INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 5 - V8I5-1243)
Available online at: https://www.ijariit.com

Sentiment analysis on COVID-19 Vaccine

Hiren Thakur
thakur12hiren@gmail.com
SRM Institute of Technology, Chennai, Tamil Nadu
Divij Singh Chauhan
divijsinghchauhan@gmail.com
SRM Institute of Technology, Chennai, Tamil Nadu
Aditya Dutta
aditya.dutta213@gmail.com
SRM Institute of Technology, Chennai, Tamil Nadu
Kaustubh Sharma
sharmakaustubh81@gmail.com
SRM Institute of Technology, Chennai, Tamil Nadu

ABSTRACT

In the current scenario of the COVID-19 pandemic, ensuring vaccination is a top priority. Our project aims to clarify people's attitudes towards vaccination against COVID-19. Various studies have already been conducted to measure mood associated withCOVID-19 vaccines, but they all suffer from a major common flaw: poor mood classification accuracy. Our project aims to increase the accuracy of mood classification for COVID-19 vaccines compared to previous studies. In our project, we used ABSA (Aspect-Based Sentiment Analysis) and TF-IDF (Term Frequency-Inverse Document Frequency) models for sentiment classification. We also tested the classification accuracy using five traditional machine learning models: Random Forest, Naive Bayes, Support Vector Machines, Logistic Regression, and Ensemble Classification. For our project, we classify sentiment into three categories called positive, negative, and neutral. This in many ways distinguishes our project from literature. First, according to the limited research we read, ABSA and TF-IDF were not used together. Second, most of the previous studies used the bag-of-words approach, an outdated model compared to ABSA. Finally, traditional machine learning models such as random forests and ensemble classification have never been used in ABSA and TF-IDF. This project provides higher accuracy than stated in our results as a mood classifier for COVID-19 vaccines, whereas previous studies have less than 67% accuracy.

Keywords: *TF-IDF, ABSA, LSTM, BiLSTM, VADER, SVM, RNN*

1. INTRODUCTION

THE COVID-19 outbreak has brought a lot of attention to modern healthcare and changed the way we think about protection in every element of our lives. At this time, safety measures such as sports masks, regular arm washing, and precautions regarding intimacy are very important. It's getting attention. In a series of tweets about vaccines, he notes that the approach taken by Eval Becker and others has shown good results [1]. We used Twitter tweets. Unlike other social media systems, Twitterspecializes in key phrases, allowing people to post to a wider audience. Additionally, the API allows users to access Tweets from a geographic location by specifying the latitude, longitude, and radius from which they would like to receive Tweets. Tweets in this paper were categorized as positive, negative, and neutral. Using this dataset, we arrived at observation after comparative observation results obtained with different algorithms.

In this study, our contribution could be doubled. First, we described a system that demonstrates best practices and model for classifying the sentiment of tweets. A side-by-side comparison was then based primarily on the overall performance of his ABSA and TF-IDF models. As vaccination sentiment is trending in the news around the world, we're following it.Due to differences in urgency and financial constraints, there may be some differences between regions. But for the most part I tried to present the actual facts based on the rough sentiments of unbiased people about vaccination. This suggests that large segments of the population on various continents are still unvaccinated. , poses the greatest difficulty for scientists and medical professionals who need to investigatethe motivations behind it. Our research also covers the timelineof such tweets. This is an important and novel contribution, as the emotions expressed change over time. The purpose and contribution of this study is to provide a concept that cleanly and correctly reflects people's feelings and thoughts about the vaccination system as opposed to COVID-19. This will help scientists, health experts and policy makers make the rightdecisions to give credibility to vaccines and empower people to do more regarding health awareness.

2. LITERATURE REVIEW

There have been several previous works that have ana- lyzed COVID-19 vaccine sentiments. These works havevarious things in common. One of the most common things among the various works is the use of the Twitter dataset [2][3][4][5]. This is something that is seen in all the articles and papers we have read so far. Detailed analysis of this observation suggests that Twitter is the best place for any kind of sentiment analysis. This is due

to Twitter's focus onkeywords and trends. Twitter's API is also noted to be goodfor extracting tweets from specific locations by specifying longitude and latitude [3]. In the work of Nwafor et al.[2], various machine learning models such as Naive Bayes, Logistic regression, random forest, Support Vector Machinesare used for baseline analysis. These models, despite their simplicity, have been proved to work well with textual data.In this work VADER which is Valence Aware Dictionary and Sentiment Reasoner is used for actual classification of sentiments of various tweets. In this work, the classification of the sentiments is done in multi-class classification. The various classes that sentiments are organized are positive, negative and neutral. A transformer model called COVID Bert v2 that produces results with high accuracy. In Mu- dassir et al. [3] work, the sentiment analysis of the tweets is done using Textblob,VADER and ABSA . While theTextblob model is used, it is known for its inaccuracies.

This was evidenced by the fact that Textblob has very low accuracy compared to other models presented in the paper. VADER was shown as having decent results. ABSA performed better than others, but it has some issues such as being computationally intensive. In this paper ABSAperformed far better than other models such as TextBlob and VADER with its accuracy being near 83% while the TextBlob and VADER have lower accuracies of around 61%and 67% respectively. In the works of Sattar et al. [4], the sentiment analysis of the tweets is again done using TextBlob and VADER just like in the previous work of Mudassir et al [3]. In this work, the number of the datasets used is two twitter data sets called Twitter Dataset 1 and Twitter Dataset 2. The first dataset is used for analysis of vaccine sentiments while the second dataset is used for analysis of sentiments regarding the precautionary measuresto COVID-19 pandemic.

In Alam et al. [5] work, the sentimental analysis is again done using VADER. In this work, Data characters were detokenized to divide the sentences into words and label them after they were checked for unique values and null values and pre-processing was completed. Then, using the VADER, a sentiment column with positive, negative, and neutral values was added and calculated. The performance of the forecasting model was then tested using deep learn- ing architectures, long short-term memory (LSTM), and bidirectional LSTM (Bi-LSTM).The RNN structure which is inclusive of lengthy short-time period memory (LSTM)

Paper	Model	Manual Labelling
Becker et al [1]	Manual Analysis	Partial
Xiong et al [10]	Multi-layer Perceptron and Convolutional Neural Network	None
D Li and J. Qian [11]	Long Short-Term Memory (LSTM) network	Partial
Raghupathi et al [12]	VADER	None
Jang et al [13]	Topic modelling and ABSA	None
Zainuddin et al [14]	Support Vector Machine	None
Kunal et al [15]	Naïve Bayes	None
Rezwanul et al [16]	K Nearest Neighbours and Support Vector Machine	None
D'Andrea et al [17]	Simple Logistic Classifier	Full
Salathé M, Khandelwal S [18]	Naïve Bayes, Maximum Entropy, Dynamic Language Model Classifier	Partial

FIGURE 1. Previous literature review and methods used for the CovidSentiment Analysis

and bidirectional LSTM (Bi-LSTM) is used to evaluate the overall performance of the predictive models, with LSTM accomplishing an accuracy which is much high than other models.

In Cotfas et al. [6] work, four approaches to text rep- resentation and classification have been examined. In thefirst approach, bag-of-words is used for text representationthat is then followed by classical machine learning model. In the second approach, Word embeddings are used first,followed by traditional machine learning. In the third ap-proach, Deep learning is followed by word embeddings. In the fourth approach, Transformers' Bidirectional EncoderRepresentations is used. The text was represented using bothBag-of Words and word embeddings techniques in order to select the best performing classification algorithm. Theaccuracy of all the methods described is less than 80%. Inthe work of Ghasiya et al. [7], There are two componentsto this work. In the first part, the top2vec model is usedto identify and analyse the most representative themes indifferent datasets. The second part is sentiment analysis.This part could be broken into two sections. Unsupervisedmachine learning algorithms are used to create a labelleddataset in one section. After that, RoBERTa is used to trainand test the labelled dataset. The Roberta model produces 80% accuracy in the classification. In the work of Lyu et al. [8], the tweets were cleaned with R software andthe tweets which had the phrases vaccination, vaccinations, vaccine, vaccines, immunization, vaccinate, and vaccinatedwere kept. The total number of tweets considered in the analysis was 1,499,421. Latent Dirichlet allocation for topicmodelling as well as sentiment and emotion analysis wasdone using R. The work separates tweets into two batches.One for the text mining and one for the sentiment analysis.In Yousefinaghani et al. [9] work, the historical tweets about the COVID-19 vaccination were collected using snscrape. Because snscrape does not supply any geograph- ical information, a user lookup Twitter API is to acquireusers' location and interaction information. This is done to acquire current tweets in this work. This work used Valence Aware Dictionary and Sentiment Reasoner (VADER) toassign each tweet a polarity of 'positive,' 'negative,' or 'neu- tral'. Sentiments were allocated based on VADER values. Any tweet with a value of 0.25 or higher was classified as positive sentiment, any tweet with a value of 0.25 or lower was classified as negative sentiment, and any value in the middle was classified as neutral sentiment. This work does not state the accuracy of the classification explicitly.

One of the biggest issues in the current works is the use of VADER model. The reason for that is its inability to classifythe tweet's overall sentiments. The second major concern is VADER's lack of accuracy when compared to other good models with higher accuracy. The use of TextBlob is next biggest issue. It uses Bag of words approach that has low precision. In our work we try to fix these issues and try to classify the COVID-19 sentiments as best as possible.

I. 3. METHODOLOGY

The purpose of this study is to categorize tweets and capture Twitter users' sentiments towards vaccines. The steps of our approach are shown in the figure 2. Describe the collection of records and tweets. Further extraction, cleaning and preprocessing of the dataset were performed. State classifications then follow three categories: for, neutral, and against. The dataset was collected from Kaggle. It contains various types of tweets pre-labeled with 0 and 4 to provide positive and negative explanations. Furthermore, the systemis divided into three main components:

A. THE COLLECTION AND PRE-PROCESSING OF TWEETS

The data set named "Sentiment140" is an open source dataset containing 1.6 million tweets about products, brand and topics [19]. It contains 6 fields as target, ids, date, flag, user, text. Then for the data pre-processing we use Stemming and Lemmatization: -

Stemming: Stemming is a process of removing inflected words to their stem words which is affixes, for example studies-study. Stemming works on some languages mainly English and Spanish.

Lemmatization: Lemmatization takes the attention of morphological eval- uation of the phrases. It reduces inflected phrases well with the foundation phrases belongs to the sentences. The disadvantage of a primarily dictionary based approach is thatit cannot cope with area and context precise orientations, that is important for gauging the emotions concerning the vaccination drive.

B. MODELS FOR TEXT ANALYSIS

The sentiment analyzer categories the given text as sentimentand will give us the result based on positive, negative and neutral. Because of the better result we basically used two models for the sentiment analysis as: -

Aspect Based Sentiment Analysis: Aspect-Based Sentiment Analysis (ABSA) is a type of text analysis that categorizes opinions by aspect and identifies the sentiment related to each aspect. E.g. "I was about to buy this product because of its design, but its price is not very good", so here we see two terms as Design: positive and Price: Negative. The main motive to use ABSA is the accuracy as it has 81 percent of accuracy [3].

TF-IDF: TF-IDF - Term Frequency Inverse Document Frequency may be described as the calculation of the way a phrase in a sequence or corpus is applicable to a textual content. It has a numerical statistic that is intended to reflect how important aword is and will increase proportionally to the wide variety of instances where that word is used. We used it as it is easyto calculate and computationally reasonably-priced and is a easy place to begin for similarity calculations. It even offers much higher accuracy than BERT, Bag of Words [20].

C. PERFORMING EVALUATION PROCESS

A support vector machine (SVM), Naive Bayes algorithm, Random forest, Logistic Regression and Ensemble Classi- fication algorithm were used to evaluate the performances and obtain results with the proper labeled predictions.

1) Support Vector Machine (SVM) for regression
The ability of SVM to solve nonlinear regression estimation problems makes SVM successful. SVM regression acknowl- edges the presence of non-linearity in the data and provides a proficient prediction model [4]

2) **Naive Bayes:** Mainly used in text classification which is a probabilistic classifier, which means it predicts based on the probability of an object. Though they are very common so people have stopped using it. But for the testing of best results we are going to use it.

3) **Random Forest:** RF regression includes numerous selection bushes and goalsthat show this is the mode of the classes' goal via way of means of person bushes. The wide variety of bushes to be grown with inside the woodland and the amount of functionsor variables selected at each node to broaden a tree are the 2 parameters [21].

4) **Logistic Regression:** Logistic Regression is the appropriate regression analysis toconduct when the dependent variable has a binary solution.

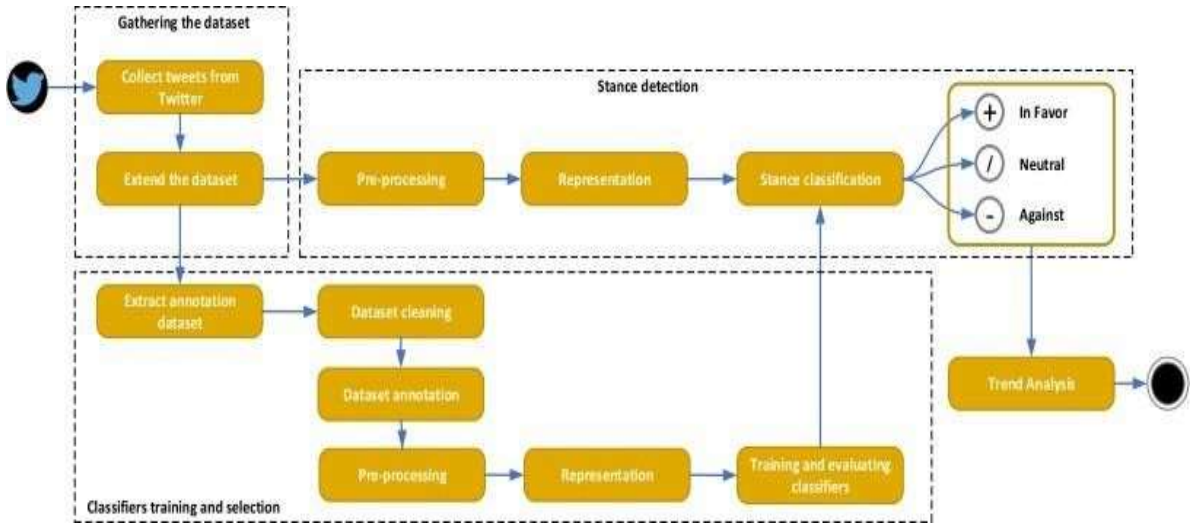


FIGURE 2. Steps of our approach.

	id	tweet	place	place_coord_boundaries	location
0	1.379175e+18	Why are hashtags always spelt incorrectly?	en	NaN	London
1	1.379174e+18	health regulator may restrict shot for younger...	en	NaN	Barquisimeto-Venezuela
2	1.379172e+18	VIDEO The effects of the four vaccines current...	en	NaN	Somalia
3	1.379171e+18	s vaccine roll out is the biggest joke in the ...	en	NaN	Alberta, canada
4	1.379169e+18	My point was that if Germany gets that many ja...	en	NaN	Birmingham

FIGURE 3. Extraction

Logistic Regression is also a type of predictive regression system mostly used to evaluate the relationship betweenone

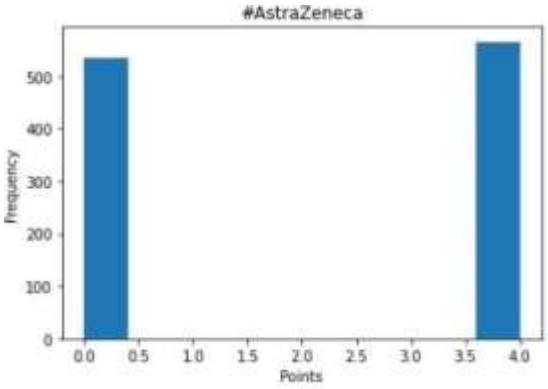
dependent binary variable and one or more independent variables [4].

Ensemble Classification: Ensemble methods are techniques that create multiple mod- els and then combine them to produce improved results. En-semble methods usually produces more accurate solutions so that is the reason we are considering Ensemble classificationto get more accurate result.

5. RESULTS

Our research focus is to predict better accuracy and outcome for analyzing the mood of the COVID-19 vaccine. This work will help health researchers to fully understand the challenges associated with vaccination. For the results, we compared the algorithms and first extracted the data, as shown in the figure. 3. ID, Tweet, City, and Location are considered for extraction. After cleaning up the text with lemmatization and stop words, import the matrix.

The random forest approach gives 72% for TF-IDF and 82%for ABSA. We can then confirm this with Naive Bayes as we get 73.7% with the TF-IDF approach, 73.8% with BOG, and 83% with TF-IDF and ABSA combined. special results. Figure 4. AstraZeneca results. Provided in source code files. SVM and Ensemble classes had similar scores, 74% on TF-



IDF and 83% on ABSA. It then retrieves tweets based on users and their locations to get vaccines for people to use. Use keywords to find frequencies and points for AstraZeneca (Figure 4), Pfizer (Figure 5), and Moderna (Figure 6). According to our tweets, Moderna is the most used vaccine and has the most negative and positive sentiment.The keywords in the tweets helped us learn about different vaccines and their frequency. Predict tweets and get results. Applying a few more techniques can further improve the results. The algorithm used was useful for comparing results. For best results, model performance evaluation is based on various metrics. This study used precision, recall,F1-score, and a confusion matrix with different values such as true positive (TP), false positive (FP), true negative (TN),and false negative (FN).

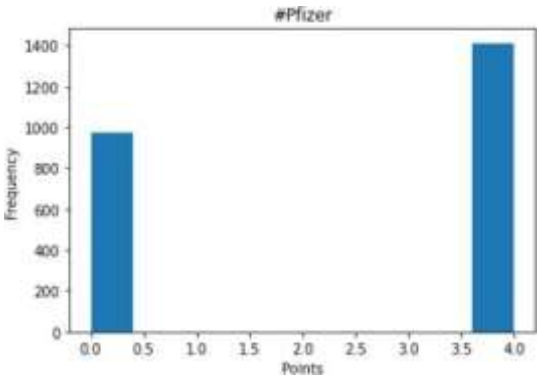


FIGURE 5. Result for the Pfizer

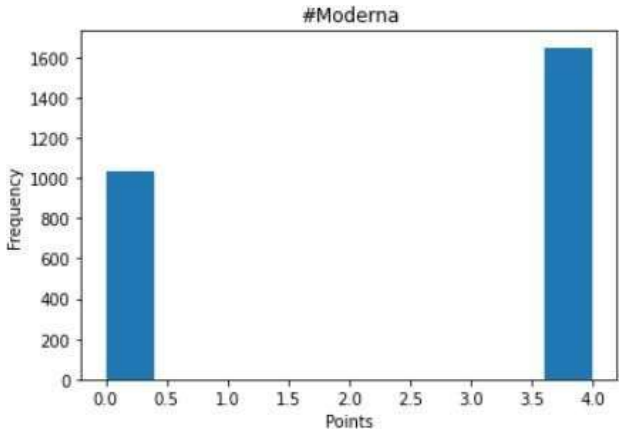


FIGURE 6. Result for the Moderna

6. DISCUSSION

In our project, we have achieved 83% accuracy in sentiment classification of COVID-19 vaccine using ABSA and TF- IDF. This places the accuracy of our work higher than the previous works which have accuracy of less than or equal to67%. In our project, the higher accuracy in sentiment classification comes because of ABSA which is a much newer model compared to Bag of Words model which isan outdated model and is used in many of the previous works. Our work has also used five traditional machinelearning models which has allowed us to test the accuracy ofour project much more reliably than previous works where usually less than 4 models are used for testing accuracy of the classification. Our work does have a shortcoming, whichis the computationally expensive nature of ABSA model used in our project. Making ABSA computationally cheap could be the direction of future work in this project.

7. CONTRIBUTION

As we have seen, we get the highest prediction model per-formance and accuracy by using RNN like LSTM and BI-LSTM in a previous literature review. So what we are doing here is using ABSA and TF-IDF together for sentiment analysis. The VADER and text blob are also well placed, but they are about 67% to 61% accurate and even outdated and losing popularity due to text context that cannot be read or understood. So we used five machine learning algorithmslike Random Forest, Naive Bayes, Support Vector Machine, Logistic Regression and Ensemble Classification together toget the best results and to perform comparison.

We used five algorithms to determine which algorithm gives us the best result .We even used ensemble sorting, which helps create multiple models and then combine them to get better results. Since we should get better performancethan RNN with LSTM and BiLSTM, we prefer to compare more algorithms. There are algorithms such as next K algorithm, but the result is insufficient. The Sentiment140 dataset is an open-source dataset containing 1.6 million tweets on products, brands, and topics. We searched the data set by region and it was labeled positive and negative with 0 and 4 respectively.

8. CONCLUSION

The validity can be ensured as the research methods we used are accurately representing the depth and format of thedata we required. The main question here is where does our project stand. The focus here is to predict the best results forthe analysis of sentiments of COVID 19 vaccine. This work will aid health

researchers in gaining a thorough understand- ing of the challenges surrounding immunisation. Companies that create vaccines, governments, health ministries from many nations, and health authorities, may all have a good notion of whether their vaccines are successful or not.

This research will aid health experts and government officials in better planning immunisation efforts. The work would also show if the government should make vaccinationmandatory in order to curb rising anti-COVID-19 attitudes. This work would also show the sentiments of the already vaccinated people. The high amount of negative sentiments regarding vaccines could indicate to the government regard-ing side effects experienced by people already vaccinated. This work would also allow the government to manage the public messaging regarding COVID-19 vaccines. High degree of negative sentiments regarding the vaccines could be deemed as failure of the Government to properly incul- cate the public regarding the hazard of avoiding COVID-19 vaccines.

For the deviation from our initial work and changes we are stuck in choosing ABSA or BOG (bag of words). Fi- nally, it's observed that ABSA, i.e., Aspect Based Sentiment Analysis performed far better than VADER and TextBlob. Itis due to to its ability to concentrate on the required aspects increased by the attention based electrical device model. In addition, it had considerably better f1-score and accuracy than other 2 models. Thus, ABSA is preferred for tasks that have a narrow focus instead of general purpose models. However, it has been noted to be a lot slower than the other same methods, which suggests that quick classificationfor basic analysis can't be obtained on most machines by the usage of ABSA. Future work will involve the event of ABSA models trained on restricted vocabulary bearing on the particular task at hand to scale back the scale of the network and therefore the computation needed to analyse every sentence. This may facilitate in reducing the amount of time taken for ABSA to run furthermore as more machines would have the ability to run it thanks to the diminished processing demands.

9. REFERENCES

[1] Becker, Benedikt Larson, Heidi Bonhoeffer, Jan van Mulligen, Erik M. Kors, Jan Sturkenboom, Miriam. (2016). Evaluation of a multinational, multilingual vaccine debate on Twitter. Vaccine. 34. 10.1016/j.vaccine.2016.11.007.

[2] E. Nwafor, R. Vaughan and C. Kolimago, "Covid Vaccine Sentiment Analysis by Geographic Region," 2021 IEEE International Confer- ence on Big Data (Big Data), 2021, pp. 4401-4404, doi: 10.1109/Big-Data52589.2021.9671854.

[3] M. A. Mudassir, Y. Mor, R. Munot and R. Shankarmani, "Sentiment Analysis of COVID-19 Vaccine Perception Using NLP," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 516-521, doi: 10.1109/ICIRCA51532.2021.9544512..

[4] N. S. Sattar and S. Arifuzzaman, "COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA," Applied Sciences, vol. 11, no. 13, p. 6128, Jun. 2021, doi: 10.3390/app11136128

[5] Kazi Nabiul Alam, Md Shakib Khan, Abdur Rab Dhruba, Moham- mad Monirujjaman Khan, Jehad F. Al-Amri, Mehedi Masud, Majdi Rawashdeh, "Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data", Computational and Mathemat-ical Methods in Medicine, vol. 2021, Article ID 4321131, 15 pages, 2021. https://doi.org/10.1155/2021/4321131

[6] Cotfas, L. A., Delcea, C., Roxin, I., Ioanas, C., Gherai, D. S., Tajariol,

[7] F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opin- ions Dynamics From Tweets in the Month Following the First Vaccine Announcement. IEEE access : practical innovations, open solutions, 9, 33203–33223. https://doi.org/10.1109/ACCESS.2021.3059821

[8] Ghasiya, Piyush Okamura, Koji. (2021). Investigating COVID-19 News across Four Nations: A Topic Modeling and Sentiment Analysis Approach.

[9] IEEE Access. 9. 36645-36656. 10.1109/ACCESS.2021.3062875

[10] Lyu, J. C., Han, E. L., Luli, G. K. (2021). COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. Journal of medical Internet research, 23(6), e24435. https://doi.org/10.2196/24435

[11] Yousefinaghani, Dara, R., Mubareka, S., Papadopoulos, A., Sharif,

[12] S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. International Journal of Infectious Diseases, 108, 256–262. https://doi.org/10.1016/j.ijid.2021.05.059

[13] Xiong, Shufeng Lv, Hailian Zhao, Weiting Ji, Donghong. (2017). Towards Twitter Sentiment Classification by Multi-Level Sentiment-Enriched Word Embeddings. Neurocomputing. 275. 10.1016/j.neucom.2017.11.023.

[14] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," 2016 First IEEE International Conference on Computer Com- munication and the Internet (ICCCI), Wuhan, China, 2016, pp. 471- 475, doi: 10.1109/CCI.2016.7778967.

[15] Raghupathi, Viju Ren, Jie Raghupathi, Wullianallur. (2020). Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets. International Journal of Environmental Research and Public Health. 17. 3464. 10.3390/ijerph17103464.

[16] Jang, Hyeju Rempel, Emily Carenini, Giuseppe Janjua, Naveed. (2020). Exploratory Analysis of COVID-19 Related Tweets in North America to Inform Public Health Institutes. 10.18653/v1/2020.nlpcovid19-2.18.

[17] Zainuddin, Nurulhuda Selamat, Ali. (2014). Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings. 333- 337. 10.1109/I4CT.2014.6914200.

[18] Kunal, Sourav Saha, Arijit Varma, Aman Tiwari, Vivek. (2018). Textual Dissection of Live Twitter Reviews using Naive Bayes. Procedia Computer Science. 132. 307-313. 10.1016/j.procs.2018.05.182.

[19] Rezwanul, Mohammad Ali, Ahmad Rahman, Anika. (2017). Sen- timent Analysis on Twitter Data using KNN and SVM. Interna- tional Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.080603.

[20] E. D'Andrea, P. Ducange and F. Marcelloni, "Monitoring negative opinion about vaccines from tweets analysis," 2017 Third International Conference on Research in Computational Intelligence and Communication Net-works (ICRCICN), Kolkata, India, 2017, pp. 186- 191, doi: 10.1109/ICR-CICN.2017.8234504

[21] Salathé M, Khandelwal S (2011) Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dy- namics and Control. PLOS Computational Biology 7(10): e1002199. https://doi.org/10.1371/journal.pcbi.1002199

[22] KazAnova. (2017, September). Sentiment140 dataset with 1.6 mil- lion tweets, Version 2. [Online] Retrieved March 4, 2022 from https://www.kaggle.com/kazanova/sentiment140

[23] Anirudha Simha, "Understanding TF-IDF for Machine Learning" ,06- Oct-2021. [Online] Available: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf. [Accessed: 4-Mar-2022]

[24] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

