# SENTIMENT ANALYSIS MODEL FOR TWITTER ON COVID-19 VACCINE

Francisca OLADIPO[73]

Prosper AKARAH[74]

Andrew OHIEKU[75]

**Abstract**
Sentiment analysis is a classification technique that specializes in categorizing a body of texts into various emotions. This categorization had proven to be handy in classifying tweets into positive, negative, or neutral emotions. The focus of this paper is to determine the sentiment analysis of Indians and Americans. Using a lexicon-based analytic architecture and a dataset used for this research work was gotten from an online database Kaggle dataset called "All COVID-19 Vaccines Tweets". The dataset contains 125,906 entries with 16 columns with every country in the world from which tweets with location marked India and USA/United States were extracted. The analysis was done in Python Programming Software with the application of a python module TextBlob. The result shows that the Americans have larger positive sentiments over the Indians with 3.26%.

**Keywords:** sentiment analysis, classification, machine learning, twitter, COVID-19 vaccines tweets

**JEL Classification**: Z00

## 1. Introduction

The social media space has evolved into extremely complex structures of information exchange platform and due to the increase in the practices of different social media platforms, there has been an increasing surge of interests in sentiment analysis as a paradigm for the mining and analysis of user opinions and sentiments based on their posts [1].

Coronavirus disease 2019 (COVID-19) is defined as illness that was caused by a novel coronavirus which is now called severe acute respiratory syndrome coronavirus 2 (SARS-COV-2); it was first identified when an outbreak of respiratory illness cases in Eagan city, Huben province in China. It was initially reported to World Health Organization (WHO) on the 30th of December 2019 and on January 30th, 2020 it was declared as a global health

[73] Professor, Federal University Lokoja, Nigeria, francisca.oladipo@fulokoja.edu.ng
[74] Student, Federal University Lokoja, Nigeria, akarahprosper@gmail.com
[75] Teacher, Capital Science Academy, Nigeria, andrewohieku@gmail.com

emergency. Finally on March 11th, 2020, the disease was declared a global pandemic by WHO which lead to lockdown all over the world [2].

Well, currently the phases of lockdown have gradually been overcome all over the world and Nigeria is not excluded and this is because medical researchers all over the world are on deck to find a vaccine to COVID-19, which now result to having different vaccines which are still under clinical trials. At the moment, the ongoing availability of COVID-19 vaccine poses a pressing need for continual monitoring and to do that we must understand the public opinions in other to develop kickoff levels of confidence in vaccines and enable us to identify early warning signals of losses in confidence, which will help us address the doubtful ones and assure trust in immunization, to realize the advantage of the immunization [3].

Traditionally, governments make use of survey processes to understand the public attitude, which is not the best process in this case because it suffers from small samples sizes, cheap questions and very limited space and time is allocated because of human heart wanting to profit greatly from every reach. So, to overcome these limitations I propose that social media (Twitter) data (tweets) can be used to enable real-time analysis of larger public sentiments and attitudes with appropriate spatiotemporal granularity.

## 2. Literature Review

This subsection summarizes and describes several related works at the introductory part of Artificial Intelligence, Machine Learning, Natural Language Processing (NLP), and sentiment analysis. The review of these works is to state qualitative approach to solving the problem, thereby pointing out improvements that can be made on the low side discovered in these works.

### 2.1 Sentiment Analysis in social media (Tweets) Classification

[1] stated that currently, sentiment analysis is a very active research domain in Artificial Intelligence with over 2,200,000 research items available in the Google Scholar search engine with the keyword ‘Sentiment Analysis’; and over 240, 000 research items with a more filtered search using ‘Sentiment Analysis AND Twitter’, also the systematic review and detailed summary is provided in the research which explored 50 research items dealing with sentiment analysis on social media showing a comprehensive review as well as a summary that contains relevant details like the author(s), the dataset used for the study, the settings/methodology, and the key findings that made them proved that sentiment analysis can be very useful in goods/services reviews and terrorism analysis. Also Sop reviewed how twitter data have been mined and analyzed for public health applications, which really showed the importance’s of using tweets for general opinion mining.

While research by [4] on approaches, tools and applications for sentiment analysis provides a classification of approaches with respect to features/techniques, advantages/limitations, and tools; [5] presented a study that investigated subject coverage and sentiment dynamics

on the hot health issue of Ebola from two different media sources: Twitter and news publications. They used vocabulary control on gathered datasets, the n-gram LDA topic modelling technique, entity extraction and entity network, and the notion of topic-based sentiment scores to conduct content and sentiment analysis. They used the Twitter stream API to collect 16,189 news pieces from 1,006 different newspapers and 7,106,297 tweets using the query word "Ebola" or "Ebola virus," then filtered out only those written in English, leaving 14,818 news articles and 4,581,181 tweets. According to the conclusions of this study, Twitter and traditional news channels work independently. Although this study was designed to determine whether differences exist in the content and sentiment of two distinct media outlets through which validated tweets are more personal and untreated, it sheds light on the content of each news medium at a time when news consumption behaviors are undergoing major changes, increasingly relying on audience participation. This study was designed to determine whether differences exist in the content

[6] used two pre-classified datasets of tweets to perform Sentiment Analysis of Tweets Using SVM to dissect the performance of Support Vector Machine (SVM) for sentiment analysis. The first dataset consisted of tweets about self-driving cars, and the second dataset dealt with apple products. The Weka tool was used to compare and analyze performance. The average precision, recall, and F-Measure for the first dataset were 55.8 percent, 59.9 percent, and 57.2 percent, respectively. For the second dataset, the average Precision, Recall, and F-Measure values are 70.2 percent, 71.2 percent, and 69.9%, respectively, illustrating that the SVM's performance is strongly dependent on the input dataset.

## 2.2 Sentiment Analysis on COVID-19 and Vaccine

[7] research on COVID-19 infection which they presented basic knowledge of the COVID-19 characteristics human coronaviruses: their origin, family, transmission, and talked briefly on animal coronavirus. This was clearly stated that as at the time of that research was done in March 2020, was there was no COVID-19 vaccine.

In a paper by [2], the researchers presented research on sentiment analysis of the Nigerian nationwide lockdown due to COVID19 outbreak. In their work, they determine the sentiment analysis of Nigerians within the period of the lockdown exercise using lexicon-based analytic architecture. A total of 22, 249 tweets were extracted from 30th March to 11th May 2020 and obtain a result with 40.7% positive against 20.7% negative polarity, which shows that Nigerian nationwide accepted the lockdown measures in good fate and are positive with the fight against COVID19.

In recent research, [8] used the Nave Bayes sentiment classification algorithm on Twitter data with the keyword 'COVID-19' filtered by the keyword 'vaccine' in Indonesian tweets. The data crawling process is performed manually using the access token received from the Twitter API and the Rapid miner tools to extract the requested information and data result containing over 6000 tweets from January 15th to 22nd 2021. During that time span, the analysis revealed 39 percent positive sentiment, 56 percent negative sentiment, and 1% favorable opinion. Because the public did not believe the vaccination was safe at the time, negative opinions were formed.

## 3. Materials and Method

Sentiment analysis have been continually applicable in various areas or fields like politics, businesses, public actions, and finances to real world problems which have yielded great results from time to time. The proposed model will help solve the stated problems. Dataset was extracted from a social media platform (Twitter) API with respect to some keywords e.g. COVID19VACCINE, COVID-19 Vaccination etc., the dataset is fitted into the model to collect the input in the column titled 'text' in the dataset CVS fil. The texts is pre-processed properly by performing the following processes: tokenization, noun phase extraction, POS tagging, words inflection and lemmatization, N-grams. Finally, the sentiment analysis is performed, and the detail summary of all opinions are analyzed and displayed based on the classification parameters -1, 0 and 1 which represent negative, neutral, and positive text respectively. For every row been analyzed a parameter is generated in place of the text, to represent it for proper displays in form of charts and percentages. The High-Level Model of the system is showed in Figure 1.

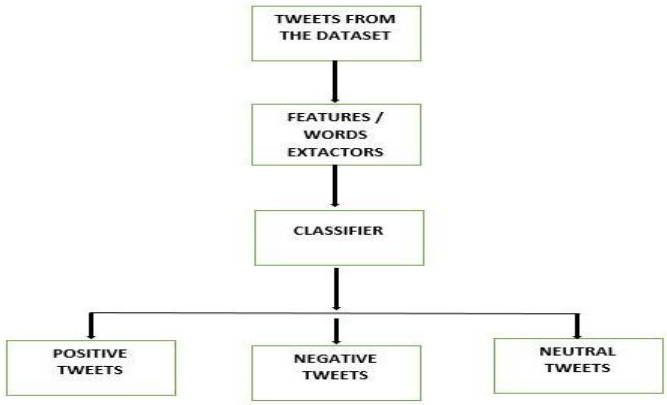


Figure 1. High Level Model of the System

### 3.1 Methodology

This section of the paper describes the stages sequentially on the ways in achieving the stated objectives of the proposed system. The selected methodology is the Cross Industry Standard Process for Data Mining (CRISP-DM). This methodology was chosen due to its sequential and iterative approach to problem solving to applying data science and machine learning algorithms which is relevant to activities carried out in this research. We systematically employ the scientific methods identified with this methodology. This research is developed with Python Programming Language.

### 3.2 Specification and Justification for the Selected Methodology

Below are the implementations of the various steps of CRISP-DM in this research.

i. **Research understanding:** In this phase of this research work we understood the topic sentiment analysis and the various COVID19 vaccines and how various vaccination processes take place.

ii. **Data Understanding**: this phase explains the dataset collection and description. The dataset used for this research work was gotten from an online database Kaggle dataset called "All COVID-19 Vaccines Tweets" (https://www.kaggle.com/datasets/gpreda/all-COVID19-vaccines-tweets). The dataset contains 125906 entries with 16 columns. The dataset features are shown below (Figure 2).

```
In [10]: df.columns

Out[10]: Index(['id', 'user_name', 'user_location', 'user_description', 'user_created',
                'user_followers', 'user_friends', 'user_favourites', 'user_verified',
                'date', 'text', 'hashtags', 'source', 'retweets', 'favorites',
                'is_retweet'],
               dtype='object')
```

Figure 2. Snippet to show the column in the dataset

iii. **Data Preparation**: this phase is where some cleaning techniques are applied, which help us to clean the dataset to fit for the modelling phase. In this paper we dropped some columns and rows to restrict the dataset to 2 countries which are title India and USA because they are the countries with many tweets. Data processing involves case swapping, removal of special character, tokenization, stop words etc., and then the dataset is split into training and testing set (Figure 3).

```
In [21]: # Cleaning Text
         df['clean_tweet'] = df['text'].apply(nfx.remove_hashtags)

In [22]: df[['text','clean_tweet']]

Out[22]:
```

| | text | clean_tweet |
|---|---|---|
| 0 | The agency also released new information for h... | The agency also released new information for h... |
| 1 | #UgurSahin #ozlemtureci the #Muslim Scientists... | the Scientists Husband-Wife are saving t... |
| 2 | Toronto to receive Ontario's 1st doses of Pfiz... | Toronto to receive Ontario's 1st doses of Pfiz... |
| 3 | More approvals to #PfizerBioNTech vaccine: #Si... | More approvals to vaccine: approves @pfize... |
| 4 | #Pfizervaccine is unlikely to be available in ... | is unlikely to be available in India for mas... |
| ... | ... | ... |
| 5347 | #CoronaVirusUpdates\n\n- India reports 42 cas... | \n\n- India reports 42 cases of\n\n- Cases... |
| 5348 | The @WHO said it had uncovered problems at a #... | The @WHO said it had uncovered problems at a ... |
| 5349 | The UN health agency inspected four #SputnikV ... | The UN health agency inspected four manufact... |

Figure 3. Snippet of Data Preparation to remove hashtag within the tweet

**iv.** **Modelling:** In this phase, the model was developed using a lexicon based or rule-based approach using Text Blob library. Text Blob is a python module and provides a simplistic API to use its methods and carry out NLP tasks. Text Blob's goal is to provide a familiar interface for common text processing operations. You can think of Text Blob objects as Python strings that have acquired the ability to perform Natural Language Processing. A nice feature of Text Blob is its resemblance to strings. As such, you can use them in the same way as strings. Few of the simpler tasks have been demonstrated below. The following code demonstrates that Text Blob is identical to a string, and the syntax is merely to illustrate the point (Figure 4).

```
[ ]  #Sentiment Analysis

[ ]  from textblob import TextBlob

[ ]  def get_sentiment(text):
         blob = TextBlob(text)
         sentiment_polarity = blob.sentiment.polarity
         sentiment_subjectivity = blob.sentiment.subjectivity
         if sentiment_polarity > 0:
             sentiment_label = 'Positive'
         elif sentiment_polarity < 0:
             sentiment_label = 'Negative'
         else:
             sentiment_label = 'Neutral'
         result = {'polarity':sentiment_polarity,
                   'subjectivity':sentiment_subjectivity,
                   'sentiment':sentiment_label}
         return result
```

Figure 4. Lexicon-based modeling using TextBlob

**v.** **Evaluation**: This section is concerned with the evaluation of the model in the context of the research objectives using different evaluation metrics. The accuracy of the model will also be considered after which it has been tested, having already trained the model with the training set. This help to visualize the result of the polarity (Figure 5).

```
In [43]: # Plot with seaborn
         sns.countplot(df['sentiment'])

Out[43]: <AxesSubplot:xlabel='sentiment', ylabel='count'>
```

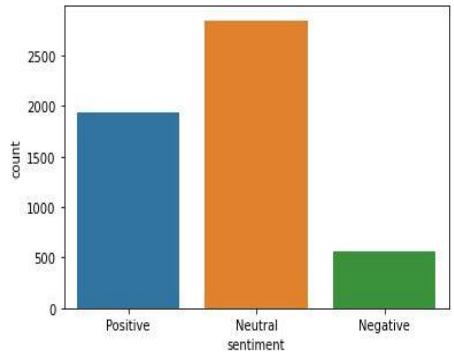


Figure 5. Model Evaluation Visualization

vi. **Deployment:** In this phase, we assess and interpret the mined pattern, rules, and reliability to the objective.

## 4. Implementation

### 4.1. Data Understanding

Our dataset was gotten from an online data science community "Kaggle". Dataset was selected in this case because our research is on text analytics and tweets from Twitter are usually in textual format and provide a randomized and raw form of data in which the tweets are present in the dataframe.

### 4.2. Data Preparation

These steps include removal of stop words, tokenization, normalization, stemming, TF-IDF weighting.

### 4.3 Packages used in the research

These are the libraries used in this research execution (Figure 6).

```
# EDA Pkgs
import pandas as pd

# Data Viz Pkg
import matplotlib.pyplot as plt
import seaborn as sns

# Hide warnings
import warnings
warnings.filterwarnings('ignore')

from textblob import TextBlob

from wordcloud import WordCloud

from collections import Counter
```

Figure 6. Packages and Libraries Visualization

### 4.4 Load Dataset

The instruction below shows the process of accessing the dataset for the research execution.

```
df = pd.read_csv("IndiaDataset.csv", index_col= 0)
```

this loads the data as shown below (Figure 7).

```
# Check Columns
df.columns
```

```
Index(['id', 'user_name', 'user_location', 'user_description', 'user_created', 'user_followers',
'user_friends', 'user_favourites', 'user_verified', 'date', 'text', 'hashtags', 'source', 'retweets',
'favorites', 'is_retweet'], dtype='object')
```

```
df.shape
```

```
(5352, 16)
```

```
df.dtypes
```

```
id                   int64
user_name           object
user_location       object
user_description    object
user_created        object
user_followers       int64
user_friends         int64
user_favourites      int64
user_verified         bool
date                object
text                object
hashtags            object
source              object
retweets             int64
favorites            int64
is_retweet            bool
dtype: object
```

Figure 7. A glimpse of the data

Next, the most useful columns are selected using the instructions below:

```
# Selecting most useful columns
df = df[['date', 'user_location', 'text', 'hashtags', 'source']]

df.head()
```

| | date | user_location | text | hashtags | source |
|---|---|---|---|---|---|
| 12 | 2020-12-12 17:45:00 | India | The agency also released new information for h... | NaN | TweetDeck |
| 75 | 2020-12-14 20:00:51 | India | #UgurSahin #ozlemtureci the #Muslim Scientists... | ['UgurSahin', 'ozlemtureci', 'Muslim', 'Pfizer... | Twitter for Android |
| 94 | 2020-12-14 18:27:23 | India | Toronto to receive Ontario's 1st doses of Pfiz... | ['Ontario'] | Twitter Web App |
| 131 | 2020-12-14 12:48:58 | India | More approvals to #PfizerBioNTech vaccine: #Si... | ['PfizerBioNTech', 'Singapore', 'CovidVaccine'] | Twitter Web App |
| 159 | 2020-12-14 06:57:09 | India | #Pfizervaccine is unlikely to be available in ... | ['Pfizervaccine', 'PfizerBioNTech'] | TweetDeck |

The most useful columns consist of the date which is used in getting the time when the text was updated, the user location which is used to validate the area where the tweet is been uploaded from, text which is the most important which the text is been pre-processed by

the model to get our sentiments. The hashtags are used to show and categorize relevant keywords within the text and finally the source is to show the device used by the user to tweet. All these processes are done to show the authentication of the dataset.

### 4.5 Distribution of the Sources

The distribution of the sources is implemented by the code below and the result is shown in Figure 8.

```
df['source'].unique()
```

```
array(['TweetDeck', 'Twitter for Android', 'Twitter Web App',
       'Twitter for iPhone', 'Echobox', 'Twitter Media Studio - LiveCut',
       'Blog2Social APP', 'Hocalwire Social Share', 'Twitter for iPad',
       'Twittimer', 'Twitter Media Studio', 'SocialPilot.co',
       'Hootsuite Inc.', 'Nonli', 'Twitter for Advertisers', 'dailyindia',
       nan, 'NDTV News Studio', 'Grabyo', 'Buffer', 'Twitter for Mac',
       'Zoho Social', 'AgoraPulse Manager', 'Samrudhi Global',
       'ETRetail.com', 'Periscope', 'omniversee', 'Birdie for Twitter',
       'SEMrush Social Media Tool', 'Cowin Vaccination Availability',
       'Publer.io', 'WordPress.com', 'The Tweeted Times', 'IFTTT',
       'LinkedIn', 'Instagram', 'C19VSNotification', "Sharoon's Bot"],
      dtype=object)
```

```
# Plot the top value_counts
plt.figure(figsize=(20,10))
df['source'].value_counts().nlargest(30).plot(kind='bar')
plt.xticks(rotation=45)
```

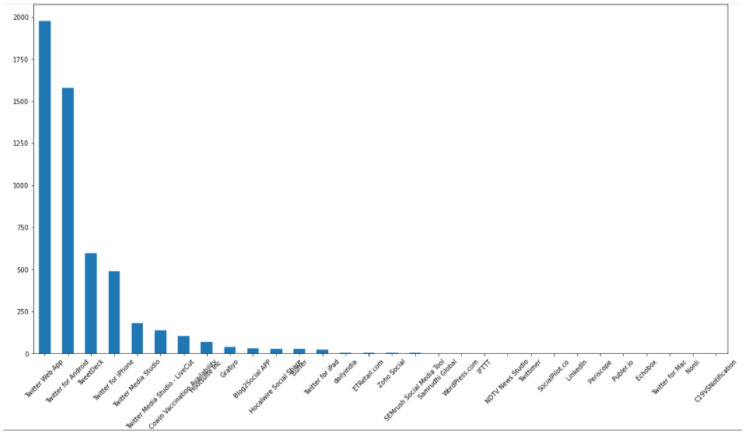


Figure 8. Sources of the tweets by devices

The display above shows that most user tweets source value on the y axis were tweet from the twitter web app with almost 2000 users, while the source of the tweets are labelled on the x axis. This is done to show the users devices for tweeting.

The tweets are then cleaned using the following processes:

```
# Load Text Cleaning Package
import neattext.functions as nfx

df['text'].iloc[2]
```

"Toronto to receive Ontario's 1st doses of Pfizer COVID-19 vaccine today://t.co/Tt7qxCQqDY#Ontario… https://t.co/vacMDknWAV"

```
#Noise remove mentions/userhandles, remove hashtags, urls, emojis, special char

df['text'].apply(nfx.extract_hashtags)
```

```
12                                                      []
75        [#UgurSahin, #ozlemtureci, #Muslim, #PfizerBio...
94                                                [#Ontario…]
131           [#PfizerBioNTech, #Singapore, #CovidVaccine,]
159                     [#Pfizervaccine, #PfizerBioNTech…]
                                ...
125865             [#CoronaVirusUpdates, #DeltaPlusVariant]
125868                               [#SputnikV, #CovidVaccine]
125890                    [#SputnikV, #Covid19vaccine, #WHO]
125891                                                [#SputnikV]
125898                                                [#SputnikV]
Name: text, Length: 5352, dtype: object
```

```
df['extracted_hashtags'] = df['text'].apply(nfx.extract_hashtags)

df[['extracted_hashtags','hashtags']]
```

| | extracted_hashtags | hashtags |
|---|---|---|
| 12 | [] | NaN |
| 75 | [#UgurSahin, #ozlemtureci, #Muslim, #PfizerBio... | ['UgurSahin', 'ozlemtureci', 'Muslim', 'Pfizer... |
| 94 | [#Ontario…] | ['Ontario'] |
| 131 | [#PfizerBioNTech, #Singapore, #CovidVaccine,] | ['PfizerBioNTech', 'Singapore', 'CovidVaccine'] |
| 159 | [#Pfizervaccine, #PfizerBioNTech…] | ['Pfizervaccine', 'PfizerBioNTech'] |
| ... | ... | ... |
| 125865 | [#CoronaVirusUpdates, #DeltaPlusVariant] | ['CoronaVirusUpdates', 'DeltaPlusVariant'] |
| 125868 | [#SputnikV, #CovidVaccine] | ['SputnikV', 'CovidVaccine'] |
| 125890 | [#SputnikV, #Covid19vaccine, #WHO] | ['SputnikV', 'Covid19vaccine', 'WHO'] |
| 125891 | [#SputnikV] | ['SputnikV'] |
| 125898 | [#SputnikV] | ['SputnikV'] |

5352 rows × 2 columns

```
# Cleaning Text
df['clean_tweet'] = df['text'].apply(nfx.remove_hashtags)
```

```
df[['text','clean_tweet']]
```

| | text | clean_tweet |
|---|---|---|
| 12 | The agency also released new information for h... | The agency also released new information for h... |
| 75 | #UgurSahin #ozlemtureci the #Muslim Scientists... | the Scientists Husband-Wife are saving t... |
| 94 | Toronto to receive Ontario's 1st doses of Pfiz... | Toronto to receive Ontario's 1st doses of Pfiz... |
| 131 | More approvals to #PfizerBioNTech vaccine: #Si... | More approvals to vaccine: approves @pfize... |
| 159 | #Pfizervaccine is unlikely to be available in ... | is unlikely to be available in India for mas... |
| ... | ... | ... |
| 125865 | #CoronaVirusUpdates\n\n• India reports 42 cas... | \n\n• India reports 42 cases of \n• Cases... |
| 125868 | The @WHO said it had uncovered problems at a #... | The @WHO said it had uncovered problems at a ... |
| 125890 | The UN health agency inspected four #SputnikV ... | The UN health agency inspected four manufact... |
| 125891 | WHO team raises concerns on #SputnikV filling ... | WHO team raises concerns on filling plant in... |
| 125898 | @1stIndiaNews @RaghusharmaINC @kashiram_journo... | @1stIndiaNews @RaghusharmaINC @kashiram_journo... |

5352 rows × 2 columns

```
df['clean_tweet'].iloc[10]
```

'WHO caution civilian of mutations of COVID19 virus.://t.co/I7Y8Uc0COn'

```
# Cleaning Text: Multiple WhiteSpaces
df['clean_tweet'] = df['clean_tweet'].apply(nfx.remove_multiple_spaces)

df['clean_tweet'].iloc[10]
```

'WHO caution civilian of mutations of COVID19 virus. https://t.co/I7Y8Uc0COn'

```
# Cleaning Text : Remove urls
df['clean_tweet'] = df['clean_tweet'].apply(nfx.remove_urls)

# Cleaning Text: Punctuations
df['clean_tweet'] = df['clean_tweet'].apply(nfx.remove_puncts)

df[['text','clean_tweet']].head()
```

| | text | clean_tweet |
|---|---|---|
| 12 | The agency also released new information for h... | The agency also released new information for h... |
| 75 | #UgurSahin #ozlemtureci the #Muslim Scientists... | the Scientists HusbandWife are saving the wor... |
| 94 | Toronto to receive Ontario's 1st doses of Pfiz... | Toronto to receive Ontarios 1st doses of Pfize... |
| 131 | More approvals to #PfizerBioNTech vaccine: #Si... | More approvals to vaccine: approves @pfizer ex... |
| 159 | #Pfizervaccine is unlikely to be available in ... | is unlikely to be available in India for mass... |

The cleaned texts displayed above are text that when through the cleaning phases to identify and remove error within the dataset to get more accurate result. The processes done were:

*remove mentions/userhandles, remove hashtags, urls, emojis, special characters.*

### 4.6. Sentiment Analysis

Sentiment analysis of the data was carried out as follows:

```
def get_sentiment(text):
  blob = TextBlob(text)
  sentiment_polarity = blob.sentiment.polarity
  sentiment_subjectivity = blob.sentiment.subjectivity
  if sentiment_polarity > 0:
    sentiment_label = 'Positive'
  elif sentiment_polarity < 0:
    sentiment_label = 'Negative'
  else:
    sentiment_label = 'Neutral'
  result = {'polarity':sentiment_polarity,
  'subjectivity':sentiment_subjectivity,
  'sentiment':sentiment_label}
  return result

# Text
df['clean_tweet'].iloc[0]
```

```
'The agency also released new information for health care providers and for patients as the US
shipped millions of d...'
```

```
get_sentiment(df['clean_tweet'].iloc[0])
```

```
{'polarity': 0.13636363636363635,
'subjectivity': 0.45454545454545453,
'sentiment': 'Positive'}
```

```
df['sentiment_results'] = df['clean_tweet'].apply(get_sentiment)
df['sentiment_results']
```

```
0       {'polarity': 0.13636363636363635, 'subjectivit...
1       {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
2       {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
3       {'polarity': 0.375, 'subjectivity': 0.41666666...
4       {'polarity': -0.04999999999999999, 'subjectivi...
                              ...
5347    {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
5348    {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
5349    {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
5350    {'polarity': -0.125, 'subjectivity': 0.375, 's...
5351    {'polarity': 0.0, 'subjectivity': 0.0, 'sentim...
Name: sentiment_results, Length: 5352, dtype: object
```

### 4.7 Evaluation

The accuracy of the model is evaluated in the context of the research objectives using different evaluation metrics. The accuracy of the model was also considered after it was tested, having already trained the model with the training set. This help to visualize the result of the polarity.

## 5. Results and Discussions

Executing the command df = df.join(pd.json_normalize(df['sentiment_results']))

df.head() gives the result of sentiment analysis on the first analysed tweets (Figure 9).The result showed a positive polarity and subjectivity of 0.1363 and 0.4545 respectively. This means that once the value is above 0, then it is a positive sentiment but if below 0 it negative and neutral if result is equal to zero.

```
df['sentiment_results'].iloc[0]
```

```
{'polarity': 0.13636363636363635,
 'subjectivity': 0.45454545454545453,
'sentiment': 'Positive'}
```

```
pd.json_normalize(df['sentiment_results'].iloc[0])
```

| | polarity | subjectivity | sentiment |
|---|---|---|---|
| 0 | 0.136364 | 0.454545 | Positive |

Figure 9. Polarity and subjectivity result

| | date | user_location | text | hashtags | source | extracted_hashtags | clean_tweet | sentiment_results | polarity | subjectivity | sentiment |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 2020-12-12 17:45:00 | India | The agency also released new information for h... | NaN | TweetDeck | [] | The agency also released new information for h... | {'polarity': 0.13636363636363635, 'subjectivit... | 0.0 | 0.0 | Neutral |
| 75 | 2020-12-14 20:00:51 | India | #UgurSahin #ozlemtureci the #Muslim Scientists... | ['UgurSahin', 'ozlemtureci', 'Muslim', 'Pfizer... | Twitter for Android | [#UgurSahin, #ozlemtureci, #Muslim, #PfizerBio... | the Scientists HusbandWife are saving the wor... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... | 0.2 | 0.3 | Positive |
| 94 | 2020-12-14 18:27:23 | India | Toronto to receive Ontario's 1st doses of Pfiz... | ['Ontario'] | Twitter Web App | [#Ontario...] | Toronto to receive Ontarios 1st doses of Pfize... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... | 0.0 | 0.0 | Neutral |
| 131 | 2020-12-14 12:48:58 | India | More approvals to #PfizerBioNTech vaccine: #Si... | ['PfizerBioNTech', 'Singapore', 'CovidVaccine'] | Twitter Web App | [#PfizerBioNTech, #Singapore, #CovidVaccine,] | More approvals to vaccine: approves @pfizer ex... | {'polarity': 0.375, 'subjectivity': 0.41666666... | 0.0 | 0.0 | Neutral |
| 159 | 2020-12-14 06:57:09 | India | #Pfizervaccine is unlikely to be available in ... | ['Pfizervaccine', 'PfizerBioNTech'] | TweetDeck | [#Pfizervaccine, #PfizerBioNTech...] | is unlikely to be available in India for mass... | {'polarity': -0.04999999999999999, 'subjectivi... | 0.5 | 0.9 | Positive |

df['sentiment'].value_counts()

```
Positive     34
Neutral      32
Negative      3
Name: sentiment, dtype: int64
```

```
# Plot with seaborn
sns.countplot(df['sentiment'])
plt.show();
```

Below is the result of the first 69 tweets that was analyzed to keep showing the result in phases which showed that in the first 69 tweets, which shows that 46% of tweets came out positive, 42% came out negative while 2% were neutral. This means that they were more positive tweets in the first 69 tweets and that shows a sign of support and improvement in the vaccination process.

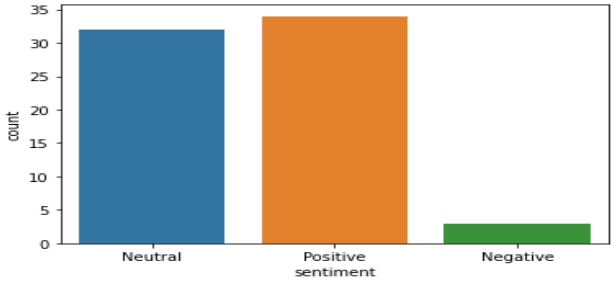


Figure 10. Initial results

The result of the total tweets analyzed for India shows that 53.23% of tweets came out neutral, 36.23% came out positive while 10.54% were negative (Figure 7). This means that they were more neutral tweets, but it still shows that they are more positive tweets over negative tweet, that a sign of support of the vaccine but there is still some level of doubt in the Indians.

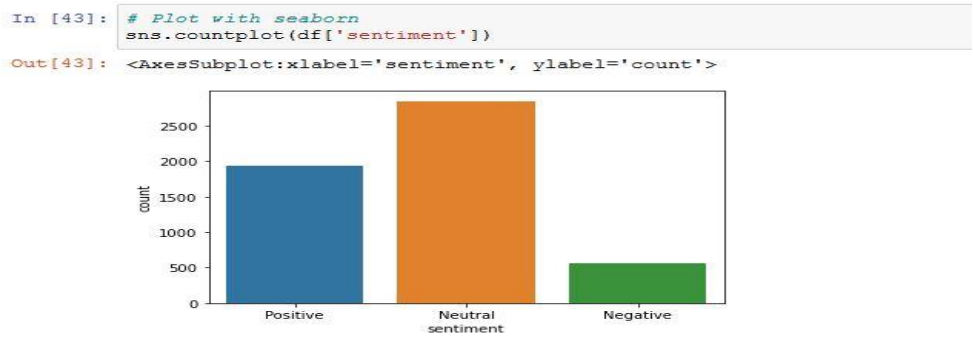


Figure 11. Sentiment analysis done on the complete Indian tweets

For the United States, Figure 8 shows that 47.03% of tweets came out neutral, 39.49% came out positive while 13.48% were negative. This means that there were more neutral tweets but still shows that there are more positive tweets over negative tweet, that as a sign of support of the vaccine but there is yet some level of doubt in the Americans.

```
In [86]: # Plot with seaborn
         sns.countplot(df['sentiment'])
Out[86]: <AxesSubplot:xlabel='sentiment', ylabel='count'>
```

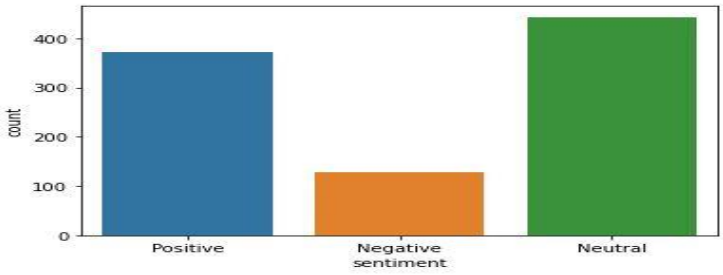


Figure 12. Sentiment analysis on the USA tweets

### 5.1 Keyword Extraction

For Positive and Negative Sentiment

```
positive_tweet = df[df['sentiment'] == 'Positive']['clean_tweet']

neutral_tweet = df[df['sentiment'] == 'Neutral']['clean_tweet']
negative_tweet = df[df['sentiment'] == 'Negative']['clean_tweet']

positive_tweet
```

```
159       is unlikely to be available in India for mass...
577      Canada: Alberta plans to send COVID19 teams to...
611      WHO caution civilian of mutations of covid19 v...
823      I have seen many vaccine videos so far But hav...
914      The video shows dancing to American singer Liz...
957      @NYTHealth So much for equity and equitable di...
977      Sahin the child of a car factory worker was in...
1046     Did US nurse faint after getting PfizerBioNTec...
1054     Im feeling dizzy: US nurse faints after gettin...
1389      begin testing their COVID19 vaccines against ...
1429     WATCH: Mutationbeating possible in six weeks: ...
1485     A nurse practitioner at ChristianaCare hospita...
1644     101yearold woman first in Germany to receive P...
1851     @IsraelinIndia @DrRonMalka @RonyYedidia @Muham...
2029     The @WHO on Thursday granted emergency validat...
2090     WHO approves Pfizer Covid19 vaccine for emerge...
2091      vaccine first to receive emergency validation...
2095     The World Health Organisation ( on Thursday De...
2101     WHOs gift on New Year clears PfizerBioNTech CO...
2113      Grants Emergency Validation For Paving Way Fo...
2168     Ps NopeThey didnt have any side effects after ...
2654     PfizerBioNTech Vaccine Appears Resistant To Ne...
2701     Pfizer's Coronavirus Vaccine Protects Against ...
2712     Pfizer/BioNTech vaccine appears effective agai...
2758              Real Testing of  results arecoming soon
3577       23 people die in after receiving officials D...
3842     Norway probes 23 elderly patients' death after...
4183     Pfizer cancels Covid19 vaccine delivery to Can...
4391      inks deal for delivery of vaccines in poor co...
Name: clean_tweet, dtype: object
```

```
# Remove Stopwords and Convert to Tokens
positive_tweet_list = positive_tweet.apply(nfx.remove_stopwords).tolist()
```

```
negative_tweet_list = negative_tweet.apply(nfx.remove_stopwords).tolist()
neutral_tweet_list = neutral_tweet.apply(nfx.remove_stopwords).tolist()
```

```
positive_tweet_list[1:20]
```

```
['unlikely available India mass distribution Read find',
 'Canada: Alberta plans send COVID19 teams hardhit areas Edmonton Calgary | Indiablooms Portal on…',
 'caution civilian mutations covid19 virus',
 'seen vaccine videos far havent seen aspirating injecting Isnt sm…',
 'video shows dancing American singer Lizzo's song celebrate the…',
 '@NYTHealth equity equitable distribution vaccine world people dev…',
 'Sahin child car factory worker introduced science books Türeci grew watching sur…',
 'nurse faint getting PfizerBioNTech's COVID19 vaccine shot',
 'Im feeling dizzy: nurse faints getting Pfizer COVID vaccine shot ▪A nurse Tennessee hospital faint…',
 'begin testing COVID19 vaccines new',
 'WATCH: Mutationbeating possible weeks: PfizerBioNTech',
 'nurse practitioner ChristianaCare hospital Delaware administered dose vaccine produced by…',
 '101yearold woman Germany receive PfizerBioNTech vaccine coronavirus',
 '@IsraelinIndia @DrRonMalka @RonyYedidia @MuhamedHeib @Orlygoldschmidt @HodayaAvzada @ronenkrausz76 @DanAlluf…',
 '@WHO Thursday granted emergency validation PfizerBioNTech vaccine paving way countries worl…',
 'approves Pfizer Covid19 vaccine emergency use',
 'vaccine receive emergency validation (novel coronavirus) outbreak beg…',
 'World Health Organisation ( Thursday December 31 granted emergency validation',
 'WHOs gift New Year clears PfizerBioNTech COVID vaccine emergency use ▪The Covid19 vacc…']
```

## 5.2 Tokenization

```
pos_tokens = [token for line in positive_tweet_list  for token in line.split()]

neg_tokens = [token for line in negative_tweet_list  for token in line.split()]

neut_tokens = [token for line in neutral_tweet_list  for token in line.split()]
```

## 5.3 Get commonest ceywords

```
def get_tokens(docx,num=30):
    word_tokens = Counter(docx)
    most_common = word_tokens.most_common(num)
    result = dict(most_common)
    return result

get_tokens(pos_tokens)
```

```
{'vaccine': 15, 'PfizerBioNTech': 7, 'emergency': 5, 'nurse': 4, 'Pfizer': 4, 'Study': 4,
'COVID19': 3, 'validation': 3, 'Covid19': 3, 'New': 3, 'Vaccine': 3, 'world': 2, 'India': 2,
'distribution': 2, '|': 2, 'Indiablooms': 2, 'Portal': 2, 'seen': 2, 'people': 2, 'getting': 2, 'shot': 2,
'COVID': 2, 'hospital': 2, 'vaccines': 2, 'new': 2, 'receive': 2, 'coronavirus': 2, 'Thursday': 2,
'granted': 2, 'countries': 2}
```

```
most_common_pos_words = get_tokens(pos_tokens)
most_common_neg_words = get_tokens(neg_tokens)
most_common_neut_words = get_tokens(neut_tokens)

# Plot with seaborn
neg_df = pd.DataFrame(most_common_neg_words.items(),columns=['words','scores'])

plt.figure(figsize=(20,10))

sns.barplot(x='words',y='scores',data=neg_df)

plt.xticks(rotation=45)
```

The seaborn plotted below (Figure 13) shows the 30 most common positive words and the number of occurrences within the dataset been classified as keywords in the dictionary. It shows the relationship between the words (x-axis) against score (y-axis) to show the parameters.

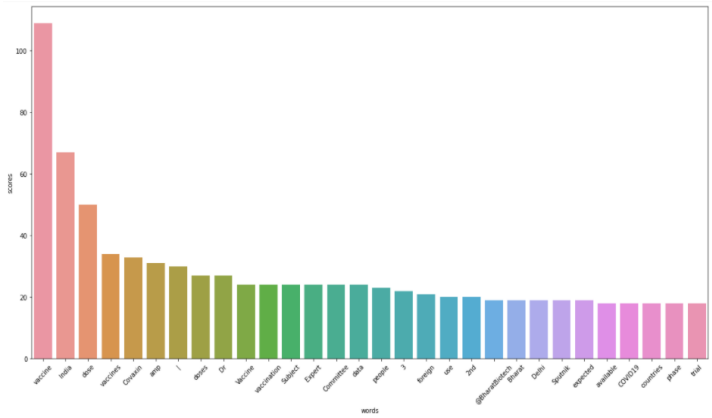


Figure 13. Most common positive words and the number of occurrences

Figure 14 shows the 30 most common negative words and the number of occurrences within the dataset been classified as keywords in the dictionary. It shows the relationship between the words (x-axis) against score (y-axis).

```
# Plot with seaborn
pos_df = pd.DataFrame(most_common_pos_words.items(),columns=['words','scores'])
plt.figure(figsize=(20,10))
sns.barplot(x='words',y='scores',data=pos_df)
plt.xticks(rotation=45)
```

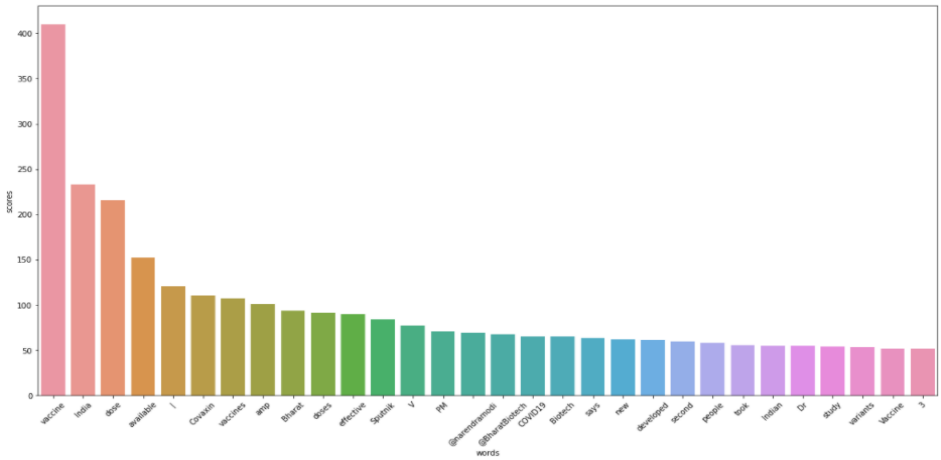


Figure 14. Most common negative words and the number of occurrences

```
# Plot with seaborn
neut_df = pd.DataFrame(most_common_neut_words.items(),columns=['words','scores'])
plt.figure(figsize=(20,10))
sns.barplot(x='words',y='scores',data=neut_df)
plt.xticks(rotation=45)
```

The seaborn plotted below (Figure 15) shows the 30 most common neutral words been classified as keywords in the dictionary.

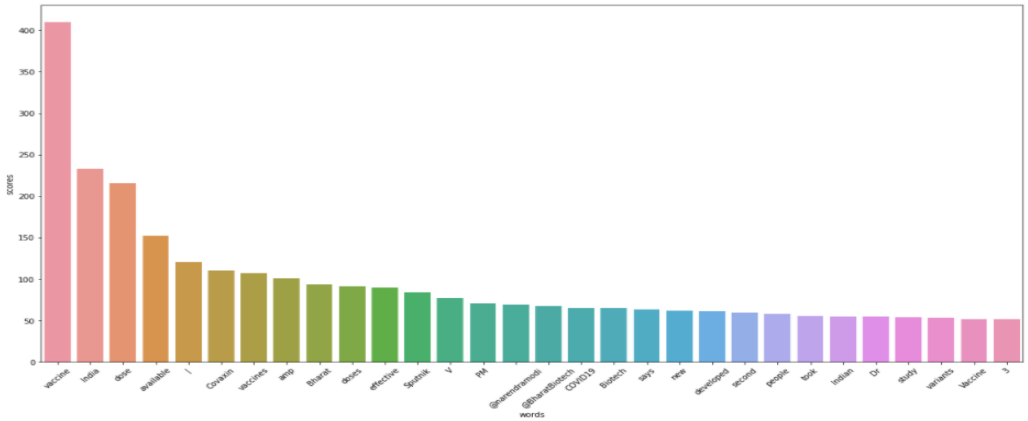


Figure 15. Most common neutral words and the number of occurrences

### 5.4 Word cloud

```
def plot_wordcloud(docx):
  plt.figure(figsize=(20,10))
  mywordcloud = WordCloud().generate(docx)
  plt.imshow(mywordcloud,interpolation='bilinear')
  plt.axis('off')
  plt.show()

pos_docx = ' '.join(pos_tokens)
neg_docx = ' '.join(neg_tokens)
neu_docx = ' '.join(neut_tokens)
```

Figure 16 shows the visualization of most positive words, which were tags as words use quickly to get quick insight of the positive words in the tweets at just a glance using *plot_wordcloud(pos_docx)*


Figure 16. Visualization of the most positive occurring words

Figure 17 show the visualization of most negative words, which were tags as words use quickly to get quick insight of the negative words in the tweets at just a glance after executing *plot_wordcloud(neg_docx)*.


Figure 17. Visualization of the most Negative occurring words

Upon executing *plot_wordcloud(neu_docx*,Figure 18 shows the visualization of most neutral words, which were tags as words use quickly to get quick insight of the neutral words in the tweets at just a glance upon.


Figure 18. Visualization of the most Neutral occurring words

## 6. Conclusions

This research is concerned with the use of Natural Language Processing and lexicon base approach for the extraction of features from social media dataset comprising of users' tweets. The aim is to analyses the tweets of the masses in other to understand how they feel about the COVID-19 vaccination. Other model has been built already, but most of these existing models were built toward analyzing the opinion of the whole world which have the major lapse on the different nature in the human ecosystem which makes so vaccines preferable than others. In this research, we developed a machine learning model to perform sentiment analysis on COVID-19 vaccination. The paper shows the response on the perception of the citizens of both countries in different ways to the ongoing COVID-19 vaccination in the countries of India and the USA. Despite having a lot of neutral tweets in both countries, our overall results shows that both countries citizens expressed positive sentiments about the vaccination. Also, for the result it shows that the Americans have larger positive sentiments over the Indians with 3.26%. This research further gave rise to the emergence of a lexicon base model was developed using Text Blob, which can be used to perform text classifications into polarity or subjectivity.

## References

[1] Oladipo, F., F. B, O., A. E., M., E. E, O., & E., A. (2020). Reviewing Sentiment Analysis at the Shallow End. Transactions on Machine Learning and Artificial Intelligence, 8(4), 47–62. https://doi.org/10.14738/tmlai.84.8274

[2] Ogbuju, E., Francisca, O., Victoria, Y., Rufai. A., Temi, O., Aliyu, A. (2020). Sentiment Analysis of the Nigerian Nationwide Lockdown Due to COVID-19 Outbreak.

[3] Pokharel, B. P. (2020). Twitter Sentiment Analysis During COVID-19 Outbreak in Nepal. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3624719

[4] Alessia, D., Fernando, F., Patrizia, G, Tiziana, G. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. 125 (3).

[5] Kim, E., Jeong, Y., Kim, Y., Kang, K., Song, M. (2016). Topic-based content and sentiment analysis of ebola virus on Twitter and in the news. Journal of Information Science 42 (6) 763-781, 2016

[6] Munir, A., Shabib, A., Iftkhar, A. (2017). Sentiment Analysis of Tweets using SVM. International Journal of Computer Applications 177(5). 975-8887 DOI: 10.5120/ijca2017915758

[7] Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. In Journal of Advanced Research (Vol. 24, pp. 91–98). Elsevier B.V. https://doi.org/10.1016/j.jare.2020.03.005

[8] Pristiyono, Ritonga, M., Ihsan, M. A. al, Anjar, A., & Rambe, F. H. (2021). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. IOP Conference Series: Materials Science and Engineering, 1088(1), 012045. https://doi.org/10.1088/1757-899x/1088/1/012045

[9] Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment Analysis of Tweets using SVM. International Journal of Computer Applications, 177(5), 25–29. https://doi.org/10.5120/ijca2017915758

[10] Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. In Expert Systems with Applications (Vol. 167). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2020.114155

[11] Andrea, A. D., Ferri, F., & Grifoni, P. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. In International Journal of Computer Applications (Vol. 125, Issue 3). http://messenger.yahoo.com/features/emoticons

[12] Chopra, H., Vashishtha, A., Pal, R., Tyagi, A., & Sethi, T. (n.d.). Mining Trends of COVID-19 Vaccine Beliefs on Twitter with Lexical Embeddings.

[13] Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine, 64(1), 17–27. https://doi.org/10.1016/j.artmed.2015.03.006

[14] Dubey, A. D. (n.d.). Public Sentiment Analysis of COVID19 Vaccination Drive in India. https://ssrn.com/abstract=3772401

[15] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89.

[16. Gomaa, W., Haggag, M. H., Badeaa, M. E., & Gomaa, W. H. (2017). Twitter Messages Sentiment Analysis Model based on Deep and Machine Learning Twitter Sentiment Analysis View project Twitter Messages Sentiment Analysis Model based on Deep and Machine Learning. In European Journal of Scientific Research (Vol. 146, Issue 1). http://www.europeanjournalofscientificresearch.com

[17] Goswami, G. (2013). Data mining and data warehousing. New Delhi: S.K. Kataria and Sons.

[18] Guzman, E., & Maalej, W. (2014). How do users like this feature? a fine-grained sentiment analysis of app reviews. 2014 IEEE 22nd International Requirements Engineering Conference (RE), 153-162.

[19] H. Manguri, K., N. Ramadhan, R., & R. Mohammed Amin, P. (2020). Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. Kurdistan Journal of Applied Research, 54–65. https://doi.org/10.24017/COVID.8

[20] Ibrohim, M. O., & Budi, I. (2018). "A dataset and preliminaries study for abusive language detection in indonesian social media. 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI), 135, 222-229.

[21] Kharde, V. A., & Sonawane, S. S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. In International Journal of Computer Applications (Vol. 139, Issue 11). http://ai.stanford.

[22] Kirtipreet, K., & Deepiinderjeet, K. (2015). "A review on automatic text summarization techniques in NLP. International Journal of Computer Sciences and Engineering, 3(7), 62-64.

[23] Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1–15. https://doi.org/10.1080/24751839.2020.1790793

[24] Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). Sentiment analysis on social media. Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, 919–926. https://doi.org/10.1109/ASONAM.2012.164

[25] Pollacci, L., Sîrbu, A., Giannotti, F., Pedreschi, D., Lucchese, C., & Muntean, C. I. (2017). Sentiment spreading: An epidemic model for lexicon-based sentiment analysis on twitter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10640 LNAI, 114–127. https://doi.org/10.1007/978-3-319-70169-1_9

[26] Raghupathi, V., Ren, J., & Raghupathi, W. (2020). Studying public perception about vaccination: A sentiment analysis of tweets. International Journal of Environmental Research and Public Health, 17(10). https://doi.org/10.3390/ijerph17103464

[27] Saif, H., He, Y., & Alani, H. (2012). LNCS 7649 - Semantic Sentiment Analysis of Twitter. www.opencalais.com

[28] Sophie, E., Sierra, E., Isaac, C.,Hai, L., King-Wa, F., Zion, T. (2018). Using Twitter for public Health Surveillance from Monitoring and Prediction to Public Response. 4 (1), 6, 2018

[29] Towers, S., Afzal, S., Bernal, G., Bliss, N., Brown, S., Espinoza, B., Jackson, J., Judson-Garcia, J., Khan, M., Lin, M., Mamada, R., Moreno, V. M., Nazari, F., Okuneye, K., Ross, M. L., Rodriguez, C., Medlock, J., Ebert, D., & Castillo-Chavez, C. (2015). Mass media and the contagion of fear: The case of Ebola in America. PLoS ONE, 10(6). https://doi.org/10.1371/journal.pone.0129179

[30] Tsai, M.-H., & Wang, Y. (n.d.). Analyzing Twitter Data to Evaluate the People's Attitudes to Public Health Policies and Events in the Era of COVID-19.