Department of Electrical and Computer Engineering

North-South University

CSE498R Directed Research

Deep Learning based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter data

Kazi Nabiul Alam **ID # 1711217042**

Md Shakib Khan **ID # 1711661042**

Abdur Rab Dhruba **ID # 1712441042**

Faculty Advisor:

Dr. Mohammad Monirujjaman Khan

Associate Professor

Electrical and Computer Engineering Department

Spring, 2021

DECLARATION

This is to certify that this Project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. Any material reproduced in this project has been properly acknowledged.

<u>Students' name & Signature</u>

1. **Kazi Nabiul Alam** ……………………………………………………
2. **Md Shakib Khan** ……………………………………………………
3. **Abdur Rab Dhruba** ……………………………………………………

APPROVAL

The Directed research project entitled "**Deep Learning based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter data**" by **Kazi Nabiul Alam (ID#1711217042), Shakib Khan (ID #1711661042)** and **Abdur Rab Dhruba (ID#1712441042),** is approved in partial fulfillment of the requirement of the Degree of Bachelor of Science in Computer Science and Engineering in July and has been accepted as satisfactory.

Approve by
Supervisor's Signature
Dr Mohammad Monirujjaman Khan **Associate Professor** Department of Electrical and Computer Engineering North South University Dhaka, Bangladesh.
Department Chair's Signature **………………………………………………………...**
Dr. Rezaul Bari **Associate Professor** Department of Electrical and Computer Engineering North South University Dhaka, Bangladesh.

Acknowledgment

First of all, we wish to express our gratitude to the Almighty for giving us the strength to perform our responsibilities and complete the report.

The project is very helpful to bridge the gap between theoretical knowledge and real life experience as part of the Bachelor of Science (BSc) program. This report has been designed to have practical experience through theoretical understanding.

We also acknowledge our profound sense of gratitude to all the teachers who have been instrumental for providing us the technical knowledge and moral support to complete the project with full understanding.

It is imperative to show our appreciation for our honorable faculty member Dr Mohammad Monirujjaman Khan for his undivided attention and help to achieve this milestone. Also, our gratitude is divine to the North South University, ECE department for providing us a course such as CSE 498R in which we could really work on this project and materialize it the way we have dreamt of.

We thank our friends and family for their moral support to carve out this project and always offer their support.

ABSTRACT

Currently or even today, the ultimate threat to mankind is the COVID-19 pandemic. It has already taken millions of lives all over the world and hasn't stopped yet. This pandemic is a matter of intense apprehension and dreadfulness that leads us to severe anxiety, phobia, and complicated feelings or emotions. Even after ‘Vaccination’ against Coronavirus had been initiated, people’s feelings became more diverse and complex, and we tried to understand and unravel their sentiments in our research. Social media is currently the best way to express feelings and emotions, and with the help of social media precisely ‘Twitter’, we can have a better idea of what's trending and what’s going on in peoples’ minds. In this research, we used a dataset of tweets, where the tweets were collected from users from all across the world, using Twitter API from December’21 to April’21, and containing tweets of most common vaccines available recently. Using a NLP tool named ‘VADER’, we analyzed the sentiments of people regarding Vaccines of all sorts. By initializing the sentiment polarities into 3 groups (positive, negative, and neutral), we tried to visualize the scenario and our findings came out 37.23% positive, 18.75% negative, and 44.02% neutral responses. Besides, we showed the timeline of these sentiments who fluctuated over time between the mentioned timeline above. Then, we evaluated the performance of the predicting model with LSTM and Bi-LSTM, where for LSTM we found the accuracy of 88.81% and found an accuracy of 89.27% for Bi-LSTM. Our belief is that this study will help us to comprehend the public sentiment related to the COVID-19 vaccines and may have an impact on our goal of exterminating Coronavirus from our beautiful world.

Table of Content

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

COVID-19 outbreak brought significant attention to the healthcare sector in recent times and it changed the entire concept of safety in every aspect of our lives. Social distancing is an effective method to reduce spreading Coronavirus, besides wearing masks, washing hands repeatedly, staying concerned about intimacy, such safety measures are very important these days. But these can only reduce spreading Coronavirus but can't eradicate it completely, and here Vaccination came into broad light, the only parameter that can fight most effectively against SARS-CoV-2 and probably to kill it. Rigorous tests have been conducted with the first mRNA vaccines to be introduced; more than 40,000 people have participated in a Pfizer vaccine trial and 30,000 in the Moderna vaccine trial. The average efficacy rate of those who received vaccines was around 94 percent for the prevention of COVID-19 disease. There were no deaths from the trials, and in both trials, the participants received vaccines. Another viral vector vaccine i.e. Johnson & Johnson, which proved to be able to fight against the COVID-19 virus to stimulate the immune response of the recipient, early findings show a rate of effective action of >85% without serious adverse effects [1].

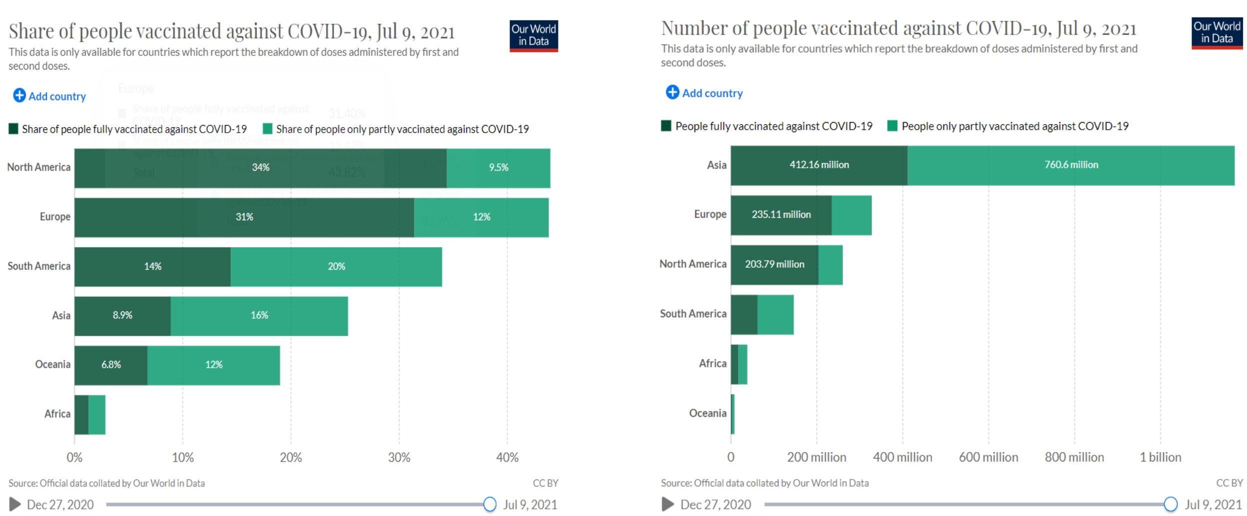


Figure 1: Vaccination Progress in percentage and numbers around the world [2]

Vaccination procedures are going in full swing all across the world, in Fig.1. There might be some reasons and conflicts between region to region in case of urgency and economic barriers which will be explained later in our paper, but we tried to show the actual data of the numbers and figures without biasness, whether the people are vaccinated or not. From this figure above, we can observe that a large amount of the population in different continents is not vaccinated yet.

1.2 Problem Statement

Production of doses of these vaccines is a major concern but we found out the lack of willingness and interest is also an alarming factor and it is a matter of great worry to the health scientists to find out the reasons. We figured out people have mixed feelings in this whole vaccination process from the very beginning, even such conflicts or questions we heard from our family members too. In different studies, researchers tried to understand the reason for such hesitancy. Some issues are discussed in scientific articles recently like what are the reasons that people are thinking more than twice to be vaccinated. Some reasons are: vaccines were invented so fast that there might not have been enough research on it, may cause cancer [3] or infertility, concerns about people getting 2nd dose or not, allergic reactions [4], blood clotting, legitimacy of the production industry, political belief, religious issues, social media, and online trends [5], some conspiracy theories [6], and so on [7], [8].

1.3 Existing Works

In [9], Snscrape was used in the period from 7 January 2020 to 3 January 2021 to collect the historical tweets for vaccination against COVID-19. In all, 4,552,652 Twitter posts have been pulled. These tweets were produced by 1,566,590 users, with 1,012,419 hashtags and 2,258,307 terms of reference. They used Valence Aware Dictionary and sEntiment Reasoner (VADER), a Python Lexicon and a rules-based sentiment analysis tool, which was developed to assess social media feelings on the basis of individual words and phrases, for assigning an amplitude of 'positive,' 'negative' or 'neutral' to each tweet. After extraction, they identified tweet vaccines and opinions and correlated their growth by period, geographic location, emerging topics, key phrases and postal engagement rate and reports. The prevalence of positive and negative feelings was slightly different with positive being the dominant direction and gaining large responses. In another research [10] we found where, throughout the pandemic, tweets from the citizens of the United Kingdom and the United States were collected through Twitter API and experiments conducted to answer three key vaccine questions: Positive, Negative and Neutral. Researchers performed relative sentiment analysis by VADER to get a dominant feeling of the citizens and introduced a modified approach, which can count the influence of the individual. In this way, they were able to take the sentimental analysis a step further and explain some of the changes in the data. The three leading companies are identified: Pfizer, AstraZeneca and Johnson and Johnson, who are involved in research on vaccines and researchers [11] extract their Instagram posts from the start of vaccination and who receive data from users using their own hashtags. The company qualitative variations on manuscripts and visual characteristics, i.e. images categorization by transfer learning, are initially presented in this research. The 'Instaloader' was used to extract the images and the image is classified with VGG-16, Inception V3 and ResNet50. By designing and conducting a controlled experiment, it confirms the accuracy ranking of the results algorithms and identifies the two best-performing algorithms. Finally, the analysis of polarity of users' posts, using a Convolutional Neural Network (CNN), clearly shows a neutral to negative feeling with highly divisive user posts. The aim of this study [12] is to perform a feeling analysis on the Twitter platform of both types of vaccines: Sinovac and Pfizer in Indonesia. Data was crawled and processed between October and November 2020 to understand the emotion. There were two types of datasets: Sinovac and Pfizer. In three classes, both data sets were marked manually: positive, negative, and neutral. With regard to the model performance evaluation, After labeling and preparing data with 'Twitter Crawling', validating with 10-fold cross-validation, they performed Support Vector Machine (SVM), Naive Bayes and Random Forest to evaluate the performance and finally to get results with the proper labeling prediction. The authors of this study [13] collected information on the Philippines' feelings about efforts by the Philippine government using the Twitter Web. In order for the government to analyze its responses, Natural Language Processing (NLP) techniques were applied to understand the overall sentiment. The feelings were trained to categorize English and Philippine language tweets as positive, negative and neutral polemics via the data science tool named 'RapidMiner' using the Naive Bayes model to classify them accurately. Another research [14] we found relevant, performed on Australian people sentiment collected from Twitter. This analysis aimed at extracting important issues and sentiments on Twitter Topics relates to COVID-19 vaccination by using machine-learning methods. They focus in particular on three factors: COVID-19 and its vaccination attitudes, the advocacy of COVID-19 infection control measures and COVID-19 control misconceptions and complaints. Between January and October of 2020, they collected 31,100 English tweets from Twitter users in Australia containing COVID-19-related keywords. In particular, tweets were analyzed by illustrating high-frequency textual data clouds and the interplay of word tokens. In order to identify the commonly mentioned subjects in a large tweet sample, they have created a Latent Dirichlet Allocation (LDA) model. Sentiment analysis was also performed to gain an idea of the overall feelings and emotions in Australia related to COVID-19.

1.4 Project Goals

The above research indicates some high satisfactory outcomes regarding COVID-19 vaccine reactions and their evaluation. Some of the research was related to sentiment related to pandemic tensions and some were related to vaccination issues. But most of the research we found was prioritized region-wise or area-wise, and mostly they did with only Sentiment Analysis tools. Besides, many focused on particular countries and were specific to the vaccines of particular vaccine producing companies. In our research, we tried to analyze the data of all the vaccines available such as Pfizer/BioNTech, Moderna, Oxford/AstraZeneca, Covaxin, Sputnik V, Sinopharm, and Sinovac. We also focused on the timeline to such tweets for understanding sentiments, which we found novel and very important findings because sentiment is very much related and changeable with the flow of time, we also showed with text inputs that our system can detect the sentiment of a sentence properly or not. Our main goal of this research is to get a clear idea about the emotions and thoughts of the mass people regarding the vaccination process of COVID-19 we showed in later parts of this paper, that help the Health Researchers and Rule makers to take proper initiative to make these vaccines more unsuspicious and credulous, to keep the people safe and aware.

In the following sections, we showed the technical components we used and their outcomes with analytical tools. Section. 2 carries the methods and technical terminologies we used to analyze the sentiments. In section 3, we showed proper visualization tools and a description of how these sentiments worked and what outcomes were found. Later we discussed what we achieved from our research and how it can be impactful and beneficial to mankind and how we can further improve this to bring betterment to the world and that’s how we concluded in Section. 4.

CHAPTER 2

Methodology

2 Methods and Materials

2.1 Full outline of the proposed system

In this research, we collected a dataset from Kaggle [15] and our dataset contains different types of tweets related to the COVID-19 vaccine. Our dataset shape is (125906, 16). After loading the dataset we checked the unique values, Null values, and distribution of our dataset. Then we visualize data using user location, source, and user name. Next, for preprocessing, we used detokenize to break the sentences into words and label them. Next, we calculate and add a sentiment column that contains (Positive, Negative, and Neutral). Then we used LSTM and Bidirectional LSTM to evaluate the performance to find out the accuracy of our model. Then we tested our model and found they can detect which sentence is positive, negative, neutral accurately or not and we visualized them all with proper figures, graphs and diagrams.

According to our proposed methodologies mentioned above, we designed the full detailed outline of the system demonstrating in Fig. 2.

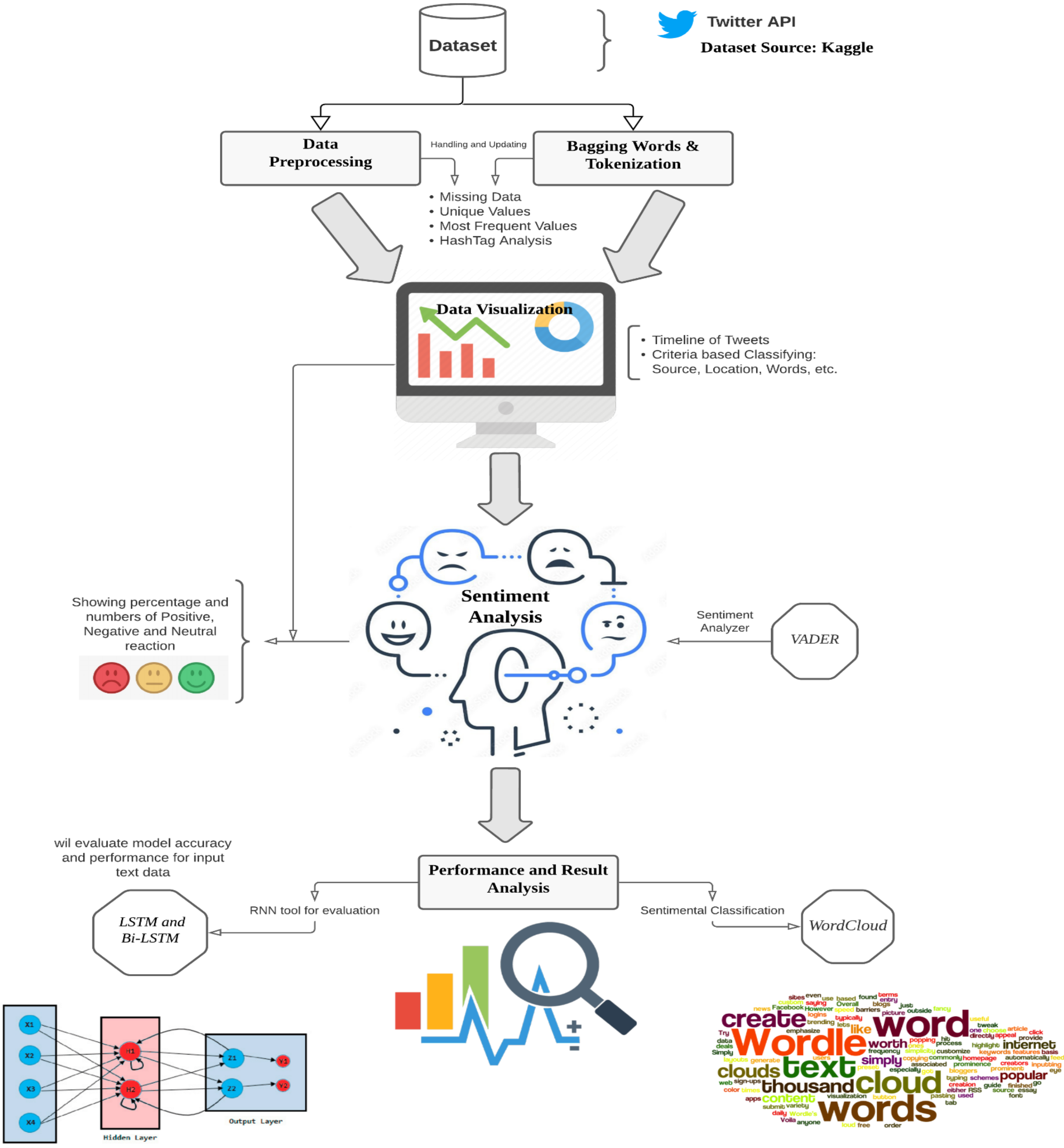


Figure 2: Outline of Sentiment Analysis procedures

2.2 Dataset

There are various datasets available for COVID-19 reactions after it broke out all over the globe, but there are limited numbers of datasets we found for 'Vaccination Reactions' related to COVID-19. From them, we chose a dataset named ' All COVID-19 Vaccines Tweets' from Kaggle [16] where we found data of almost all the renowned vaccines such as Pfizer/BioNTech, Oxford/AstraZeneca, Moderna, Covaxin, Sputnik V, Sinopharm, Sinovac, that are available and accepted to be implemented. Here, the dataset shape is (125906 by 16) that includes username, date, location, number of friends, retweets, hashtags, sources, etc. Our motivation behind choosing this dataset is we found data of all the vaccines available currently and it helps us to generate precise and clear knowledge about vaccination reactions.

2.3 Data Preprocessing, Handling and Tokenization

We collected the dataset from Kaggle, but the CSV file didn't have the process data. So we need to process it first before applying any algorithm. So, first, we dropped some unnecessary columns, then removed all the URLs, emails from the tweets. Then we took out all the new line characters, alimenting all the double, single quotes, and deleted all the punctuation signs. For this type of processing, we first needed to tokenize all the tweets, then apply all the methods of removing those texts. After doing that we detokenize and convert them as a NumPy array.

2.4 Sentiment Analyzer tool (VADER)

VADER (Valence Aware Dictionary for sEntiment Reasonable), proposed by C.J. Hutto [vADER] in 2014, is a lexicon-based sentiment analyzer that is a pre-trained model that uses rules-based values tailored to the perceptions of social media expressions and which works well on texts from other fields. It analyzes a message text and appraises the intensity of this emotion like: positive, negative, and neutral. It is polarity and reflectance Intelligent Dictionary available in the NLTK package. In particular, it has impeccable performance in the area of social media text. VADER can perform a sentimental analysis of assorted lexical characteristics based on its comprehensive rules, shown in Fig. 3. VADER could therefore tackle the problems of indirect use of language groups and inherent action of sentiment.

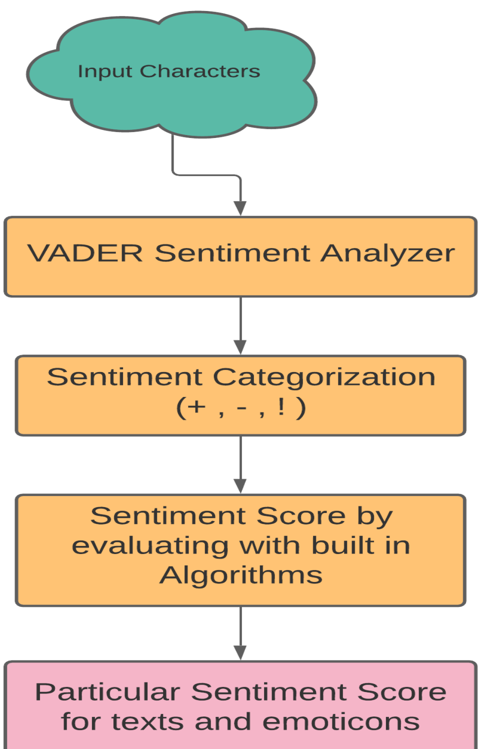


Figure 3: Diagram of VADER

As the valence values for each word in the lexicon, VADER provides a percentage for text ratios that crumble into a positive, negative or neutral category and sums up a probability value of 1. The compound score for sentiment analysis is the most frequently used measure where a float value in the interval [-1,+1] is a compound score, whose index is determined by adding in the lexicon the value values of each word, adapted according to rules and then standardized to its range. No training data is required. It can fully comprehend the vibe of a text which contains emotions, slang words, conjunctions, keywords, punctuation marks, etc. from different particular domains.

2.5 Data Visualization tools

Understanding patterns and correlations between numbers is at the heart of data visualization. Knowing structures, trends, and correlations in groupings of numbers is more important than understanding individual numbers. It may involve detection, measurement, and comparison from the user's perspective and is strengthened by interactive techniques and information from many perspectives and through multiple methodologies. In this work, we do various types of analysis to visually see how our data correlates with another. We used different kinds of diagrams like Bar plots, Line graphs, WordCloud to understand patterns between our datasets. To do visualization, we used lots of prebuilt libraries which are available in python. Tools like Matplotlib, Seaborn, and WordCloud help us to visualize what is hidden in the data.

2.6 Performance Evaluation

In our research, we evaluate our model with a Recurrent Neural Network (RNN) based network called Long Short-Term Memory (LSTM). Here, basic LSTM and Bidirectional LSTM (Bi-LSTM) are being implemented in our evaluation procedure. LSTM is a method of an adept

RNN that addresses RNNs with additional cells, inputs, and outputs. The problems related to vanishing gradients are intuitively rectified by additional additives and gate activations are forgotten, which enable the gradients to pass through the network architecture without eroding so rapidly. A Bidirectional LSTM or Bi-LSTM is a sequence processing model made up of two LSTMs. One takes the input forward and the other backward. Bi-LSTMs efficiently improve the volume of network information in order to improve computational accuracy.

2.6.1 LSTM

An LSTM's control flow is similar to that of a recurrent neural network (RNN) showed in Fig. 4. As it moves forward, it processes data and passes it forward. The LSTM's cells have a variety of actions. These processes are used by the LSTM to recall or forget information. The four gates that make up the LSTM architecture are the input gate, forget gate, control gate, and output gate.

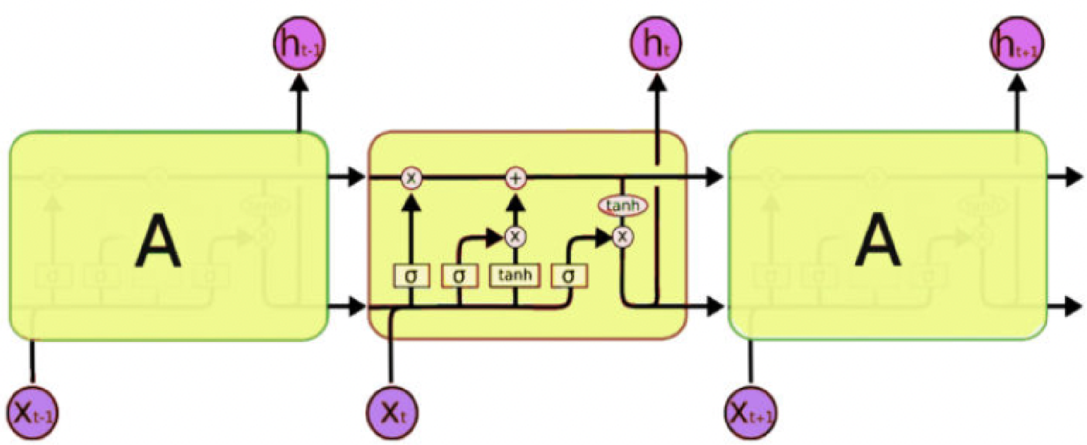


Figure 4: Architecture of Long Short-Term Memory (LSTM) [17]

A series of equations describes the gates of the LSTM [17]. Before attempting to describe the equation, it is necessary to first comprehend some of the variables used in these calculations. The sigmoid activation function is used, as well as the W is weight matrices. The previous LSTM block's output is represented by h_{t-1}, and the preference for the corresponding gates is represented by b_i. Finally, x_t is the existing timestamp's input. Now, The Input gate i_t described as in equation (1)

$$i_t = \sigma\,(W_i * [h_{t-1}, x_t] + b_i) \tag{1}$$

The data that can be given to the cell is chosen using this equation. The forget gate f_t decides which data from the previous memory's input side should be ignored using the equation. (2)

$$f_t = \sigma\,(W_i * [h_{t-1}, xt] + b_i) \tag{2}$$

The following formula (3), where tanh is used to normalize the values into the range -1 to 1, and C is the candidate for cell state at the timestamp, controls the updating of the cell (t).

$$\underline{C} = tanh\,(W_c * [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t * C_{t-1} + i_t * \underline{C} \tag{3}$$

The output layer (o_t) upgrades both the hidden layer h_{t-1} as well as the output layer according to the formula (4)

$$o_t = \sigma (W_o * [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * tanh (C_t) \tag{4}$$

Here is our proposed LSTM configuration in Fig. 5.

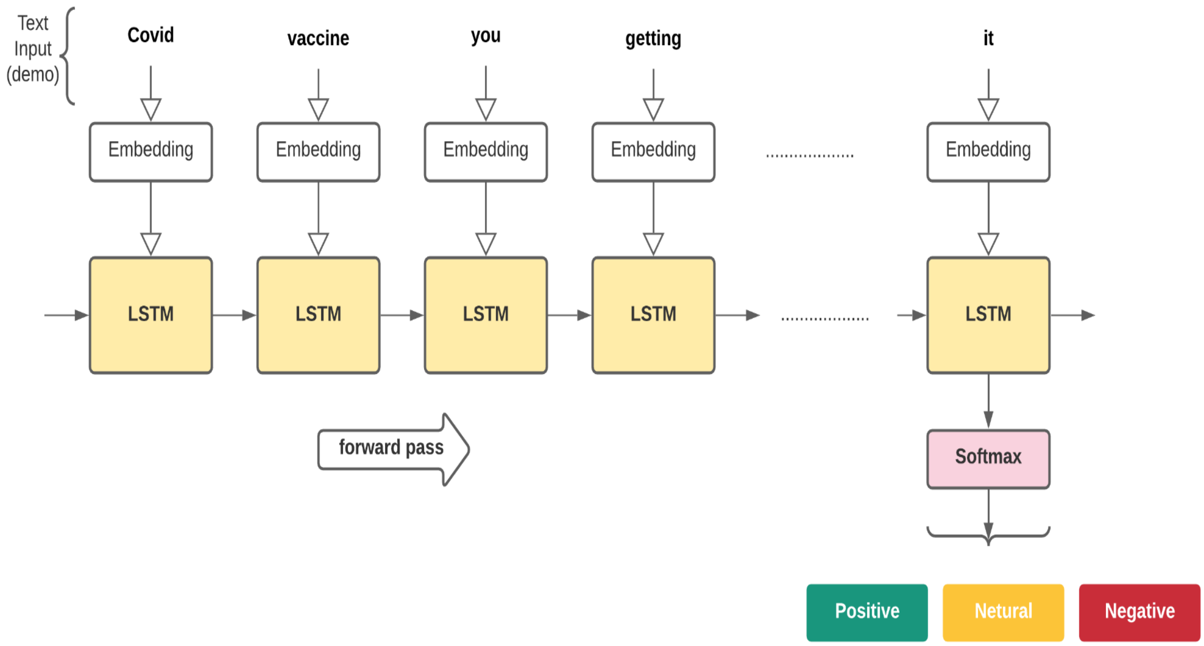


Figure 5: LSTM Architecture for proposed tasks

2.6.2 Bi-LSTM

The idea of Bi-LSTMs comes from bidirectional RNN [18], which analyzes sequence input in both forward and backward directions with two different hidden layers, and is the inspiration for BDLSTMs. The two hidden layers are connected by BDLSTMs to the same output layer. In various disciplines, such as phoneme classification [19] and speech recognition [20], bidirectional networks have been shown to outperform unidirectional networks. The structure of an unfolded BDLSTM layer, which contains a forward LSTM layer and a backward LSTM layer, is described in this section and depicted in Fig. 6.

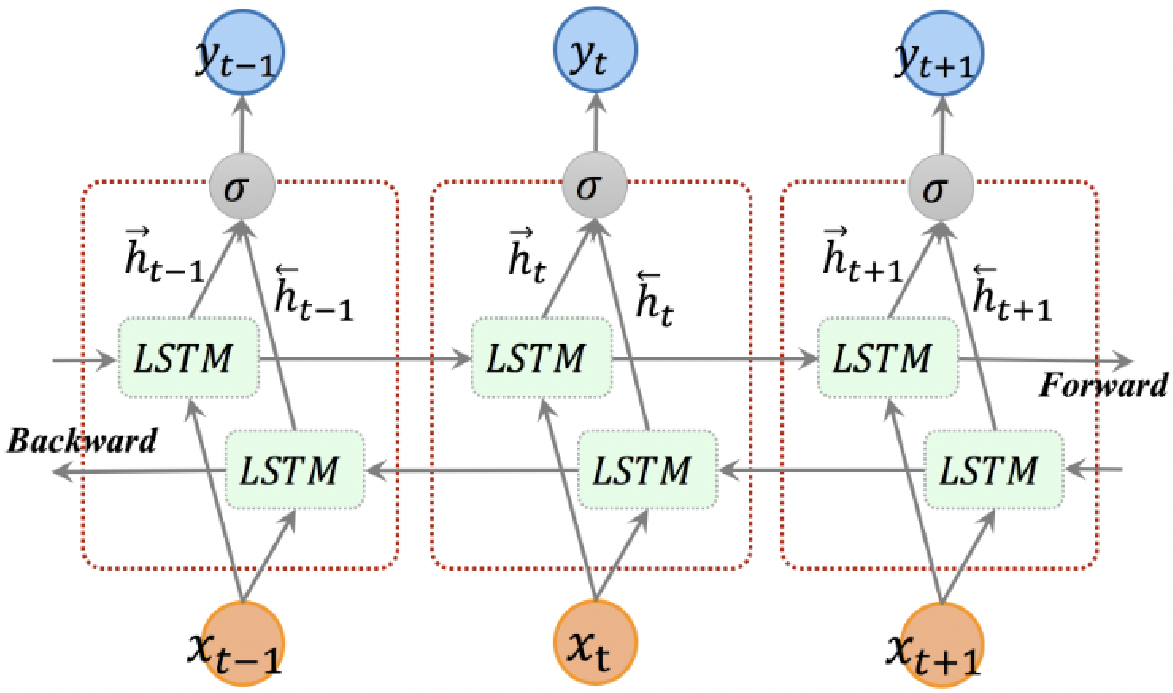


Figure 6: Bidirectional LSTM with three consecutive steps [17]

The forward layer output sequence $h^{\rightarrow}$ is created repeatedly using positive sequence inputs from time $T-n$ to $T-1$, whereas the backward layer output sequence $h^{\leftarrow}$ is calculated using reverse inputs from time $T-n$ to $T-1$. The basic LSTM equations (1) - (4) are used to calculate both the forward and backward layer outputs. The BDLSTM layer generates an output vector Y_T, in which each element is calculated by using the following equation (5):

$$y_t = \sigma\,(\,h^{\rightarrow}\,, h^{\leftarrow}\,) \tag{5}$$

The two output sequences are combined using the σ function. It could be a concatenating, summing, averaging, or multiplying function. A BDLSTM layer's final output can be represented as a vector in the same way that an LSTM layer's final output can which is $[Y_T = Y_{T-n,\,......\,,} Y_{T-1}]$, in the last element, Y_{T-1}, is predicted for the next iteration.

Our Proposed Bi-LSTM architecture is designed below in Fig. 7: for proposed tasks

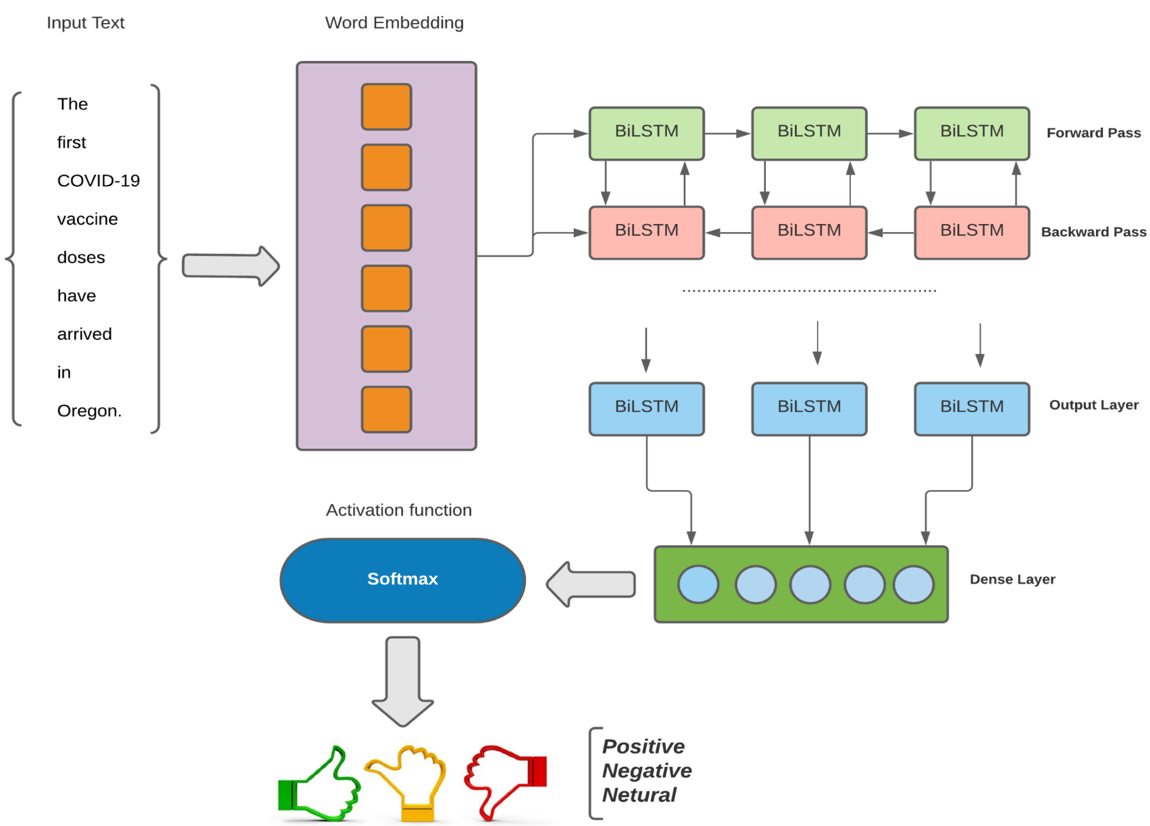


Figure 7: Bi-LSTM architecture

In the following section, we visualized and discussed all the outcomes and their analytical results.

CHAPTER 3

Results and Analysis

3 Result and Analysis

Several outcomes we found and tried to dissect in our research. A brief discussion of such findings we are presenting in the following sections.

3.1 Overview of outcomes

Vaccines available today didn't come with ease. Researchers and Scientists tried hard from the very earliest time of detecting Novel Coronavirus to find out its killer antibody. Pfizer/BioNTech, Oxford/AstraZeneca, Moderna, Covaxin, Sputnik V, Sinopharm, Sinovac such vaccines are the most common and acceptable ones nowadays. And after its production and so many trials on different occasions, there was a timeframe to become available for the mass. So here timeline is important because sentiment and thought change with time, especially in this pandemic situation. Another key thing to keep in mind that, how efficiently we can analyze the fact that people are still or ever interested to be vaccinated, how their psychology works, how they interpret their feelings, and their defense regarding it. People's reactions vary from country to country that we found in this research. The reason behind is the stability of the government, faith in vaccine producing companies, nationalism, and other factors. We tried to cover some major countries where people are active in sharing thoughts on Twitter and they have proper availability of vaccines. Hashtag analysis is another important feature of Sentiment Analysis that we did in our experiment. The most important part was to classify the tweets in the form of ratio analysis to show the numbers and percentage of the reaction: Positive, Negative, and Neutral that we showed in our experiment. Then we validate our model using special sort of RNNs such: LSTM and Bi-LSTM. Both architectures are pretty accurate invalidating and give better accuracy in our model. We analyzed the relative metrics and accuracy-loss and other equivalent indicators of performance evaluation methods and showed them graphically. Using the LSTM and Bi-LSTM architecture, we also showed how well our model can predict particular characters' inputs (tweets).

Wordcloud is a frequently used visualization tool in Sentiment Analysis tasks that we used to show words are categorized as 3 different types of sentiments, to understand the psychological analysis from these tweets. In later parts, we will see all the outcomes sequentially.

3.2 Tweets according to user locations and sources

Tweets are being collected from different sources from different parts of the globe, we analyzed them with facts and figures below.

In Fig. 8, we showed the sources of such vaccine reactions and we found out Twitter was the most used platform for delivering these thoughts in the COVID-19 situation. And as our data contains tweets from all around the world, we tried to show in Fig. 9 some major regions where tweets came from.

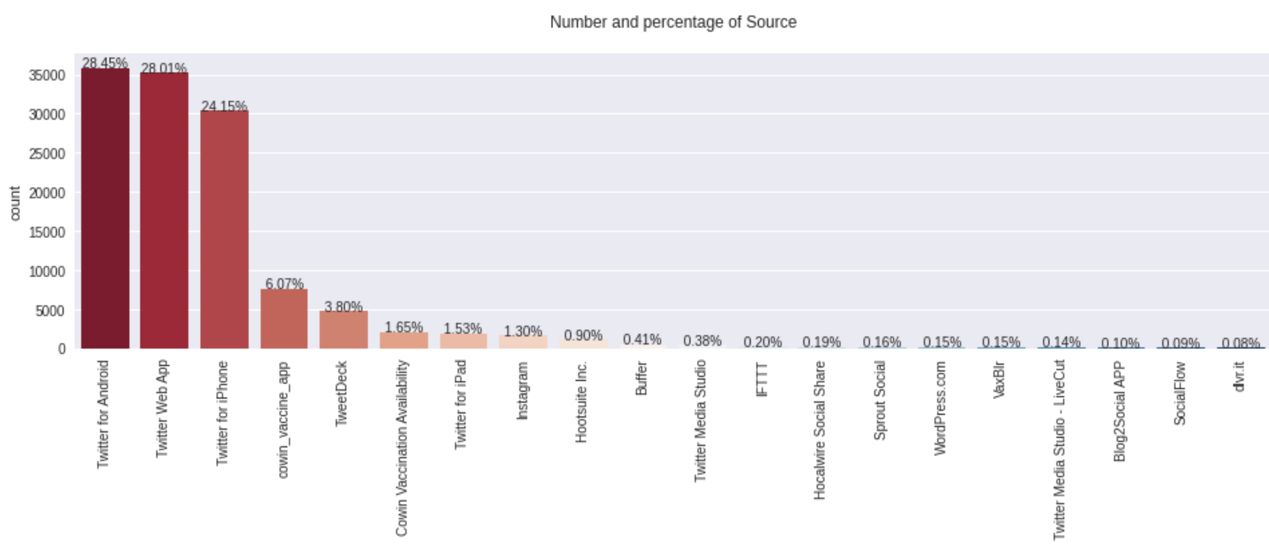


Figure 8: Sources of vaccine reactions

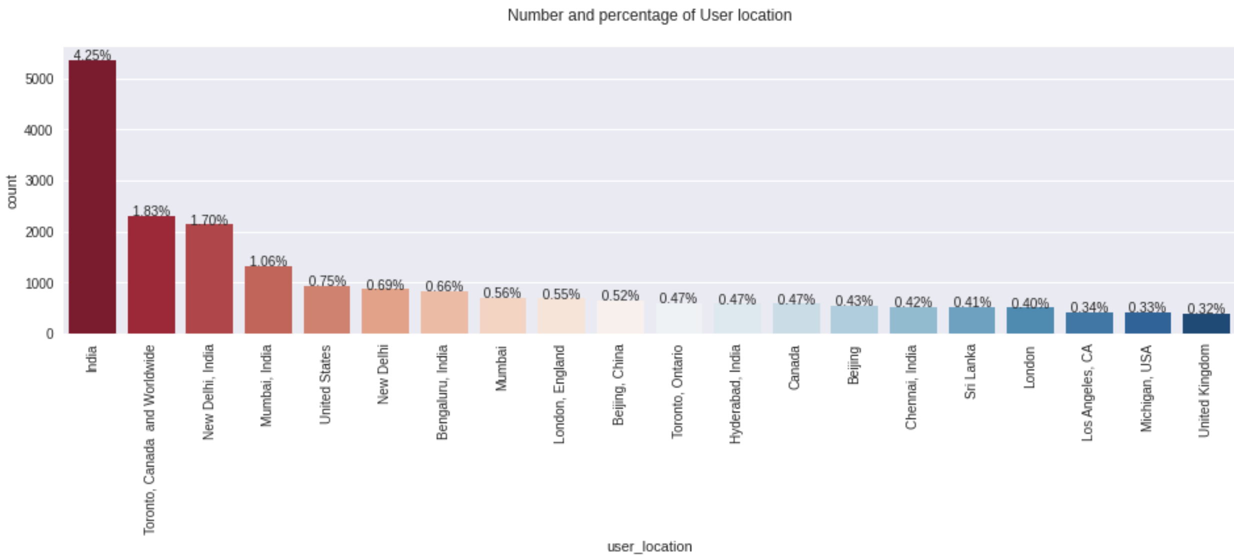
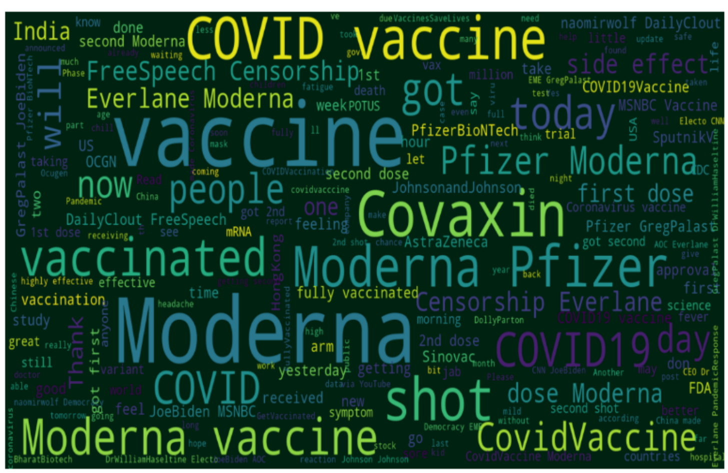


Figure 9: User Location of tweets

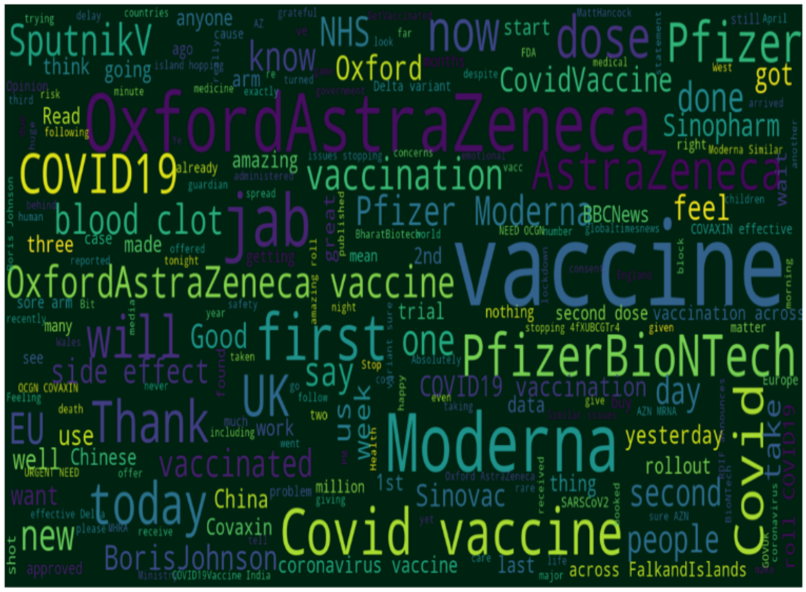
3.3 Country wise 'Prevalent Word' usage

Here in collecting major utterances of vaccine-related sentiments, we prioritize the users who have most of the tweets according to the location. So using word clouds we visualized the terms that were most used in the USA, UK, Canada, and India respectively in Fig. 10, Fig. 11, Fig. 12, and Fig. 13, to have a clear concept of the overall scenario.



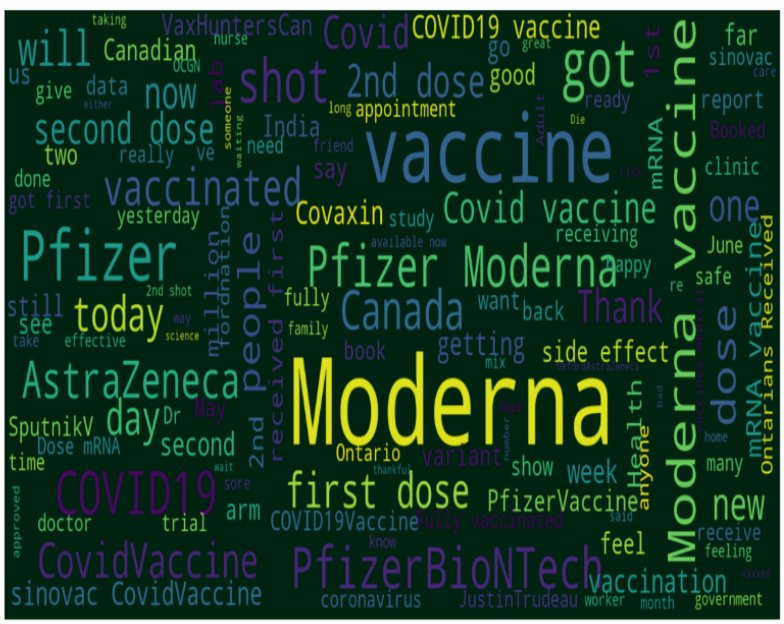
Prevalent words in tweets from US

Figure 10: Terms used most in the US



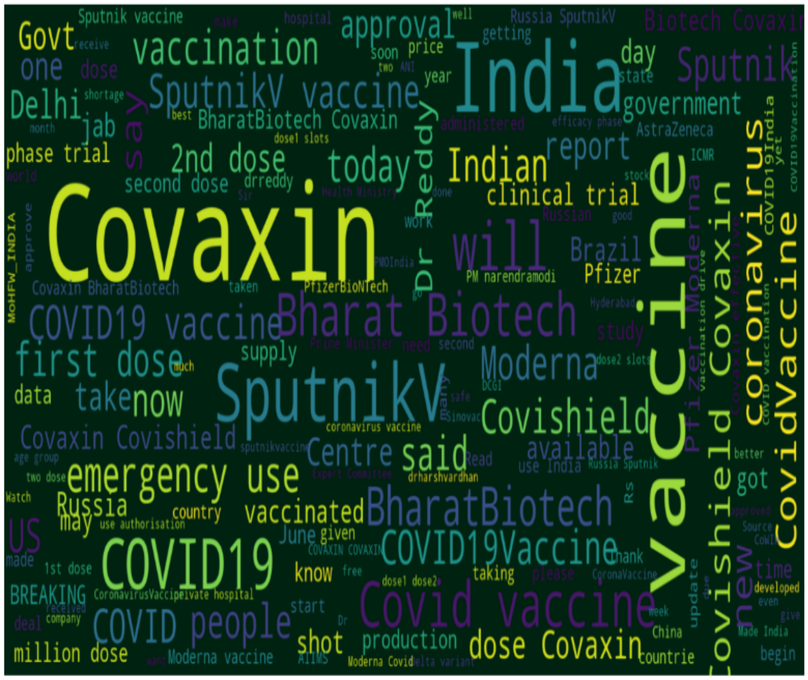
Prevalent words in tweets from UK

Figure 11: Terms used most in the UK



Prevalent words in tweets from Canada

Figure 12: Terms used most in Canada



Prevalent words in tweets from India

Figure 13: Terms used most India

Here, the USA, UK are the producers of particular vaccines so their names came into the spotlight in their tweets. Some alarming terms like ‘blood clot’, ‘2nd dose’, ‘Moderna’, ‘Covaxin’, ‘Pfizer’, ‘death;, ; hope’, ‘emergency’, ‘ joebiden’, ‘narendramodi’, ‘approved’ can be seen in these clouds.

3.4 Hashtag counts per tweet

Hashtags are important and highly used by maximum users on Twitter, and sometimes hashtags carry some significant meanings of particular events or trends. In Fig. 14 we showed the usage of hashtags.

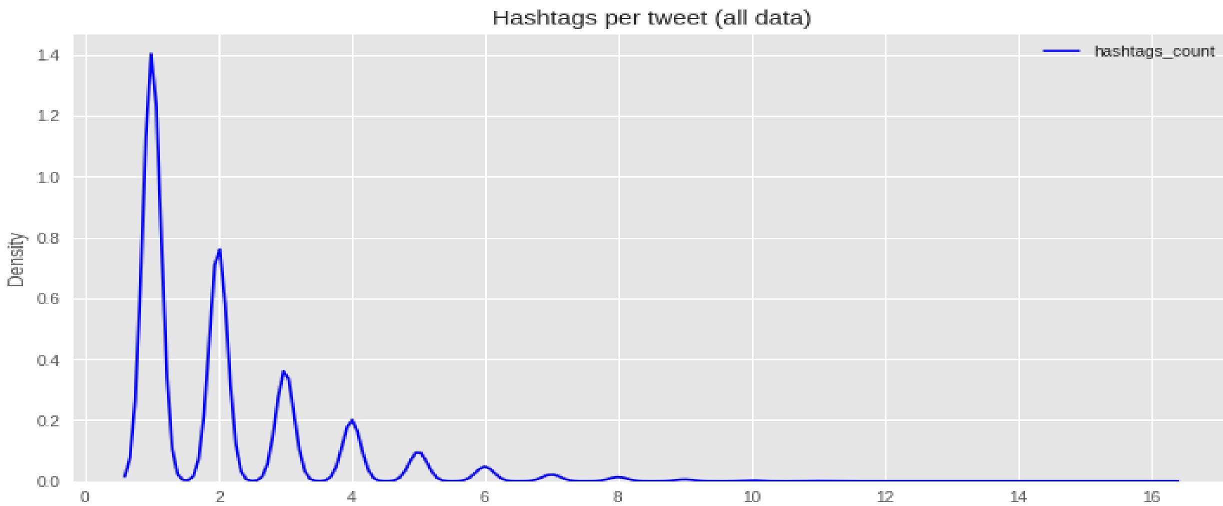


Figure 14: Hashtags per tweet for all data

3.5 Timeline of tweet reactions

The number of tweets varies from time to time. We tried to show the changes in the number of tweets during the start of the vaccination process till the latest time frame in Fig. 15.

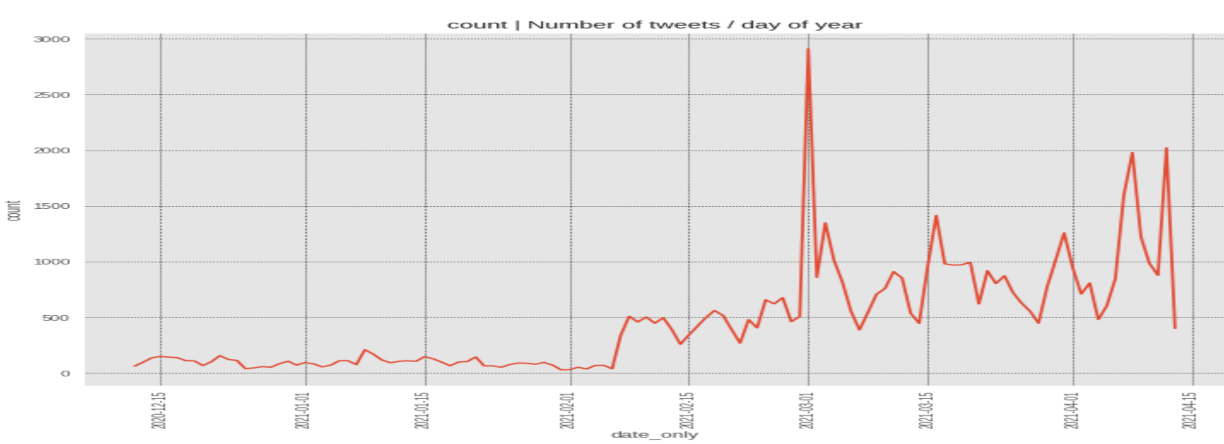


Figure 15: Day wise of ‘Number of Tweets

3.6 Sentiment Analysis and Evaluation

This is one of the vital parts of our research, we are focusing on the Positive, Negative and Neutral percentages to understand how emotions vary.

3.6.1 Numbers and Percentage of Sentiment criteria

In Fig. 16 we demonstrated the numbers of tweets classified as Positive, Negative, And Neutral by Green bar charts and percentages of these 3 categories with Blue bar charts. For the charts colored Green, the x-axis determined the numbers of tweets and the y-axis determined the sentiment classes. And in Blue charts, probability distributions are shown in the x-axis and in the sentiment classes on the y-axis.

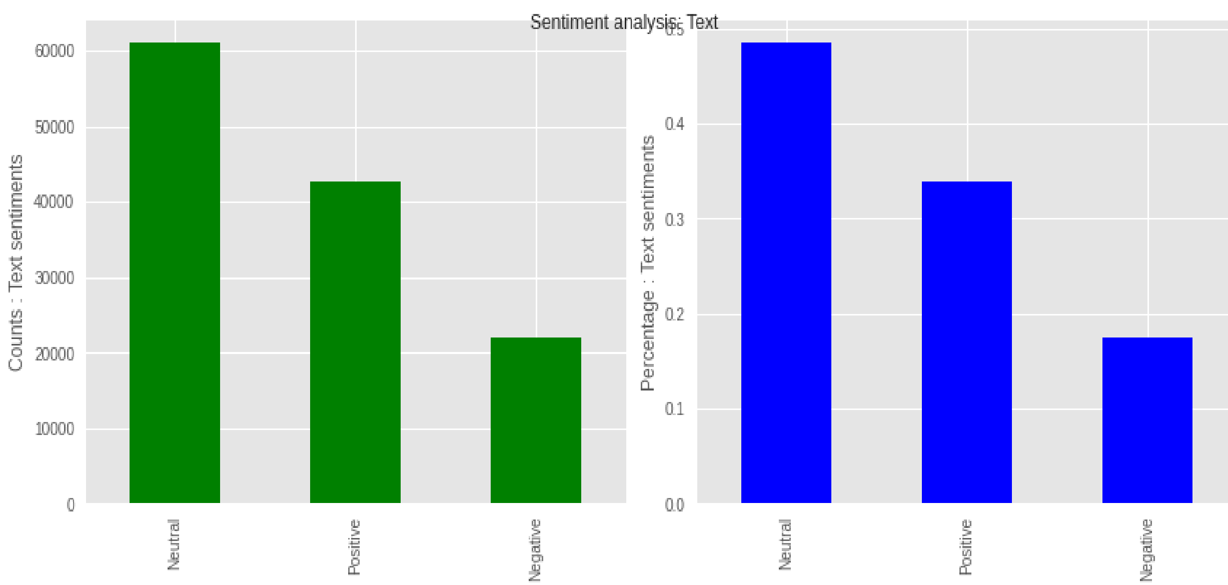


Figure 16: Numbers and Percentages of Sentiment criteria.

From the datasets, there were 125,906 tweets that we analyzed through lexicon-based VADER and segregated them into 3 parameters: Positive, Negative, and Neutral. There were 61047 numbers of Neutral tweets with a percentage of 48.49%.

We found 42,765 numbers of Positive tweets with a percentage of 33.96% and 22,094 numbers of Negative tweets with a percentage of 17.55%. Neutral tweets are the majority here, negative reactions are also not that less and that indicates that confusions, conflicts, and uncertainties related to COVID-19 vaccination procedures are still there. Optimistically a good amount of people are still positive about these vaccines and that may motivate the vaccine producers and the rule-makers to control the coronavirus.

3.6.2 Timeline of these sentiments

As we said earlier, sentiment or emotions evolve and transform with the time period. In this research, we showed how the reactions changed with time and how it fluctuated in different time intervals. We showed in Fig. 17 how sentiment changes or shuffles as time went by. 3 sentiment classes are shown in 3 colors. For positive, negative, and neutral the colors we used are green, red, and blue respectively.

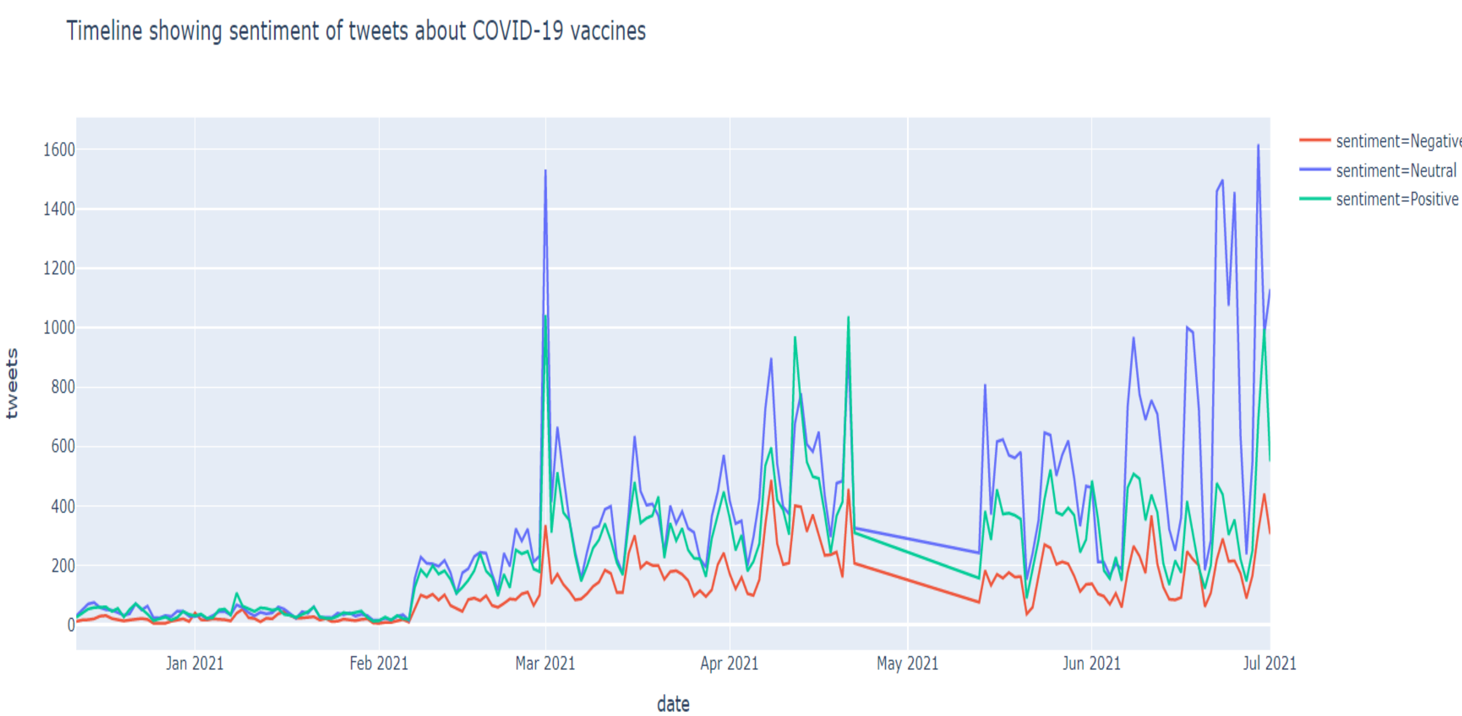
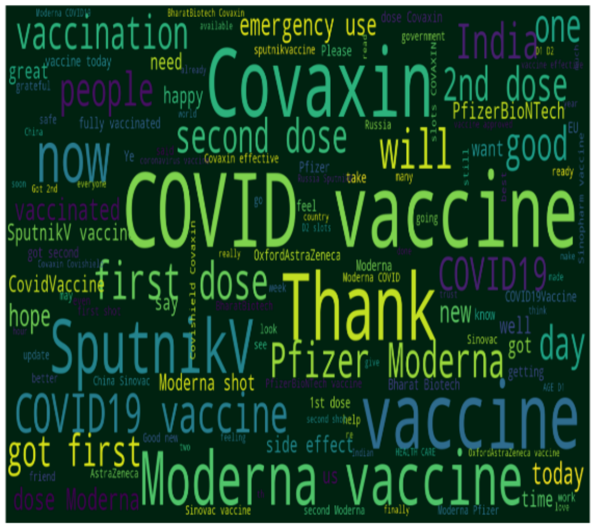


Figure 17: Sentiment variation with Time

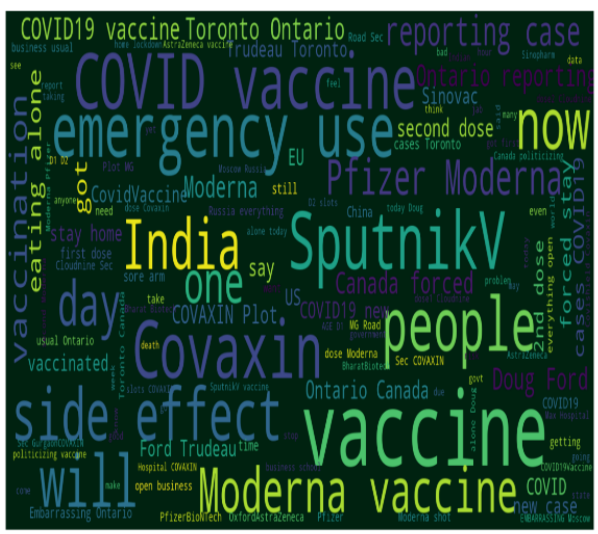
3.6.3 Sentiment words according to polarities

Using textcloud, the specific words or terms are classified in 3 polarity groups are shown in Fig. 18.



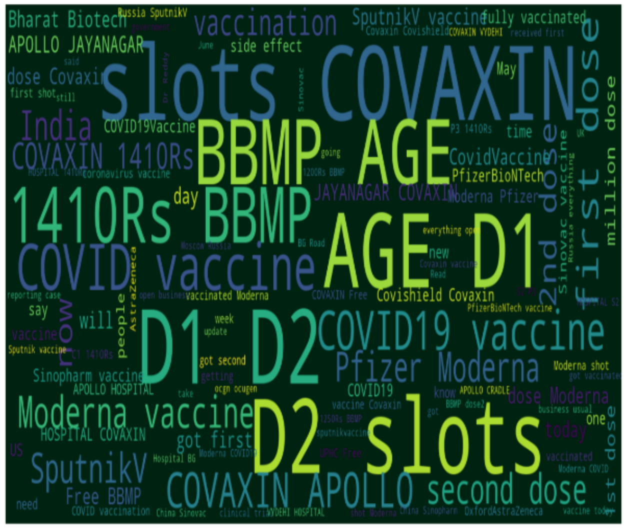
Prevalent words in tweets (Positive sentiment)

(p)



Prevalent words in tweets (Negative sentiment)

(q)



Prevalent words in tweets (Neutral sentiment)

(r)

Figure 18: Words according to Positive (a), Negative (q), and Neutral (r)

3.7 Performance Evaluation

For evaluating model performance, we used RNN based architectures LSTM and Bi-LSTM that performed efficiently in our experiment. In the Data sequencing and splitting part, we convert processed data into vectors using tokenizing and transform values into target labels. We used the train test split method from the scikit-learn library. Our splitting data shapes are x_train: 94429, x_test: 31477, y_train: 94429, y_test: 31477. In this section, we presented the performance metrics of both models and their prediction capabilities.

3.7.1 Performance Analysis with LSTM and Bi-LSTM

Long Short-term Memory (LSTM) played a vital role in evaluating model performance in our experiment. To train our model with both LSTM and Bi-LSTM, we used these factors listed in Table. 1 below.

Table 1: Training Parameters for LSTM and Bi-LSTM

Training factors	Components
Platform	Google COLAB
GPU	Colab GPU (NVIDIA Tesla K80)
Optimizer	RMSProp
Loss	Categorical Cross-Entropy
Epoch	10
Activation	Softmax

In Fig. 19 and Fig. 20, we exhibited model accuracy and model loss for both training & test data of the LSTM model. We found the accuracy of the validation set is 90.59% for the LSTM architecture.

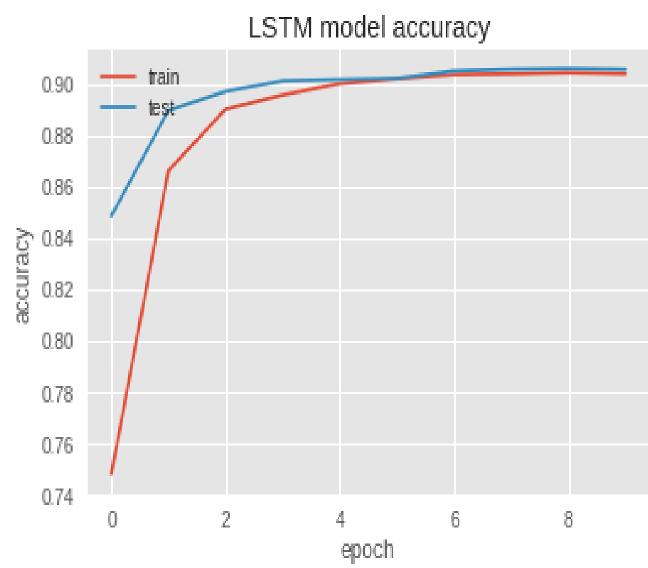


Figure 19: Accuracy of LSTM model

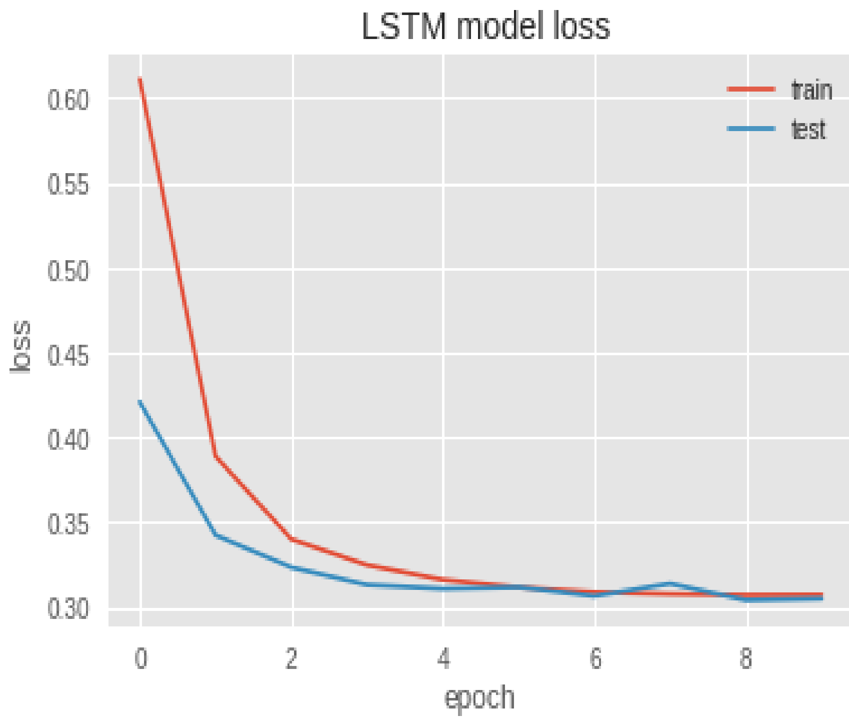


Figure 20: Loss of LSTM model

We also showed model accuracy and model loss for both training & test data of the Bi-LSTM model in Fig. 21 and Fig. 22. Here, for the Bi-LSTM model, we got an accuracy of 90.83% for both the validation set and the test set.

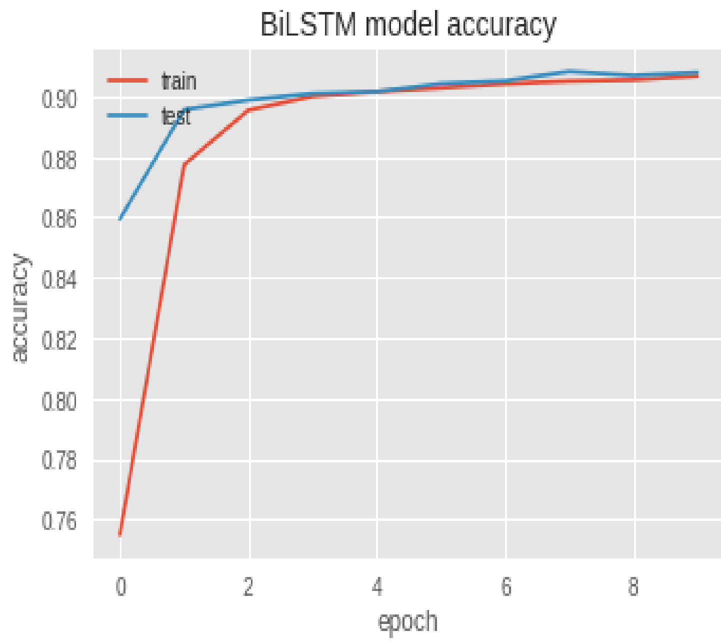


Figure 21: Accuracy of Bi-LSTM model

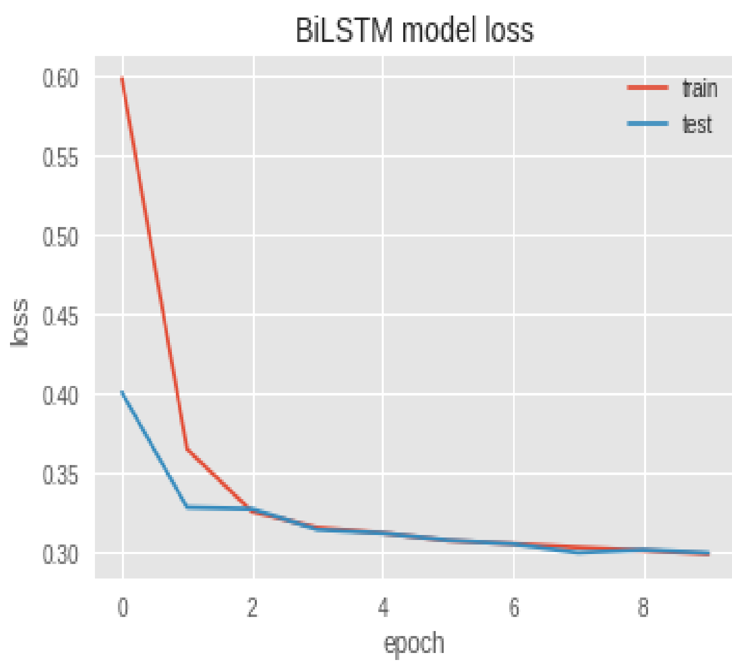


Figure 22: Loss of Bi-LSTM model

After implementing B-LSTM, the accuracy of the model increased a little bit from 90.59% to 90.83%. Here, we also showed some more performance analysis, below we showed some performance classification metrics like Precision, Recall, and F-1 score for both models in Fig. 23 and Fig. 24, and we demonstrated confusion matrix in Fig. 25.

	precision	recall	f1-score	support
0	0.94	0.94	0.94	15145
1	0.85	0.79	0.82	5495
2	0.88	0.92	0.90	10837
accuracy			0.91	31477
macro avg	0.89	0.88	0.89	31477
weighted avg	0.91	0.91	0.91	31477

Figure 23: Performance metrics for LSTM

	precision	recall	f1-score	support
0	0.94	0.95	0.94	15145
1	0.85	0.79	0.82	5495
2	0.89	0.91	0.90	10837
accuracy			0.91	31477
macro avg	0.89	0.88	0.89	31477
weighted avg	0.91	0.91	0.91	31477

Figure 24: Performance metrics for Bi-LSTM

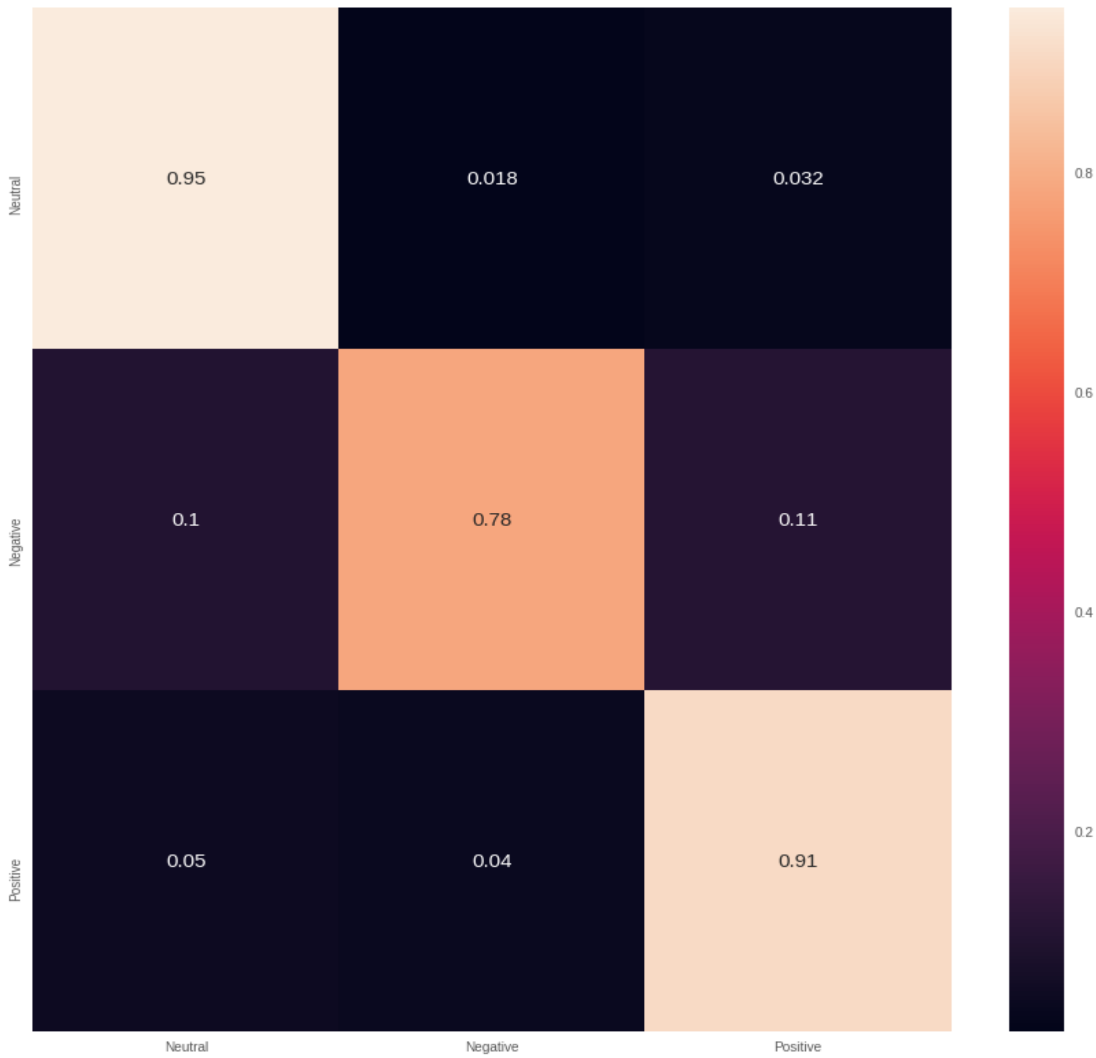


Figure 25: Confusion Matrix for Sentiments

3.7.2 Sentiment prediction table

We took the input from the tweets randomly and checked how well they can predict. In Table. 2 We showed some outcomes.

Table 2: Assessment of sentiment prediction

	Text	Sentiment
0	The same folks said daikon paste could treat a cytokine storm #PfizerBioNTech	Positive
1	While the world has been on the wrong side of history this year hopefully the biggest vaccination effort weve ev	Negative
2	#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid19 Russian vaccine is created to last 2 4 years	Positive
3	Facts are immutable, Senator, even when you're not ethically sturdy enough to acknowledge them 1 You were born i	Neutral
4	Does anyone have any useful advice/guidance for whether the COVID vaccine is safe whilst breastfeeding	Neutral
5	it is a bit sad to claim the fame for success of #vaccination on patriotic competition between USA Canada UK and	Positive
6	There have not been many bright days in 2020 but here are some of the best	Positive
7	Covid vaccine You getting it #CovidVaccine #covid19 #PfizerBioNTech #Moderna	Neutral
8	#CovidVaccine States will start getting #COVID19Vaccine Monday #US says #pakustv #NYC #Healthcare #GlobalGoals	Neutral

4. Future Work and Conclusion

In recent times, we are observing some confusion or conflicts regarding the vaccination of COVID-19 among the mass of people. Many people took the decision to vaccinate themselves and a good number of people are still confused, many of them are frightened and many of them directly refused to be vaccinated. In our research, we've analyzed the sentiment of these people regarding the vaccination process. We used Deep Learning techniques to measure the basic stats about people who are positive, who are negative, and who are neutral in this vaccination procedure from the Covid-19 vaccination available datasets. We'd also theoretically analyzed the reason behind these conflicts or confusions.

This research will help the health researchers to get the proper knowledge of the issues regarding the vaccination process. The companies who produce vaccines, Governments or Health Ministries of different countries, or policymakers of this sector like WHO or others, can have a proper idea about whether their vaccine is effective or not, and the percentage of this effect as well. They can understand that, in which sector, they have to improve so that people can have faith in this vaccination process. Moreover, they may have a clear idea about how many doses they should be ready with according to the wishes of the masses. In Health Science Research, our project will be an added benefit to finding out the proper scenario and pros & cons of the vaccination process of Covid-19. We believe we'll be a little but effective part to help the upfront fighters against these NOVEL coronaviruses and keep our lives healthy and safe.

Acknowledgment: Authors would like to thank the Department of Electrical and Computer Engineering of North South University.

Funding: Authors received no funds from any external sources, it’s a self-funded project and all related costs belonged to the authors of this paper.

Conflicts of Interest: “The authors declare that they have no conflicts of interest to report regarding the present study.”

Bibliography

References

[1] V. A. Grossman, “The COVID-19 Vaccine: Why the Hesitancy,” *Journal of Radiology Nursing*, vol. 40, no. 2, pp. 116–119, Jun. 2021.

[2] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao, “Author Correction: A global database of COVID-19 vaccinations,” *Nature Human Behaviour*, Jun. 2021.

[3] D. Trapani and G. Curigliano, “COVID-19 vaccines in patients with cancer,” The Lancet Oncology, vol. 22, no. 6, pp. 738–739, Jun. 2021.

[4] B. Cabanillas and N. Novak, “Allergy to COVID-19 vaccines: A current update,” Allergology International, vol. 70, no. 3, pp. 313–318, Jul. 2021.

[5] S. Pullan and M. Dey, “Vaccine hesitancy and anti-vaccination in the time of COVID-19: A Google Trends analysis,” Vaccine, vol. 39, no. 14, pp. 1877–1881, Apr. 2021.

[6] W. Jennings, G. Stoker, H. Bunting, V. O. Valgarðsson, J. Gaskell, D. Devine, L. McKay, and M. C. Mills, “Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy,” *Vaccines*, vol. 9, no. 6, p. 593, Jun. 2021.

[7] G. Troiano and A. Nardi, “Vaccine hesitancy in the era of COVID-19,” *Public Health*, vol. 194, pp. 245–251, May 2021.

[8] A. Bendau, J. Plag, M. B. Petzold, and A. Ströhle, “COVID-19 vaccine hesitancy and related fears and anxiety,” *International Immunopharmacology*, vol. 97, p. 107724, Apr. 2021.

[9] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, “An analysis of COVID-19 vaccine sentiments and opinions on Twitter,” *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, May 2021.

[10] T. Na, W. Cheng, D. Li, W. Lu, H. Li, “Insight from NLP Analysis: COVID-19 Vaccines Sentiments on Social Media,” *arXiv preprint*, arXiv:2106.04081, Jun. 2021.

[11] D. Amanatidis, I. Mylona, I. (Eirini) Kamenidou, S. Mamalis, and A. Stavrianea, “Mining Textual and Imagery Instagram Data during the COVID-19 Pandemic,” *Applied Sciences*, vol. 11, no. 9, p. 4281, May 2021.

[12] D. A. Nurdeni, I. Budi and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer," *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pp. 122-127, 2021.

[13] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, “Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes,” *Information*, vol. 12, no. 5, p. 204, May 2021.

[14] S. W. Kwok, S. K. Vadde, and G. Wang, “Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis,” *Journal of Medical Internet Research*, vol. 23, no. 5, May 2021.

[15] “Kaggle: Your Machine Learning and Data Science Community,” *Kaggle*. [Online]. Available: https://www.kaggle.com/. [Accessed: 10-Jul-2021].

[16] G. Preda, “All COVID-19 Vaccines Tweets,” *Kaggle*. [Online]. Available: https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets. [Accessed: 13-Jul-2021].

[17] S. Hochreiter and J. Schmidhuber. “LONG SHORT-TERM MEMORY.” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Jan. 1997.

[18] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp.2673–2681, 1997.

[19] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, Jul. 2005.

[20] A. Graves, N. Jaitly and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *Proc. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273-278, 2013.

Implementation

Used Libraries

```
import warnings
warnings.simplefilter("ignore")
import re
import numpy as np
import pandas as pd
import matplotlib
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('seaborn')
import gensim  import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
from textblob import TextBlob
from wordcloud import WordCloud, STOPWORDS
from nltk.stem import WordNetLemmatizer
from nltk.stem import LancasterStemmer
from nltk.tokenize.treebank import TreebankWordDetokenizer

import tensorflow as tf
import plotly.express as px
```

```
from nltk.corpus import stopwords
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from nltk.stem import LancasterStemmer
from nltk.tokenize.treebank import TreebankWordDetokenizer
from collections import Counter
from wordcloud import WordCloud
from nltk.corpus import stopwords
from gensim.utils import simple_preprocess
from nltk.corpus import stopwords
import gensim
from sklearn.model_selection import train_test_split
import spacy
```

```
import pickle
from keras.preprocessing.text import Tokenizer
import seaborn
from sklearn.metrics import confusion_matrix
import keras
```

 # Show Word Cloud

```
# Make a word Cloud from dataset
stopwords = set(STOPWORDS)

def show_wordcloud(data, title = None):
    wordcloud = WordCloud(
        background_color='white',
        stopwords=stopwords,
        max_words=50,
        max_font_size=40,
        scale=5,
        random_state=1
    ).generate(str(data))

    fig = plt.figure(1, figsize=(10,10))
    plt.axis('off')
    if title:
        fig.suptitle(title, fontsize=20)
        fig.subplots_adjust(top=2.3)

    plt.imshow(wordcloud)
    plt.show()

def show_wordcloud(data, title=""):
    text = " ".join(t for t in data.dropna())
    stopwords = set(STOPWORDS)
    stopwords.update(["t", "co", "https", "amp", "U"])
    wordcloud = WordCloud(stopwords=stopwords, scale=4,
max_font_size=40,
max_words=700,background_color="#002210").generate(text)
    fig = plt.figure(1, figsize=(20,20))
```

```
    plt.axis('off')
    fig.suptitle(title, fontsize=20)
    fig.subplots_adjust(top=2.3)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.style.use('ggplot')
    plt.show()
```

Hashtag Analysis

```
def plot_features_distribution(features, title, df, isLog=False):
    plt.figure(figsize=(12,6))
    plt.title(title)
    for feature in features:
        if(isLog):
            sns.distplot(np.log1p(df[feature]),kde=True,hist=False,
bins=120, label=feature,color='blue')
        else:
            sns.distplot(df[feature],kde=True, hist=False, bins=120,
label=feature,color='blue')
    plt.xlabel('')
    plt.legend()
    plt.show()
tweets_df['hashtags'] = tweets_df['hashtags'].replace(np.nan,
"['None']", regex=True)
tweets_df['hashtags'] = tweets_df['hashtags'].apply(lambda x:
x.replace('\\N',''))
tweets_df['hashtags_count'] = tweets_df['hashtags'].apply(lambda x:
len(x.split(',')))
plot_features_distribution(['hashtags_count'], 'Hashtags per tweet
(all data)', tweets_df)
tweets_df['hashtags_individual'] = tweets_df['hashtags'].apply(lambda
x: x.split(','))
from itertools import chain
all_hashtags =
set(chain.from_iterable(list(tweets_df['hashtags_individual'])))
print(f"There are totally Hastags of : {len(all_hashtags)}")
```

Process Data

```
def depure_data(data):

    #Removing URLs with a regular expression
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    data = url_pattern.sub(r'', data)

    # Remove Emails
    data = re.sub('\S*@\S*\s?', '', data)

    # Remove new line characters
    data = re.sub('\s+', ' ', data)

    # Remove distracting single quotes
    data = re.sub("\'", "", data)

    return data

temp = []
#Splitting pd.Series to list
data_to_list = tweets_df["text"].values.tolist()
for i in range(len(data_to_list)):
    temp.append(depure_data(data_to_list[i]))
list(temp[:5])
def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence),
deacc=True))
data_words = list(sent_to_words(temp))
print(data_words[:10],'\n')

def detokenize(text):
    return TreebankWordDetokenizer().detokenize(text)
data = []
for i in range(len(data_words)):
    data.append(detokenize(data_words[i]))
print(data[:5])
```

Convert Labels

```
data = np.array(data)
#Create a sentiment column
tweets_df["sentiment"] = np.nan

# borrowed from
https://www.kaggle.com/pashupatigupta/sentiments-transformer-vader-em
bedding-bert
from nltk.sentiment import SentimentIntensityAnalyzer
import nltk
nltk.download('vader_lexicon')

sia = SentimentIntensityAnalyzer()
def find_sentiment(post):
    if sia.polarity_scores(post)["compound"] > 0:
        return "Positive"
    elif sia.polarity_scores(post)["compound"] < 0:
        return "Negative"
    else:
        return "Neutral"

tweets_df['sentiment'] = tweets_df['text'].apply(lambda x:
find_sentiment(x))
tweets_df.head(10)
def plot_sentiment(df, feature, title):
    counts = df[feature].value_counts()
    percent = counts/sum(counts)

    fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(12, 5))

    counts.plot(kind='bar', ax=ax1, color='green')
    percent.plot(kind='bar', ax=ax2, color='blue')
    ax1.set_ylabel(f'Counts : {title} sentiments', size=12)
    ax2.set_ylabel(f'Percentage : {title} sentiments', size=12)
    plt.suptitle(f"Sentiment analysis: {title}\n")
    plt.tight_layout()
    plt.show()
```

Labels Encoding

```
labels = np.array(tweets_df["sentiment"])
y = []
for i in range(len(labels)):
    if labels[i] == 'Neutral':
        y.append(0)
    if labels[i] == 'Negative':
        y.append(1)
    if labels[i] == "Positive":
        y.append(2)
y = np.array(y)
labels = tf.keras.utils.to_categorical(y, 3, dtype="float32")
del y

#Convert the word to vec
max_words = 5000
max_len = 200

tokenizer = Tokenizer(num_words=max_words)
tokenizer.fit_on_texts(data)
sequences = tokenizer.texts_to_sequences(data)
tweets = pad_sequences(sequences, maxlen=max_len)
print(tweets)

print(labels)

#Splitting the data
X_train, X_test, y_train, y_test = train_test_split(tweets,labels,
random_state=0)
print (f"Our data split form:\n")
print(f"X_train: ",len(X_train))
print(f"X_test: ",len(X_test))
print(f"y_train: ",len(y_train))
print(f"y_test: ", len(y_test))
```

LSTM Model

```
model1 = Sequential()
model1.add(layers.Embedding(max_words, 20))
model1.add(layers.LSTM(15,dropout=0.5))
model1.add(layers.Dense(3,activation='softmax'))

model1.compile(optimizer='rmsprop',loss='categorical_crossentropy',
metrics=['accuracy'])
checkpoint1 = ModelCheckpoint("best_model1.hdf5",
monitor='val_accuracy', verbose=1,save_best_only=True, mode='auto',
period=1,save_weights_only=False)
history1 = model1.fit(X_train, y_train,
epochs=10,validation_data=(X_test, y_test),callbacks=[checkpoint1])

# summarize history for accuracy
plt.plot(history1.history['accuracy'])
plt.plot(history1.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()

# summarize history for accuracy
plt.plot(history1.history['loss'])
plt.plot(history1.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper right')
plt.show()
lstmpredict = model1.predict(X_test)
print(lstmpredict)
test_loss, test_acc = best_model.evaluate(X_test, y_test, verbose=2)
print('Model accuracy: ',test_acc)

print(classification_report(np.argmax(y_test,
axis=1),np.argmax(predictions, axis=1)))
```

Bidirectional LSTM Model

```
model = Sequential()
model.add(layers.Embedding(max_words, 40, input_length=max_len))
model.add(layers.Bidirectional(layers.LSTM(20,dropout=0.6)))
model.add(layers.Dense(3,activation='softmax'))
model.compile(optimizer='rmsprop',loss='categorical_crossentropy',
metrics=['accuracy'])
checkpoint2 = ModelCheckpoint("best_model2.hdf5",
monitor='val_accuracy', verbose=1,save_best_only=True, mode='auto',
period=1,save_weights_only=False)
# Train the Model
history = model.fit(X_train, y_train,
epochs=10,validation_data=(X_test, y_test),callbacks=[checkpoint2])
# summarize history for accuracy
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
# summarize history for accuracy
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper right')
plt.show()
test_loss, test_acc = best_model.evaluate(predictions, y_test,
verbose=2)
print('Model accuracy: ',test_acc)

# Plot the classification_report
from sklearn.metrics import classification_report
print(classification_report(np.argmax(y_test,
axis=1),np.argmax(predictions, axis=1)))
```

Confusion Matrix & Prediction

```
matrix = confusion_matrix(y_test.argmax(axis=1),
np.around(predictions, decimals=0).argmax(axis=1))
conf_matrix = pd.DataFrame(matrix, index =
['Neutral','Negative','Positive'],columns =
['Neutral','Negative','Positive'])

#Normalizing
conf_matrix = conf_matrix.astype('float') /
conf_matrix.sum(axis=1)[:, np.newaxis]
plt.figure(figsize = (15,15))
seaborn.heatmap(conf_matrix, annot=True, annot_kws={"size": 15})

sequence = tokenizer.texts_to_sequences(['the trump administration
failed to deliver on vaccine promises shocker covidiots coronavirus
covidvaccine'])
test = pad_sequences(sequence, maxlen=max_len)
sentiment[np.around(best_model.predict(test),
decimals=0).argmax(axis=1)[0]]

sequence = tokenizer.texts_to_sequences(['this data science article
is the best ever'])
test = pad_sequences(sequence, maxlen=max_len)
sentiment[np.around(best_model.predict(test),
decimals=0).argmax(axis=1)[0]]

sequence = tokenizer.texts_to_sequences(['Facts are immutable Senator
even when youre not ethically sturdy enough to acknowledge them.'])
test = pad_sequences(sequence, maxlen=max_len)
sentiment[np.around(best_model.predict(test),
decimals=0).argmax(axis=1)[0]]
```

END