Covid19 Vaccine Sentiment Analysis

AKSHAY B K
Computer Science Dept.
PES UNIVERSITY
Bengluru, India
SRN: PESUG19CS030

ANUP OMKAR
Computer Science Dept.
PES UNIVERSITY
Bengluru, India
SRN: PESUG19CS051

KEDARNATH K BHAT
Computer Science Dept.
PES UNIVERSITY
Bengluru, India
SRN: PESUG19CS180

NIKITHA MK
Computer Science Dept.
PES UNIVERSITY
Bengluru, India
SRN: PES2UG19CS261

***Abstract* — Web-based media, like Twitter, is a wellspring of trading data and assessment on worldwide issues, for example, COVID-19 pandemic. In this review, we work with an information base of around 8k tweets gathered across five weeks of December 2020 – April 2021 to reach inferences about open opinions towards the immunization viewpoint when inoculations become generally accessible to the populace during the COVID-19 pandemic. We convey regular language handling and opinion examination procedures to uncover bits of knowledge about COVID-19 immunization mindfulness among the general population. Our outcomes show that individuals have positive opinions towards taking COVID-19 antibodies rather than some unfavorable impacts of a portion of the antibodies. We additionally investigate individuals' perspectives towards the security proportions of COVID-19 in the wake of getting the immunizations. Once more, the positive feeling is higher than that of negative as far as keeping up with security measures against COVID-19 among the inoculated populace. This review will help to comprehend public response and help the policymakers to project the inoculation crusade also as wellbeing and security measures in the continuous worldwide wellbeing emergency.**

Keywords — Tweet; Sentiment; Covid; Natural Language processing; Time series; EDA; Visualization, Random Forest, Logistic Regression, Naïve Bayes, Support Vector Classifier

I. INTRODUCTION

AI (ML) is the latest technique in information science that has cleared the way for innovative achievements and instruments that would have been inconceivable as a few years prior. Image recognition, sentiment Analysis [1–4], item suggestions, spam/extortion discovery [5], online media highlights, and so forth are a portion of this present reality machine learning applications that are clearing the world. Distinctive online based web-based media has been comprehensively used as a method for exchanging information by both the populace and associations from one side of the planet to the other. The amount of online media clients has begun to augment rapidly, especially in the earlier decade. Facebook, Twitter, YouTube, LinkedIn, and Pinterest saw gigantic augmentations over the earlier year. Facebook is the most well-known web-based media with 2.8 billion month to month active clients [6], while Twitter has around 300 million month to month active clients [7]. Twitter is experiencing quick turn of events and is quickly obtaining notoriety wherever on the planet. The Twitter interface is used by specific clients to help unique perspectives, for example as a medium to battle, political missions, and data spreading, furthermore it is accepting a critical part in friendly turn of events.

Covid is one of the moving topics on Twitter since December 2020 and has kept on being analysed to date. A bunch of pneumonia cases in Wuhan, China, was accounted for to the World Health Organization (WHO) on 31 December 2019 and the reason for the pneumonia cases (the infection named as COVID-19) was distinguished as an original beta Covid, the 2019 novel Covid (2019-nCoV, renamed as SARS-CoV-2) [8]. In March 2020, COVID-19 was proclaimed as a pandemic by WHO thinking about in excess of 118,000 cases in 114 nations [9]. By 1 June 2021, there has been 3.57 million affirmed passings and 171.19 million affirmed. Coronavirus cases [10]. The circumstance has improved since the immunization of COVID-19 began to increase. As we acquire proof of the positive effects of inoculation on transmission, it will assist with reinforcing public trust [11]. Thinking about this, examining the general assessment or then again feeling is vital for spurring individuals to be immunized against COVID-19.

This paper targets investigating public opinion on COVID-19 immunization. Investigating the Twitter content enables wellbeing specialists, policymakers to find out with regards to the public's response to inoculation during the Coronavirus pandemic. It additionally clarifies individuals' perspectives on the wellbeing rules for the counteraction of COVID-19 subsequent to getting inoculated. Disclosures from this investigation related to wellbeing are valuable as principal assessments for building more careful models, which can be used to make proposition for the bigger public and set up significant procedures and arrangements. The tweets in this review have gotten conversations about inoculation what's more wellbeing rules during COVID-19 in various countries. Web-based media data grants researchers and scientists to have a worldwide perspective, which is especially quick during an overall pandemic. This review can be repeated by scratching tweets consistently until the COVID-19 pandemic reaches a conclusion for understanding the by and large public opinion while the immunization crusade is continuous. This type of study will be useful for the health and government officials to get insights about any newly discovered disease with early invention of vaccine for that particular disease.

II. PREVIOUS WORK

An analysis of Previous works and predecessors is detailed below:

1. *"Sentiment Analysis of COVID-19 Vaccination from Survey Responses in Bangladesh"* [12]

This paper essentially manages examination of feelings from study reactions. Anyway, procedures and results utilized in this specific paper filled in as essential motivation for our work.

This paper manages the reactions gathered by the client physically either through discussions or the face to face. The principal procedures utilized here depend on the tremendous sea that is AI. AI is regularly a monster pool of different regulated and unaided procedures of forecast. In particular, this paper manages numerous models subsequent to resampling and scaling back information through include extraction. A random forest classifier is utilized to order the information gathered. After this cross-validation model was worked to investigate the examples and give a precision of expectation of feelings.

The outcome was a respectable level of accuracy - 75% on a test set. At long last, the finish of this paper and model is that it's verification of how genuinely amazing AI is utilized and how it tends to be utilized cleverly to address various issues. The opinion of individuals towards vaccination being a huge matter can be investigated and individuals' psyche can be changed through meetings.

2. *"Using Twitter for sentiment analysis towards Pfizer COVID-19 vaccine"* [13]

Unlike previous paper this paper deals with the tweets collected for sentiment analysis. The Twitter academic Application Programming Interface was used to retrieve all English-language tweets mentioning Pfizer vaccines in 4months from 1 December 2020 to 31 March 2021. Sentiment analysis was performed using the AFINN lexicon to calculate the daily average sentiment of tweets which was evaluated longitudinally and comparatively for each vaccine throughout the 4months.

The Analysis techniques used in this paper, helped us to analyse the tweets regarding all vaccines and helped us to get a better visualisation of the sentiments of people on vaccine.

The sentiment regarding Pfizer vaccines appeared positive and stable throughout the 4 months, with no significant differences in sentiment between the months. This paper also analyses the impact of Pfizer vaccine in different parts of the world. The Conclusion of this paper is, Lexicon-based Twitter sentiment analysis is a valuable and easily implemented tool to track the sentiment regarding COVID-19 vaccines. High vaccine uptake is paramount for ending the pandemic, while identification of events that impact the sentiment around vaccines also allows for better planning and implementation of specific interventions.

Limitations with predecessor work and our own Modifications

The first paper referenced above centres around the model structure just and the survey is done physically which emerges the bias of doing the overview just in a specific region or city. Performing a survey is a tedious task, and individuals who are not able to answer a study can give false information and henceforth mislead the structure of the model. The subsequent second paper settles the issue of the first, by analysing the Twitter tweets, on the grounds that individuals who are tweeting tweet really what they feel, however the subsequent paper doesn't focus on model building to classify the tweets. The emphasis is simply on the analysis itself.

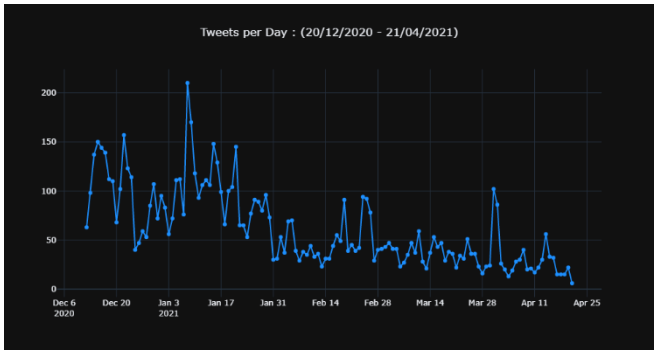
Now we aim to take the insights obtained from each paper and resolve the limitations occurred from aforementioned papers by analysing the tweets (resolving the limitation occurred from paper1) and building the model to classify the sentiments as positive, negative and neutral (resolves the limitation of paper2).

III. PROPOSED SOLUTION

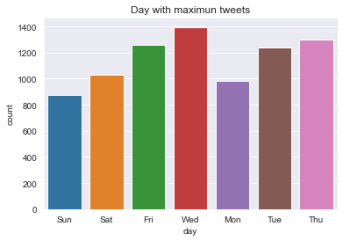
We collected the pre extracted tweet dataset from Kaggle. As mentioned above our main focus is on analyzing the tweets and build the various models using sklearn library to classify the tweet sentiments and conclude which model best classifies the model.

A. EDA and Visualization

- The dataset consists of following attributes:
  - user_name, user_location, user_description, user_followers, user_friends, user_favourites, user_verified, date, text, hashtags, source, retweets, favorites, is_retweet
- The dataset contains 8082 rows and 16 attributes. But extra attributes for label is added while performing exploratory analysis.
- user_location, user_description has most missing values. There are around 4k missing values in 129.3k values (whole dataset) , missing values will be handled during visualisation, since there is no much effect of missing values on EDA and model building process.



- The histogram and line chart shows the number of tweets by the date from Dec – 20 -2020 to Apr – 11 – 2021. We can see that the number of tweets were maximum when the covid vaccine was launched (i.e At the end of 2020 and beginning of 2021)



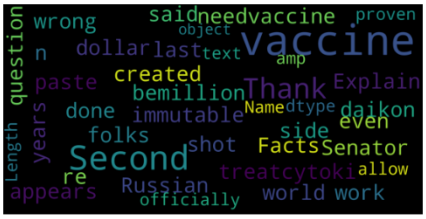
- The above bargraph shows that most of the tweets related to vaccine were done on Wednesday.

B. Text Preprocessing

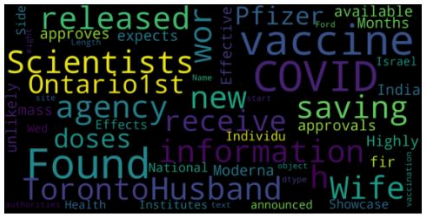
Text Preprocessing is traditionally an important step for Natural Language Processing (NLP) tasks. It transforms text into a more digestible form so that machine learning algorithms can perform better.

The Preprocessing steps taken are:

Conert to lowercase; Removing Twitter handles; Remove Twitter Hashtags; Remove URL; Removing Non-Alphabets; Removing Short Words; Removing Consecutive letters; Removing Multiple Spaces.



Prevalent words in tweets

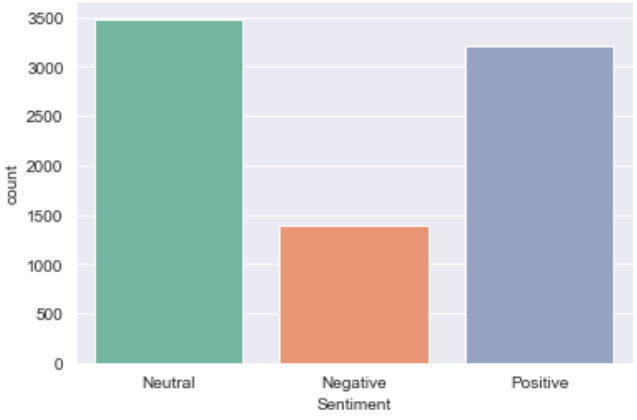


Prevalent words in tweets from India

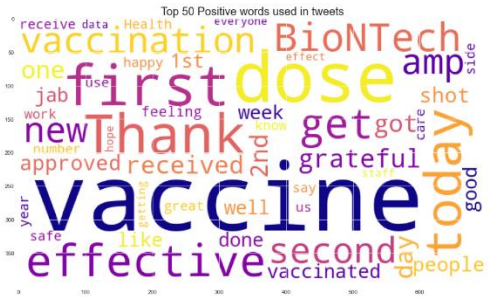
- The word cloud shows the prevalent words in the tweets across the globe and across India.

C. Apply VADER Sentiment to the tweets to get labels

- The VADER Sentimental analysis module is used to label the example / datapoint as the positive, negative or neutral. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up each word's intensity in the text.
- After labelling the sentiments, the following bar chart and funnel chart are plotted to count how many tweets are positive, negative and neutral



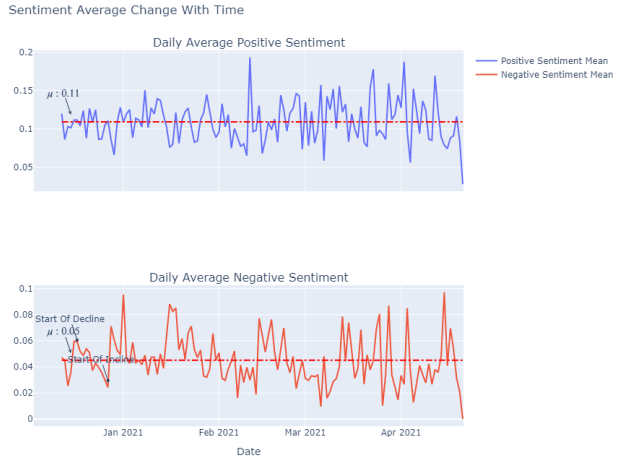
- It can be seen that around 3205 tweets are positive, 3483 are neutral and 1394 are negative



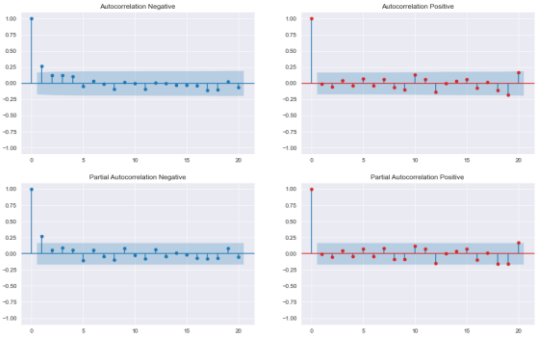
- The above Word Cloud shows the most common words in Positive sentimental tweets.

- The above Word Cloud shows the most common words in Negative sentimental tweets.

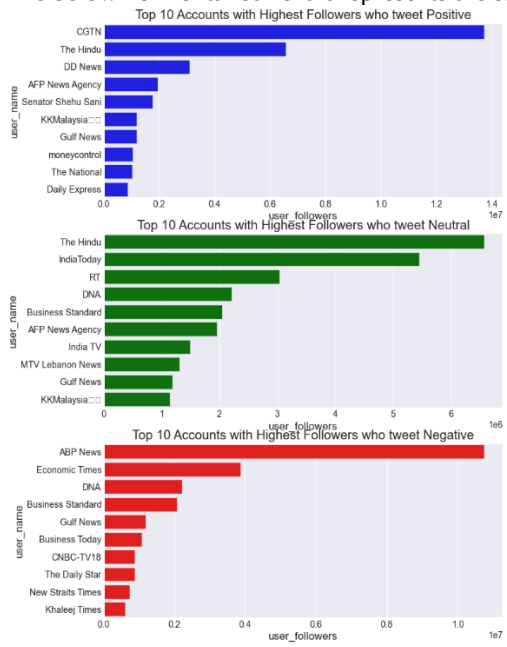
D. Time series Sentiment Analysis



- The plot shows the daily average change in Positive and negative sentiments over time.
- Here we can see that the there is no trend or cycles or seasonality observed with Time in Positive and Negative Sentiments of tweets. So we can conclude that **time series analysis on Sentiments is pointless**



- Here, from graphs we can observe that the acf and pacf values for positive and negative sentiments are nearly zero and there is no exponential decrease in acf and pacf plots. Hence, the p and q values are 0. Hence using time series forecasting models like ARMA or ARIMA doesn't make any sense.
- As, it's obvious that verified users or media channels make a significant impact on people.
- ABP news, Economic Times, Business Standard which are most famous media channels tweeted negatively about the vaccine. while, the news channels Hindu, DD news, CGTN tweeted positively.
- The below horizontal bar chart represents the same.



E. Stop Words Removal, Lemmatization and Feature Extraction.

- *Removing Stopwords*: Stopwords are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. (eg: "the", "he", "have")
- *Lemmatizing*: Lemmatization is the process of converting a word to its base form. (e.g: "Great" to "Good")
- *TF-IDF* indicates what the importance of the word is in order to understand the document or dataset. TF-IDF Vectoriser converts a collection of raw documents to a matrix of TF-IDF features. The Vectoriser is usually trained on only the train dataset.
    - *ngram_range* is the range of number of words in a sequence. (e.g "very expensive" is a 2-gram that is considered as an extra feature separately from "very" and "expensive" when you have a n-gram range of (1,2))
    - *max_features* specify the number of features to consider. (Ordered by feature frequency across the corpus)
- So, we convert out tweet to the vector form which consists of numbers (that's what all machine learning models deal with).

F. Model Building and Evaluation

- We have used 4 different models for predicting the sentiment as negative, neutral or positive encoded as 0, 1 & 2 respectively.

➢ The first model chosen is *Logistic Regression.* ***Logistic regression***, by default, is limited to two-class classification problems. Some extensions like one-vs-rest can allow logistic regression to be used for multi-class classification problems, although they require that the classification problem first be transformed into multiple binary classification problems. Instead, the multinomial logistic regression algorithm is an extension to the logistic regression model that involves changing the loss function to cross-entropy loss and predict probability distribution to a multinomial probability distribution to natively support multi-class classification problems.

➢ The second model chosen is ***Multinominal Naïve Bayes***. In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features (see Bayes classifier). They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.

➢ The third model chosen is ***Linear Support Vector Classifier,*** it is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVC algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

➢ The fourth model chosen is ***Random Forest Classifier***, A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

Evaluation metrics:

Logistic Regression

Sentiment	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Negative	*0.93*	*0.51*	*0.66*	80%
Neutral	*0.75*	*0.90*	*0.82*	
Positive	*0.84*	*0.82*	*0.83*	

Naïve Bayes

Sentiment	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Negative	*0.91*	*0.14*	*0.24*	72%
Neutral	*0.74*	*0.84*	*0.79*	
Positive	*0.69*	*0.85*	*0.76*	

Linear Support Vector Classifier

Sentiment	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Negative	*0.86*	*0.61*	*0.72*	83%
Neutral	*0.81*	*0.92*	*0.86*	
Positive	*0.86*	*0.85*	*0.85*	

Random Forest Classifier

Sentiment	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Negative	*0.92*	*0.47*	*0.62*	80%
Neutral	*0.72*	*0.98*	*0.83*	
Positive	*0.90*	*0.74*	*0.81*	

G. Experiment Results, insights and shortcomings

We used a range of models to predict the sentiment regarding tweets on covid-19 vaccine. The best accuracy was given by Logistic regression, Linear SVC and Random Forest Classifier. Since ***the data is skewed*** i.e. the number of labels are not equal in dataset, the ***F1 score may be the best measure possible*** for this type of data. Looking at the metrics, the overall F1 score is better for Support vector Classifier then comes Logistic Regression.

The Naïve Bayes performs poorly compared to other models. Since the words are converted to the vector using TF-IDF vectorizer, the vectors have real values and only ensemble models (here Random Forest classifier), Support vector classifier and Regression Models (Logistic Regression) performs well which is as expected.

The Naive Bayes Model works well when the vector is in integer format, because the Naïve Bayes only deals with probability which requires frequency of words and can be obtained through CountVectorizer.

The tweets which are sarcastic, irony, negation, exaggeration in nature can't be handled by our model which is one of the shortcomings of this paper.

Many People tweet in multilingual format, while pre-processing most of the important matter which are other than English language will be removed. These research on these things can be carried forward.

The methodology we used to classify sentiments may have missed a few posts as we didn't audit the whole corpus to track down phrases. A comprehensive audit of the corpus physically would have not been imaginable as far as work and time given.

Despite the effort we made to standardize the geographical information of users, the user-defined profile locations cannot necessarily represent the actual locations that tweets were posted from.

IV. CONCLUSIONS

In the present study, sentiment and opinion analyses of approximately 8K tweets concerning COVID-19 vaccines. The Twitter platform that was used in the present study may be a valuable tool for public health promotion to reinforce vaccine acceptance and decrease vaccine hesitancy and opposition.

Overall, understanding sentiments and opinions toward vaccination can help public health authorities reinforce positive language and comments within the positive posts while dispelling combative language promoting misinformation within negative posts.

Also, general wellbeing organizations might have the option to deal with Twitter and different news sources to expand positive informing, lessen negative and contradicting messages and favorable to effectively suspend hostile to inoculation records, for example, bots to energize and upgrade the take-up of a vaccination.

The results here revealed that the patterns of defined sentiments and opinions have changed in response to vaccine-related events during the pandemic. In general, the positive sentiment about the COVID-19 vaccine was the dominant polarity on Twitter. The main topics in positive tweets included hope, support and faith while negative tweets were usually related to fear, discouragement, anger and politics.

High vaccine uptake is paramount for ending the pandemic, while identification of events that impact the sentiment around vaccines also allows for better planning and implementation of specific interventions. Finally, it is worrisome that the sentiment regarding the vaccine appears to be decreasing in positivity over time. In March 2021, it was on average negative, and if this trend continues, it may boost hesitancy rates towards this specific COVID-19 vaccine.

REFERENCES

[1] Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 2014, 5, 1093–1113.

[2] Patel, R.; Passi, K. Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. IoT 2020, 1, 218–239.

[3] Dandannavar, P.; Mangalwede, S.; Deshpande, S. Emoticons and their effects on sentiment analysis of Twitter data. In EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing; Springer: Berlin/Heidelberg, Germany, 2020; pp. 191–201.

[4] Naseem, U.; Razzak, I.; Khushi, M.; Eklund, P.W.; Kim, J. COVIDSenti: A Large-Scale Benchmark

Twitter Data Set for COVID-19 Sentiment Analysis. IEEE Trans. Comput. Soc. Syst. 2021.

[5] Sattar, N.S.; Arifuzzaman, S.; Zibran, M.F.; Sakib, M.M. Detecting web spam in webgraphs with predictive model analysis. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 4299–4308.

[6] Facebook Reports Fourth Quarter and Full Year 2020 Results. Available online: https://investor.fb.com/investor-news/pressrelease-details/2021/Facebook-Reports-Fourth-Quarter\-and-Full-Year-2020-Results/default.aspx (accessed on 1 June 2021). [CrossRef] [PubMed] Appl. Sci. 2021, 11, 6128 30 of 32.

[7] Twitter Revenue and Usage Statistics (2021).

[8] Boldog, P.; Tekeli, T.; Vizi, Z.; Dénes, A.; Bartha, F.A.; Röst, G. Risk assessment of novel coronavirus COVID-19 outbreaks outside China. J. Clin. Med. 2020, 9, 571.

[9] WHO Director-General's Opening Remarks at the Media Briefing on COVID-19, 11 March 2020

[10] Roser, M.; Ritchie, H.; Ortiz-Ospina, E.; Hasell, J. Coronavirus pandemic (COVID-19). In Our World in Data; 2020.

[11] Mathieu, E.; Ritchie, H.; Ortiz-Ospina, E.; Roser, M.; Hasell, J.; Appel, C.; Giattino, C.; Rodés-Guirao, L. A global database of COVID-19 vaccinations. Nat. Hum. Behav. 2021, 1–7.

[12] Sentiment Analysis of COVID-19 Vaccination from Survey Responses in Bangladesh.

[13] Using Twitter for sentiment analysis towards Pfizer COVID-19 vaccine