COVID-19 Vaccine Tweet Sentiment Analysis

Gourav Verma

Data Science, Bellevue University, NE, USA

DSC680-T302: Applied Data Science

Prof Fadi Alsaleem

05/02/2021

Abstract

In this project, sentiment analysis of tweets about COVID-19 vaccines is conducted. To achieve this latest dataset containing COVID-19 vaccine tweets across the world was used.[1] Natural Language Processing (NLP) is a branch of artificial intelligence that enables the machine to read, understand, communicate, and derive meanings from human's natural languages.[2] Natural languages can be in the form of text or speech. The Semantic (Text) analysis aspect of NLP enables the machine to understand the meaning conveyed by a Text. It involves applying computer algorithms to understand the meaning and interpretation of words and how sentences are structured. With the combination of NLP and machine learning algorithms, sentiment analysis can be done, to automatically identify the emotions attached to the communication. 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. To overcome this issue dataset with existing sentiment labels was used to train the models.[8] Tweet sentiments are classified into three types Positive, Negative, and Neutral. RNNs (Single LSTM and Bidirectional LSTM) and 1D CNN were applied to train the tweet training dataset. With Bidirectional LSTM, I achieved 78% accuracy, hence it was used to implement on our COVID-19 vaccine tweets dataset.

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

Table of Contents

Introduction

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. Being a highly popular social platform, Twitter provides a large scale of text data for various research such as sentiment analysis. It has many real-world applications. For word embedding, unsupervised and semi-supervised techniques are more popular, however, many sentiment analyses use handcrafted features. 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 data present in the Kaggle for COVID-19 vaccine tweets is getting updated daily. However, the text seems to be incomplete, as many tweets end with '…' proceeding, with a link for the tweet. This dataset contains tweets from all over the world. To train the model, tweets dataset with existing sentiment classifications, then after training the model, the best model was applied to the COVID-19 vaccine tweet dataset. It will help the model to understand the tweet language.

The year 2020 was full of COVID-19 spread across the world, multiple lockdowns, and burnout of healthcare 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.[5] As everyone wanted to go back to normal life, people were closely monitoring vaccine developments. Due to the overflowing COVID-19 news and speedy development of a vaccine, people rationally have many questions on vaccines. Some of them are:[6]

- 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, [7] every day, there are about 4 million tweets a day related to COVID-19. Hence, I planned to utilize tweets to understand people's sentiments. This analysis will help understand the difference in sentiments for different vaccines and the change in sentiments over time. 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.

Prior Studies

Sosa, 2017[11] 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.[12] In the sentiment analysis of COVID-19 tweets using supervised machine-learning authors Rustam Et al., 2020 [13] 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, I will be using tweets about the COVID-19 vaccines used in the entire world.[2] As per the Keggle data contributor, tweets are collected using the tweepy Python package to search Twitter API using relevant search terms. The dataset is continuously updated once a day, during morning hours (GMT). The tweets are about Pfizer/BioNTech, Sinopharm, Sinovac (both Chinese-produced vaccines), Moderna, Oxford/Astrazeneca, Covaxin, and Sputnik V vaccines.

The tweets dataset contains below columns –

- Id: Total 60.3k values
- User_name: 60.3k values ,32.6 unique values
- User_location: 13.8k missing values
- User_description: 4158 missing values
- User_created: user creation date.
- User_followers: follower count of user
- User_friends: friends count of user.
- User_favourites
- User_verified: True-6421, False-53.9k
- **Date**: Tweet date
- **Text**: Tweet text, 60.3k 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. 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, I will be comparing sentiments trend with COVID-19 vaccination progress.[3] This data is collected from https://ourworldindata.org/ GitHub repository.[4]

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. However, the text seems to be incomplete, as many tweets end with '…' proceeding, with a link for the tweet. It might be due to the way of data extraction from Twitter API. It might result in incorrect sentiment identification. Another dataset with complete tweet text, if available, can be preferred for the 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 training dataset was split into test and train.

Modelling

Models

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, I 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, I 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,[8] Bidirectional LSTM, and 1D CNN [9]. 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.[10] 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:1.

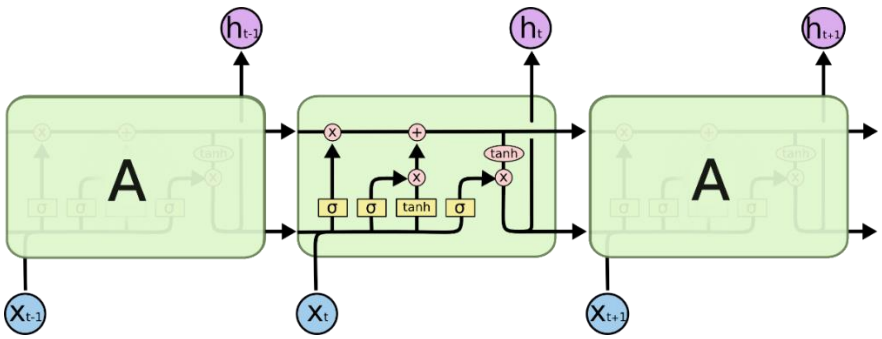


Figure 1: LSTM Network

For the LSTM classification model, I 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.

```
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 2 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:3 Bi-LSTM has two parallel layers to circulate in two directions.

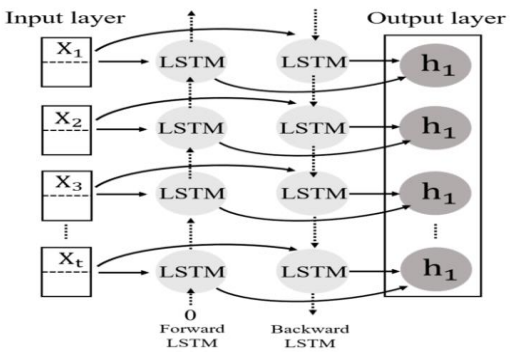


Figure 3: Structure of Bi-LSTM

For the Bi-LSTM classification model, I 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 72.7% accuracy.

```
645/645 [==============================] - 52s 76ms/step - loss: 0.8849 - accuracy: 0.5927 - val_loss: 0.6527 - val_accurac
y: 0.7274

Epoch 00001: val_accuracy improved from -inf to 0.72737, 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 4: 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.

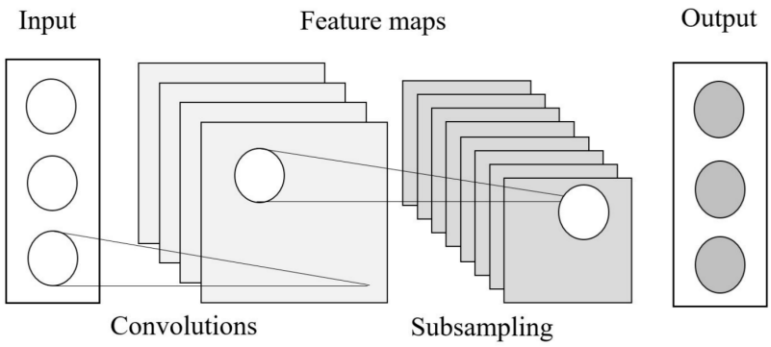


Figure 5: Structure of 1D CNN

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.

```
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 6: 1D CNN Model Outcome

Model Deployment

Among all the three models, Bidirectional LSTM shows the highest accuracy of 72.7%. 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.

Figure 7: 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.

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

```
Sentiments
Negative     7456
Neutral     60925
Positive     1337
Name: id, dtype: int64
```

Visualizations using Tableau.

Further, the outcome dataset was used to visualize the sentiment trend using tableau. Below are some observations. Overall, I see a higher number of negative tweet sentiments towards vaccines than positive tweets(Figure:8). 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, I was further interested in the sentiment trend for individual vaccines. Figure:9 shows the trend in the sentiments. Even though negative sentiment has a higher count, Moderna, Sinovac, Sputnik, and Johnson & Johnson show an increasing trend in negative sentiment. Pfizer seems to have a constant level of negative sentiment. Among all vaccines, people tweeted more about the Moderna vaccine.

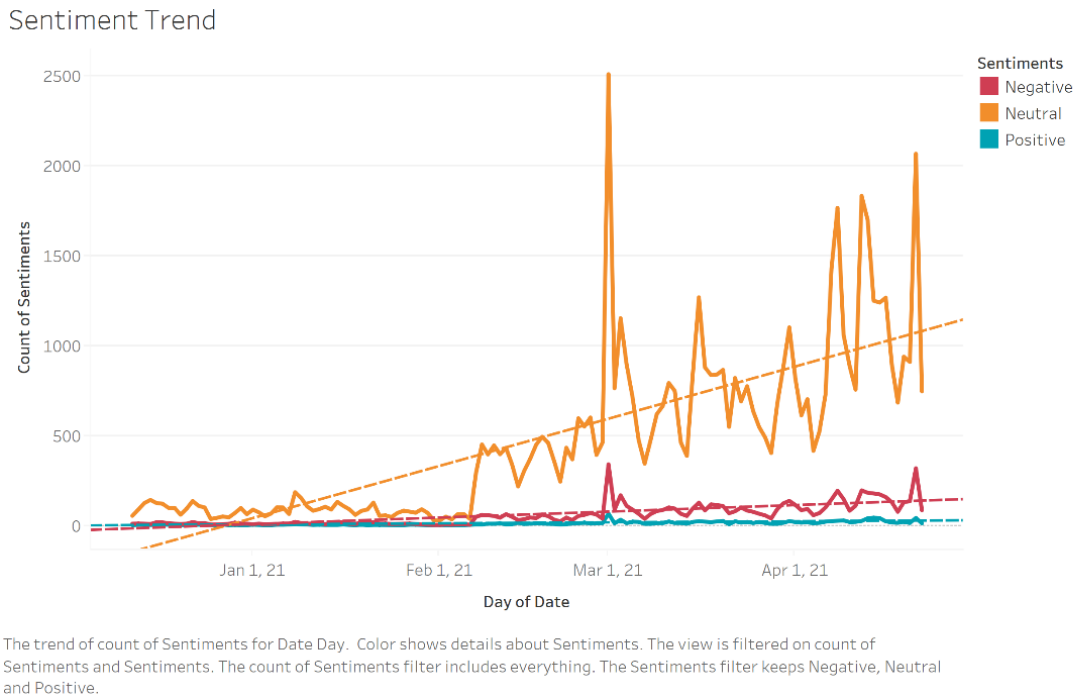
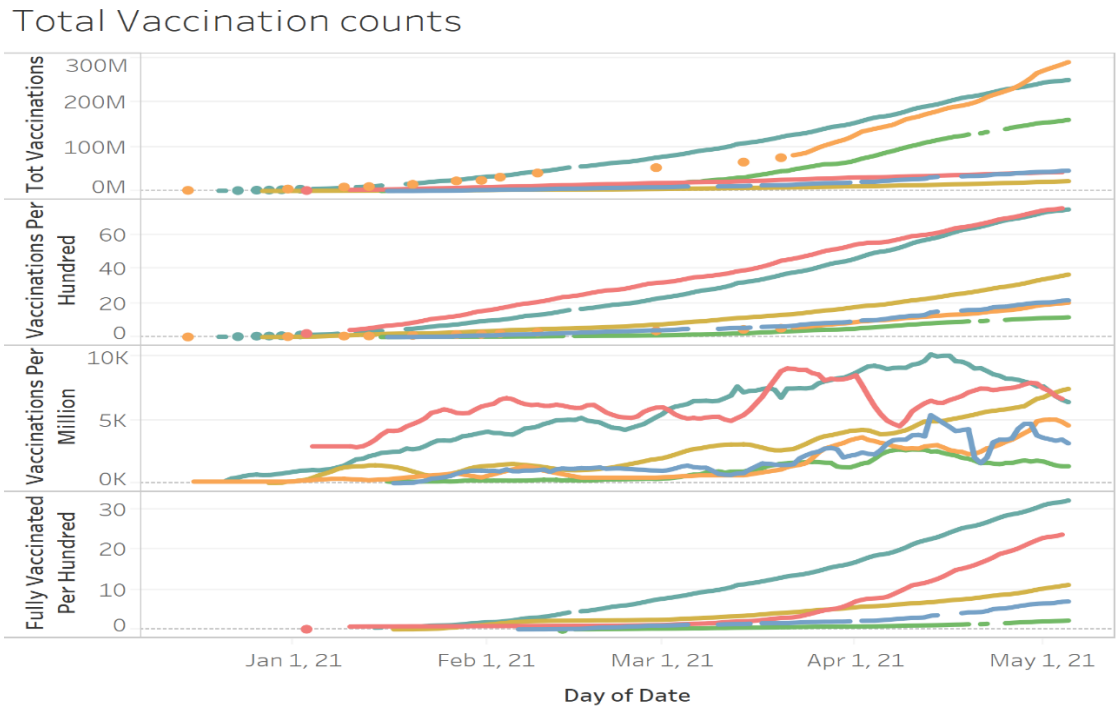


Figure 8: Sentiment Trend

Figure 9: Vaccine Sentiment Trend

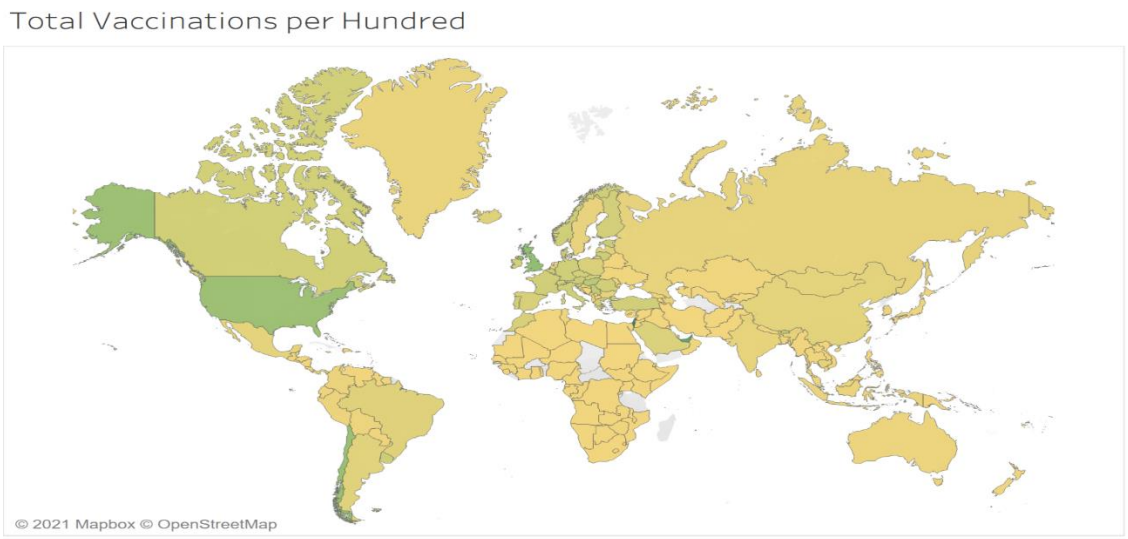
As of May-1st China has done highest number of vaccinations done than any other country. So far United States has a greater number of fully vaccinated population per hundred. More than 30% of US population are fully vaccinated and around 70% of people has received at least one dose of a vaccine. Even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



The trends of sum of Total Vaccinations, sum of Total Vaccinations Per Hundred, sum of Daily Vaccinations Per Million and sum of People Fully Vaccinated Per Hundred for Date Day. Color shows details about Country and Vaccines. The view is filtered on Country, which keeps 6 of 195 members.

Country, Vaccines
- Brazil, Oxford/AstraZeneca, Sinovac
- China, Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac
- England, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech
- India, Covaxin, Oxford/AstraZeneca
- Italy, Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech
- United States, Johnson&Johnson, Moderna, Pfizer/BioNTech

Figure 10: Vaccination count trend



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Vaccinations Per Hundred. Details are shown for Country. The view is filtered on sum of Total Vaccinations Per Hundred, which includes values greater than or equal to 0.

0 10,700

Figure 11: World map of Total Vaccinations per Hundred

Conclusion

It's 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. Highest accuracy of 72.7% was achieved by Bi-LSTM and it was used for implementation on COVID-19 vaccine tweets dataset. Trends of different sentiments were drawn using Tableau. I see 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. USA has highest number of vaccinations done as of May 1st. As the 2nd wave of virus has hit India hard, it needs to ramp up its vaccination per hundred. For further analysis dataset with complete tweet text, if available, can be utilized.

References

1. Wikipedia contributors. (2021, April 7). *Natural language processing*. Wikipedia. https://en.wikipedia.org/wiki/Natural_language_processing
2. *All COVID-19 Vaccines Tweets.* (2021, April 15). Kaggle. https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets
3. *COVID-19 World Vaccination Progress.* (2021, April 17). Kaggle. https://www.kaggle.com/gpreda/covid-world-vaccination-progress
4. O. (2021). *owid/covid-19-data*. GitHub. https://github.com/owid/covid-19-data
5. 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
6. https://www.ucsf.edu/news/2021/01/419691/covid-19-vaccine-fact-vs-fiction-expert-weighs-common-fears
7. *PanaceaLab - COVID19 Twitter Dataset Homepage.* (2021). Panacea Lab. http://www.panacealab.org/covid19/
8. *Tweet sentiment extraction.* (n.d.). https://www.kaggle.com/c/tweet-sentiment-extraction.
9. Garg, P., & Bassi, V. G. (2016). *Sentiment analysis of twitter data using NLTK in python* (Doctoral dissertation).
10. *Understanding LSTM Networks*. Understanding LSTM Networks -- colah's blog. (n.d.). http://colah.github.io/posts/2015-08-Understanding-LSTMs/.
11. Sosa, P. M. (2017). Twitter sentiment analysis using combined LSTM-CNN models. *Eprint Arxiv*, 1-9.
12. 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
13. 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.