7449. URL https://www.mdpi.com/2036-7449/13/2/32

[21]. Raj Gupta, Ajay Vishwanath, and Yinping Yang. Covid-19 twitter dataset with latent topics, sentiments and emotions attributes. 09 2020

[22]. Vishwanath Ajay Gupta, Raj and Yinping Yang. Covid-19 twitter dataset with latent topics, sen timents and emotions attributes. Inter-university Consortium for Political and Social Research [distributor], 2020. URL https://doi.org/10.3886/E120321V5

[23]. Shravan I.V. Sentiment analysis in python using nltk. OSFY - OpensourceForYou, 12 2016

[24]. Gabriel Prada. Covid-19 all vaccines tweets. 2020. URL https://www.kaggle.com/ gpreda/all-covid19-vaccines-tweets

[25]. Jeff Raven. Covid-19 vaccine tracker. 2021. URL https://www.raps.org/ news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker