# COVID-19 Vaccine Tweet Sentiment Analysis

Gourav Verma & Shani Kumar

Data Science, Bellevue University, NE, USA

DSC680-T302: Applied Data Science

Prof Fadi Alsaleem

06/02/2021

**Abstract**

In this research, we performed sentiment analysis of COVID-19 vaccine tweets using a supervised machine learning approach. Social media is extensively used as a common platform to spread news and opinions. Identification of opinions towards COVID-19 vaccines would allow evidence-based decision-making for better counsel and distribution of vaccines. Using social networking site Twitter tweets from across the world were extracted for a set of vaccine-related keywords. These tweets were gathered from Nov-2020 to March-2021. A random sample of the whole dataset was cleaned by applying text pre-processing techniques. As tweets contain emoticons and short words, it is difficult to correctly recognize the context of the tweets. To overcome this issue dataset with existing sentiment labels was used to train the models. Tweet sentiments are classified into three categories Positive, Negative, and Neutral. The sentiments were annotated and supervised learning methods RNNs (Single LSTM and Bidirectional LSTM), and 1D CNN was applied to train the tweet training dataset. The performance of the classifiers is evaluated on accuracy. With the Bidirectional LSTM model, we achieved 74% accuracy, and it was used to implement on the COVID-19 vaccine tweets dataset.

**Keywords:** Tweets, NLP, Sentiment analysis, RNN, LSTM, Bi-LSTM, 1D CNN

Table of Contents

## Introduction

The year 2020 was full of COVID-19 spread across the world, multiple lockdowns, and burnout of essential workers. So far, 2021 has been focused on vaccine distribution. March 11, 2021, was the first anniversary since WHO declared COVID-19 as a global pandemic [3]. As everyone wanted to go back to normal life, people were closely monitoring vaccine developments. During the lockdown, COVID-19 news spread was enormously amplified in social networking platforms. Being a highly popular social platform, Twitter provides a large scale of text data for various research such as sentiment analysis [12]. Involvement of individual views and biasness in the vast fragment of this news is giving growth to (un)intended fake information, negativity, and ambiguity in the human society [13]. These circumstances are gathering the attention of researchers to carry out computational studies for forming a complete representation. This research is focused on sentiment analysis of COVID-19 vaccine tweets using a supervised machine learning approach. Due to the overflowing COVID-19 news and the speedy development of a vaccine, people rationally have many questions on vaccines. Some of them are [4]:

- Do the mRNA vaccines change your DNA?
- Did the vaccine clinical trials skip steps to be completed faster?
- Can the vaccine give you COVID-19?
- Will we need new vaccines if the virus continues to mutate?

Such questions and limited reliable answers lead to confusion and doubts overtaking the vaccine. As per Panacea lab, [5] every day, there are about 4 million tweets a day related to COVID-19. This study utilizes tweets to understand people's sentiments. It will help understand the difference in sentiments for different vaccines and the change in sentiments over time.

These days Natural Language Processing (NLP) is the breeding ground of research in Data Science. Sentiment analysis is one of the most common divisions of NLP. This domain has diversified the way businesses work due to extensive application usage in creating market strategies, opinion polls, chatbots, etc. For sentiment analysis, NLP has made the processing of thousands of text documents in seconds, which will take hours to process manually. The major work in this project is to clean the text data and train the model to understand the language of Twitter. As tweets contain

emoticons and short words, it is difficult to correctly recognize the context of the tweets. Tweet sentiments are classified into three types Positive, Negative, and Neutral. The COVID-19 vaccine tweets dataset contains tweets from all over the world. To train and help the model to understand the tweet language, a tweets dataset with existing sentiment classifications was used. After training the model, the best model was applied to the COVID-19 vaccine tweet dataset.

Applications:

- Policy implementations for public awareness.
- Biotech companies can utilize this analysis to understanding people's response to vaccination for future vaccine rollouts.
- Social media monitoring.
- Market research.

The rest of the paper is organized as follows. The prior studies section contains different related works relevant to this research. Description of datasets is present in the data section. The modeling section contains a methodology and proposed methods. Results are discussed at the end with concluding remarks.

**Prior Studies**

An increase in false propaganda through social networking sites is a global issue. Spreading false news directly impacts the sentiments of individuals and challenges the solutions implied by governments [14]. In some words of Chakraborty K et al. 2020 [15], most of the COVID-19 tweets have positive sentiments, but many people are often engaged in spreading negative sentiments.

Sosa, 2017[9] in his research of tweet sentiments compared different models along with the combined model of CNN+LSTM. The CNN+LSTM model showed higher performance but lower than LSTM. However, the LSTM+CNN model showed the highest accuracy. It concludes that sequencing is important while combining features of different models. Tweet sentiment analysis is an open field with broad analysis approaches from lexicon-based to involving big data platforms [10].

In the sentiment analysis of COVID-19 tweets using supervised machine-learning authors Rustam Et al., 2020 [11] found that LSTM, Bi-LSTM, and CNN-LSTM models on a small dataset give poor performance because of not enough learning path for a stable system. Further, the analysis showed that the extra tree classifiers model showed the highest accuracy of 93%.

## Datasets

### Data Selection

#### *COVID-19 Vaccine Tweets*

For this project, we used tweets about the COVID-19 vaccines distributed in the entire world. Tweets are collected using the tweepy Python package to search Twitter API for the keywords relevant to COVID-19 vaccines. The tweets are about Pfizer/BioNTech, Sinopharm, Sinovac (both Chinese-produced vaccines), Moderna, Oxford/Astrazeneca, Covaxin, and Sputnik V vaccines. For the analysis we selected a 10% random sample of 6.7 Million tweets.

The tweets dataset contains below columns –

- Id: Total 6.7M values
- User_name: Name of user
- User_location: Location of tweet
- User_created: user creation date.
- User_followers: follower count of user
- User_friends: friends count of user.
- **Date**: Tweet date
- **Text**: Tweet text, 6.7M tweets
- Hashtags: hashtags
- Source: web-31%, phone-29%
- Retweets: retweet count

***Twitter Sentiment Dataset***

This dataset was available as part of twitter sentiment analysis competition on Kaggle [6]. This training dataset contains approx. 22K tweet texts with existing sentiment labels. This dataset will be used to train the models. This dataset contains below columns:

- textID - unique ID for each piece of text
- text - the text of the tweet
- sentiment - the general sentiment of the tweet
- selected_text - the text that supports the tweet's sentiment.

***COVID-19 Vaccine Progress Dataset***

For the time-series analysis, we compared sentiments trend with COVID-19 vaccination progress [1]. This data is collected from https://ourworldindata.org/ GitHub repository [2]. We used Tableau to plot vaccination progress and world map.

**Bias and Limitations**

In the dataset, there was no biasness with demographic, gender, and age characteristics. The dataset contains tweets from across the world and about all vaccines getting distributed in different countries. This sentiment analysis is only focused on tweets, so it cannot be applied to the whole population who does not use Twitter or the internet. Due to hardware constraints, only a 10% random sample of the data was used for the analysis. It might cause limited results and not depict the whole picture.

## Methodology

This section provides an overview of the methodology applied in this research. The complete pipeline is shown in Figure 1. Broadly the experiment is carried out in three different phases including *(i)* the Data preparation phase, *(ii)* the Modelling phase, and *(iii)* the Implementation phase. In the first step, we scrapped the tweets related to COVID-19 vaccines from Twitter, cleaned, annotated, and randomly selected for implementation. The second phase involves building a sentiment

classifier model using a training dataset. In the final implementation phase, the best model was applied to the COVID-19 vaccine tweet dataset, and the outcome was used to analyze the sentiment trends.

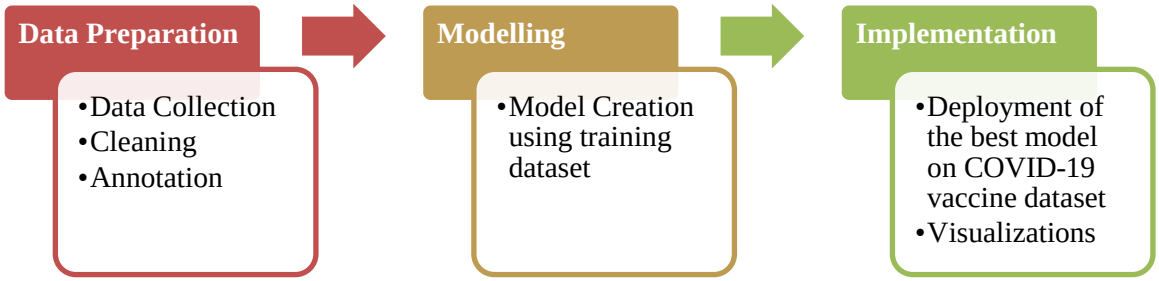


*Figure 1: Pipeline for Sentiment Analysis*

**Data Preparation**

Sentiment analysis helps governments and businesses to understand people's opinions. Tweet texts on both the datasets (COVID-19 vaccine and tweet training) were cleaned, by removing null values, URLs, newline characters, single quotes, and punctuation signs. Post that, text sentences were tokenized into a list of tokens using *gensim* utility. It also converted words into lowercase. Then, with the use of NLTK *TreebankWordDetokenizer*, words were detokenized. The training dataset had three categorical labels for sentiments neutral, negative, and positive. To make the categories understandable for the model, these labels were converted into float values of 0, 1, and 2 for neutral, negative, and positive, respectively. Using *the to_categorical* method from Keras, these float values were converted to a categorical binary class matrix. Similarly, to make it understandable for the model, text data was transformed into 3D float data using Keras tokenizer. Further, the tweet training dataset was split into train and test of 75% and 25% respectively.

## Modelling

Tweets texts are informal language forms. In tweets there could be multiple meanings of the same word and sentences might not be grammatically correct. Hence, traditional rule-based models such as word2vec, Tfidf, and BoW cause week performance for sentiment analysis of tweet texts. For the project, we decided to use LSTM models, which is one of the best RNN models in NLP. RNN takes a word as an input instead of the entire sample. It enables RNN to work with sentences of variable lengths.

For the text classification, we took a machine learning approach to sentiment analysis. The cleaned text was fed, to classifier return categories, e.g., positive, negative, and neutral. For the modeling, the dataset was split to train-test and applied to Single LSTM,[4] Bidirectional LSTM, and 1D CNN [5]. Depending upon the accuracy and speed of each model, the best model was selected to apply to COVID dataset for sentiment classification.

### *RNNs*

#### Single LSTM

In 1997, Hochreiter & Schmidhuber introduced Long Short-Term Memory (LSTM) networks that are RNN networks capable of learning long-term dependencies.[6] Simply LSTM is an improvement over traditional RNN. The main reason behind its design is to avoid the long-term dependency problem of traditional neural networks. Practically it can remember information for long periods. Visually, it looks like shown in Figure 2.

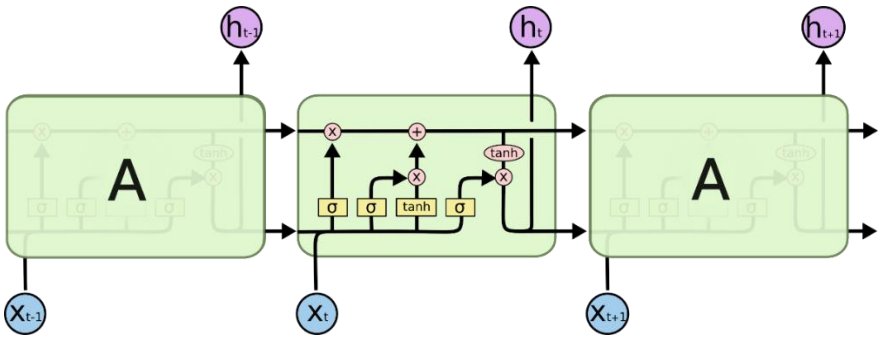


*Figure 2: LSTM Network*

For the LSTM classification model, we used *softmax* activation. It creates discrete probability distribution over the target class for each sample. A dropout of 0.5 was used to avoid overfitting. *Rmsprop* optimizer neutralizes the gradients by employing the scales of recent gradients. As the output consists of more than two categories, for the model, *categorial_crossentropy* was used. Below is the output and model summary of the simple LSTM model. The model showed 71.8% accuracy on the test dataset.

```
645/645 [==============================] - 35s 53ms/step - loss: 0.9119 - accuracy: 0.5929 - val_loss: 0.6578 - val_accurac
y: 0.7182

Epoch 00001: val_accuracy improved from -inf to 0.71820, saving model to mod1.hdf5
```

```
mod1.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 20)          100000
_________________________________________________________________
lstm (LSTM)                  (None, 15)                2160
_________________________________________________________________
dense (Dense)                (None, 3)                 48
=================================================================
Total params: 102,208
Trainable params: 102,208
Non-trainable params: 0
```

*Figure 3 Simple LSTM Model outcome*

**Bidirectional LSTM**

By combining a forward and a backward RNN, a bi-directional RNN is formed. At a given time, prediction is made with a combination of results of both the RNNs. It is an advanced version of traditional LSTM. It is known as bidirection because it trains two LSTMs instead of one and operates in both the direction to frame information from past and future. It enables to provide exceptional sequential modeling performance. As shown in Figure 4 Bi-LSTM has two parallel layers to circulate in two directions.

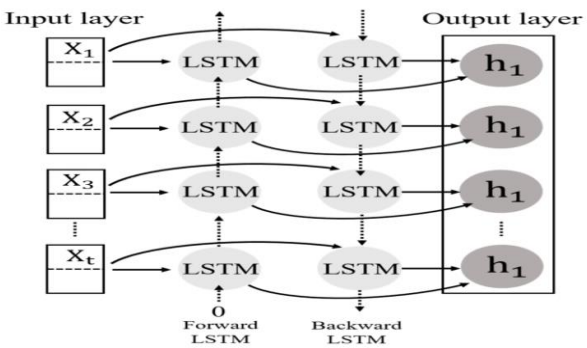


*Figure 4: Structure of Bi-LSTM*

For the Bi-LSTM classification model, we used *softmax* activation. It creates discrete probability distribution over the target class for each sample. A dropout of 0.6 was used to avoid overfitting and to drop the neurons during the training. *Rmsprop* optimizer neutralizes the gradients by employing the scales of recent gradients. As the output consist of more than two categories, for the model, *categorial_crossentropy* was used. Below is the output and model summary of the Bi-LSTM model. The model showed 74.4% accuracy on the test dataset.

```
645/645 [==============================] - 48s 71ms/step - loss: 0.8841 - accuracy: 0.5878 - val_loss: 0.6427 - val_accurac
y: 0.7444

Epoch 00001: val_accuracy improved from -inf to 0.74440, saving model to mod2.hdf5
```

```
mod2.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 200, 40)           200000
_________________________________________________________________
bidirectional (Bidirectional (None, 40)                9760
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 123
=================================================================
Total params: 209,883
Trainable params: 209,883
Non-trainable params: 0
```

*Figure 5: Bi-LSTM Model outcome*

***1D CNN***

Convolutional Neural Networks (CNN) is a standard neural network. CNN uses a convolutional and pooling layer instead of using fully connected hidden layers. Feature maps need to be created from input data to feed into the CNN. It can do a lot of good things. As it consists of multiple layers, the first layer is feed with a bunch of signals. The second layer is feed with some discrete features. Rather than considering the whole set of features, CNN matches part of a signal. Along with learning from high-dimensional data, CNN also learns from small variations. It leads to the requirement for large storage at the time of development. Hence, a down-sampling layer is added in between the layers of convolution.  1D CNN was trained on tweet text in the form of 3D float to identify sentiments of the text. The superior power of CNN helps models learn complicated patterns efficiently in a short timeframe. It results in futuristic capabilities in speech emotion recognition systems.

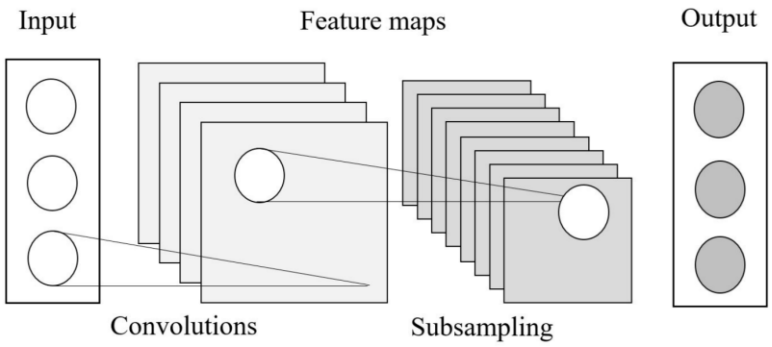


*Figure 6: Structure of 1D CNN*

The two-layered 1D CNN model was trained, for one epoch with a batch size of three. The *rmsprop* optimizer was used, with default parameters. As the research data was normally distributed, accuracy was used as valid matric to assess the model's performance. The model was saved by monitoring the accuracy of the test set. The resulting model was able to make classification with 61.5% accuracy on the test dataset.

```
645/645 [==============================] - 6s 8ms/step - loss: 1.2703 - acc: 0.4789 - val_loss: 0.8991 - val_acc: 0.6150
WARNING:tensorflow:Can save best model only with val_accuracy available, skipping.
```

```
mod3.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 200, 40)           200000
_________________________________________________________________
conv1d (Conv1D)              (None, 195, 20)           4820
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 39, 20)            0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 34, 20)            2420
_________________________________________________________________
global_max_pooling1d (Global (None, 20)                0
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 63
=================================================================
Total params: 207,303
Trainable params: 207,303
Non-trainable params: 0
_________________________________________________________________
```

*Figure 7: 1D CNN Model Outcome*

**Model Deployment**

Among all the three models, Bidirectional LSTM shows the highest accuracy of 74.4%. Below is the confusion matrix for the Bi-LSTM model. Overall, the model's performance is poor. However, through hyperparameter tuning, performance can be improved. Further, the Bi-LSTM model was used for implementation on the COVID-19 vaccine tweet dataset to identify sentiments of tweet texts.

| Algorithms | Activation | Batch Size | Optimizer | Accuracy |
|---|---|---|---|---|
| LSTM | Softmax | 3 | Rmsprop | 71.8% |
| Bi-LSTM | Softmax | 3 | Rmsprop | 74.4% |
| 1D CNN | Relu | 3 | Rmsprop | 61.5% |

*Table 1: Model Comparison*


*Figure 8: Confusion Matrix for Bi-LSTM model*

Before the model deployment, the COVID-19 vaccine tweet dataset was cleaned, tokenized, and detokenized. For the implementation, the 'tweet text' variable was used, and a new sentiment column was created, to store the resulting outcome category. Below are the tweet sentiment counts of different categories after applying the Bi-LSTM model to COVID-19 vaccine tweet dataset sample.

```
# Sentiment Counts
tweet_sent.groupby('Sentiments').nunique()
```

| Sentiments | created_at | text |
|---|---|---|
| **Negative** | 3165 | 3142 |
| **Neutral** | 644539 | 661396 |
| **Positive** | 1703 | 1704 |

*Table 2: Outcome Sentiment Counts*

***Visualizations using Tableau.***

Further, the outcome dataset was used to visualize the sentiment trend using tableau. Below are some observations. Overall, we see a higher number of negative tweet sentiments towards vaccines than positive tweets (Figure-9). On March 1st, there was a spike in the number of tweets related to vaccines. A similar trend can be seen for tweets related to, Covaxin. In India, Prime Minister Narendra Modi announced a kick-start of 2nd round of vaccinations for the age group of 45 and above.

All the vaccines are manufactured by different pharmaceutical companies, and all have different storage conditions, dose amounts, and side effects. Hence, we were further interested in the sentiment trend for individual vaccines. Figure-10 shows the trend in the tweets. Among all vaccines, people tweeted more about the Pfizer vaccine. During the launch of each vaccine, we see an increase in tweet counts.

Sentiment word clouds in Figure-11 show that ‘covid vaccine’ is the most used phrase among all. Majorly ‘thank’, ‘love’, hope’, and ‘happy’ were used in positive tweets. ‘vaccine passport’, ‘johnson’, and ‘rollout covid’ were among negative tweets. ‘wear mask’, ‘million doses’, ‘public health’, and ‘vaccine trial’ are some of the neutral phrases people used in tweets.

As of May-1st China has done the highest number of vaccinations done than any other country.  So far the United States has a greater number of fully vaccinated populations per hundred. More than 30% of the US population are fully vaccinated and around 70% of people have received at least one dose of a vaccine (Figure-12). Even though India is in the 3rd spot of the highest number of vaccinations, considering the overall population of India, vaccination per hundred is still at 10%. Approximately 10% per hundred people in England are fully vaccinated.

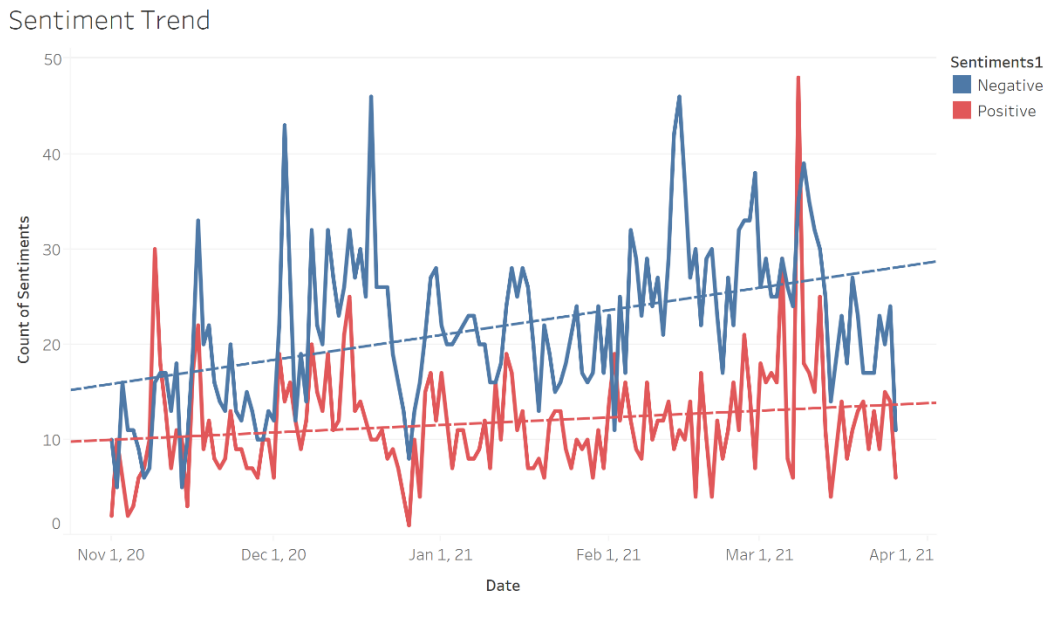


*Figure 9: Sentiment Trend*

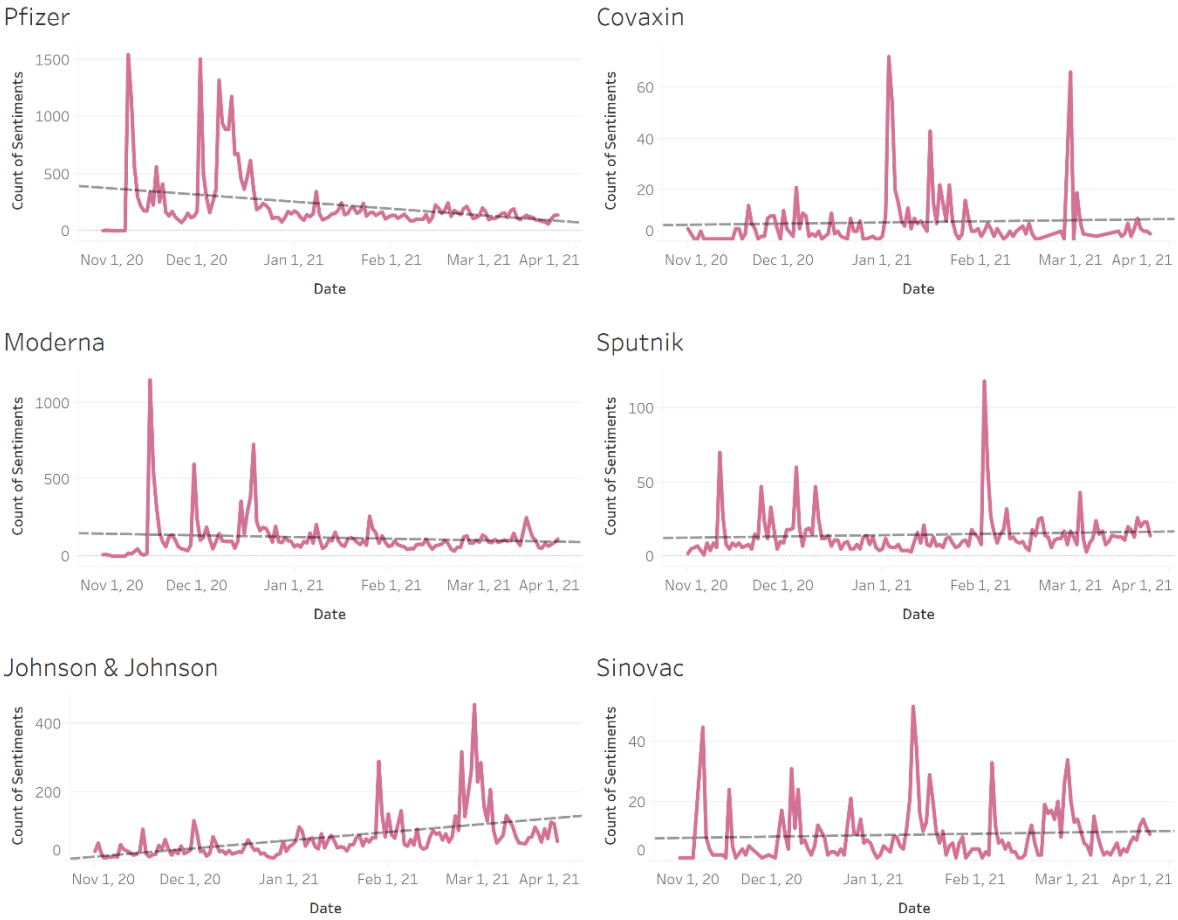


*Figure 10: Vaccine Tweet Trend*

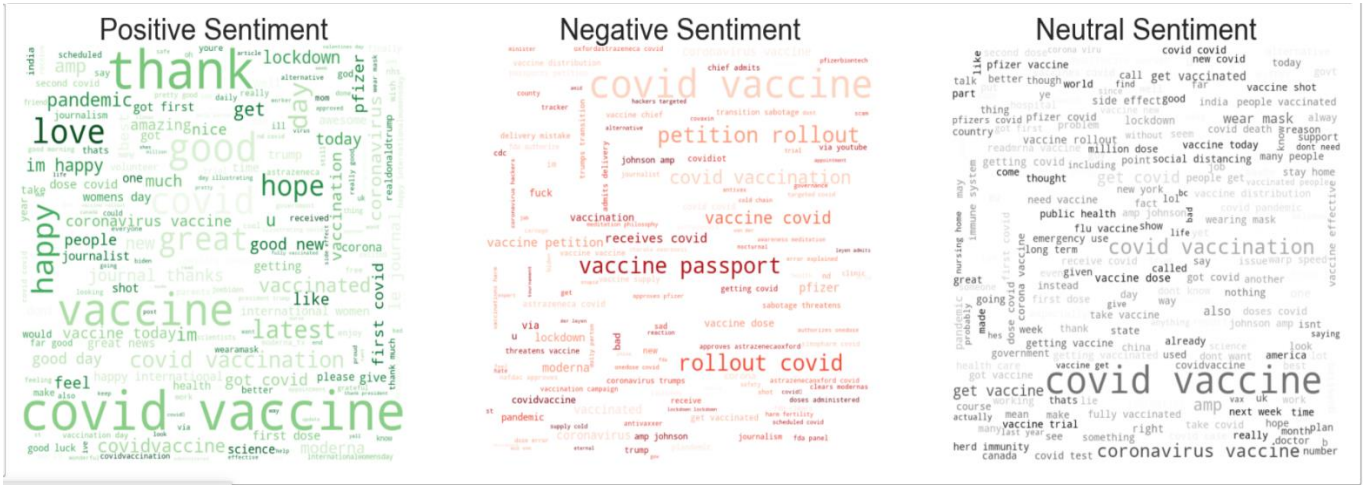


*Figure 11: Sentiment Word Clouds*

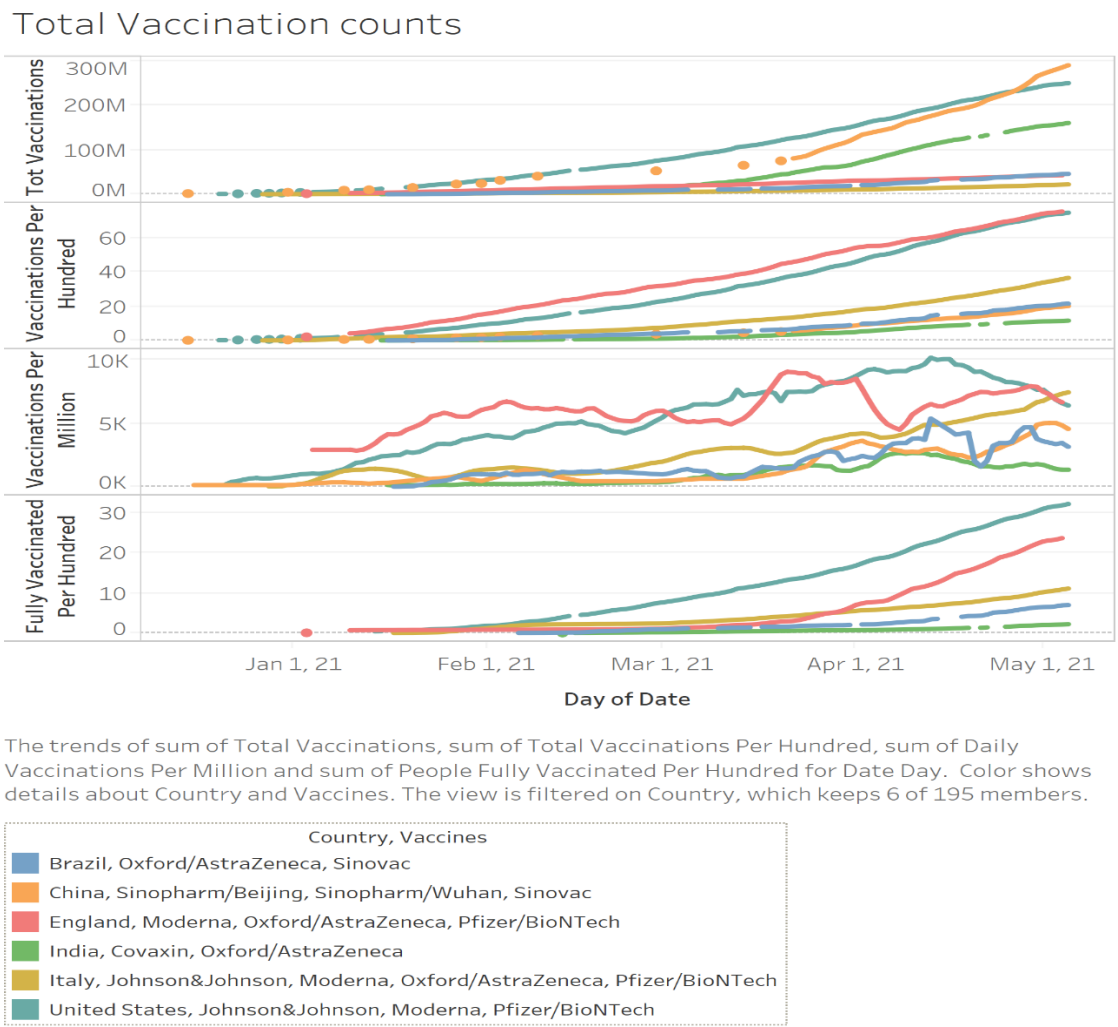


*Figure 12: Vaccination count trend*

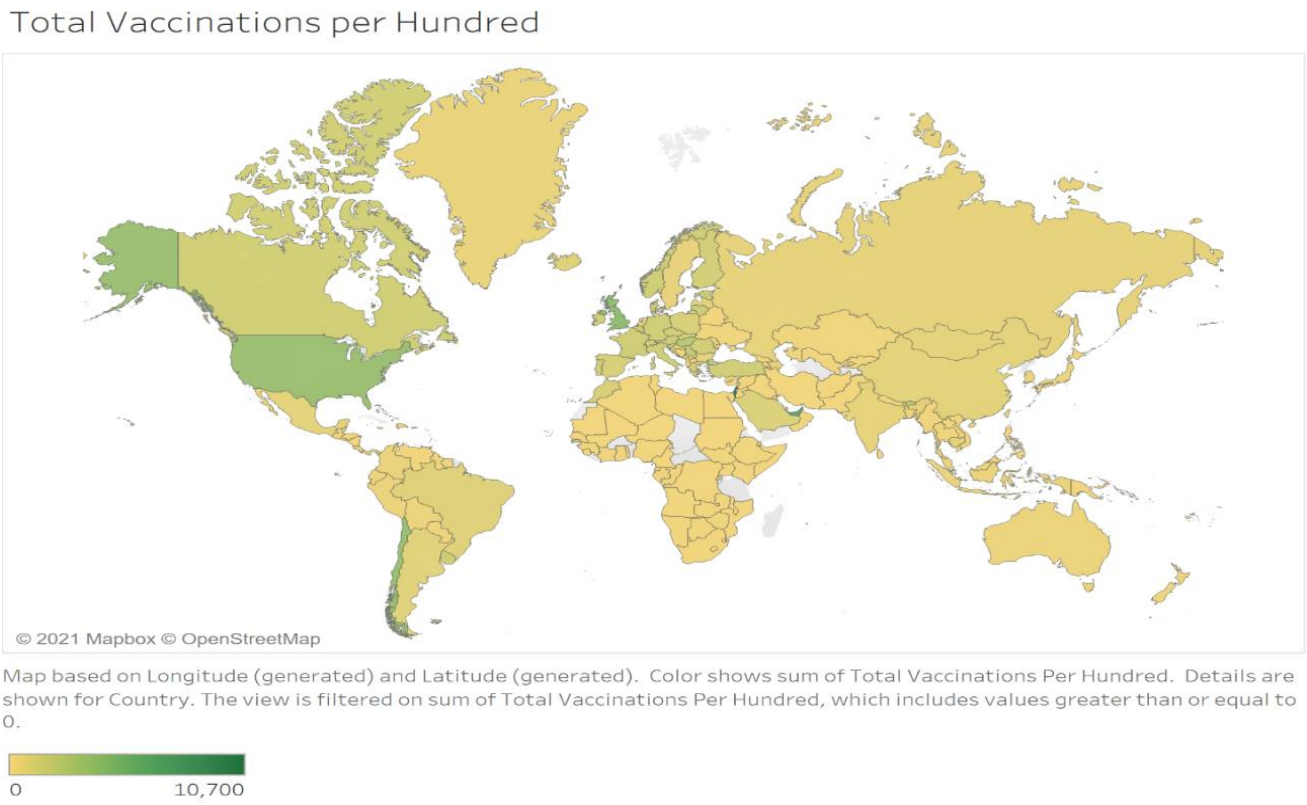


*Figure 13: World map of Total Vaccinations per Hundred – Till May-1st, 2021*

## Conclusion

It has been a long timeframe of sickness, sadness, hopelessness, and distress, but the rollout of COVID-19 vaccination globally has given rise to feelings of relaxation for so many. The information about vaccination, side effects, and efficiency is ongoing and circulating on social platforms. This project utilized the power of NLP on Twitter data to understand the sentiments of people over-vaccination. Python package NLTK was used for tokenization/detokenization. LSTM, Bi-LSTM, and 1D CNN models were trained and evaluated for accuracy. The highest accuracy of 74.4% was achieved by Bi-LSTM and it was used for implementation on the COVID-19 vaccine tweets dataset. Trends of different sentiments were drawn using Tableau. We see a higher number of negative tweets as compared to positive tweets. Moderna, Sinovac, Sputnik, and Johnson & Johnson show an increasing trend in negative sentiment. Pfizer seems to have a constant level of negative sentiment. The USA has the highest number of vaccinations done as of May 1st. As the 2nd wave of the virus has hit India hard, it needs to ramp up its vaccination per hundred. For further analysis complete tweet text dataset can be utilized, instead of small sample.

**References**


1. *COVID-19 World Vaccination Progress.* (2021, April 17). Kaggle. https://www.kaggle.com/gpreda/covid-world-vaccination-progress
2. Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, Cameron Appel, Lucas Rodés-Guirao and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org. Retrieved from*: 'https://ourworldindata.org/coronavirus' [Online Resource] O. (2021).
3. Staff, A. (2021). *A Timeline of COVID-19 Vaccine Developments in 2021*. AJMC. https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021
4. https://www.ucsf.edu/news/2021/01/419691/covid-19-vaccine-fact-vs-fiction-expert-weighs-common-fears
5. *PanaceaLab - COVID19 Twitter Dataset Homepage.* (2021). Panacea Lab. http://www.panacealab.org/covid19/
6. *Tweet sentiment extraction.* (n.d.). https://www.kaggle.com/c/tweet-sentiment-extraction.
7. Garg, P., & Bassi, V. G. (2016). *Sentiment analysis of twitter data using NLTK in python* (Doctoral dissertation).
8. *Understanding LSTM Networks*. Understanding LSTM Networks -- colah's blog. (n.d.). http://colah.github.io/posts/2015-08-Understanding-LSTMs/.
9. Sosa, P. M. (2017). Twitter sentiment analysis using combined LSTM-CNN models. *Eprint Arxiv*, 1-9.
10. Adwan, O. Y., Al-Tawil, M., Huneiti, A. M., Shahin, R. A., Zayed, A. A. A., & Al-Dibsi, R. H. (2020). Twitter Sentiment Analysis Approaches: A Survey. *International Journal of Emerging Technologies in Learning*, *15*, 79–93. https://doi-org.ezproxy.bellevue.edu/10.3991/ijet.v15i15.14467
11. Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE, 2.

12. Dubey AD. Twitter Sentiment Analysis during COVID19 Outbreak. Available at SSRN 3572023. 2020.
13. Koohikamali M, Sidorova A. Information Re-Sharing on Social Network Sites in the Age of Fake News. Informing Science. 2017;20
14. Apuke, O.D.; Omar, B. Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users. *Telemat. Inform*. 2021, 56, 101475.