Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm

F. M. Javed Mehedi Shamrat[1], Sovon Chakraborty[2], M. M. Imran[3], Jannatun Naeem Muna[4], Md. Masum Billah[5], Protiva Das[6], Md. Obaidur Rahman[7]
[1,5]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2,4,7]Department of Computer Science and Engineering, European University of Bangladesh, Dhaka, Bangladesh
[3]Cefalo Bangladesh Limited
[6]BRAC University, Dhaka, Bangladesh

Article Info

Article history:

Received Mar 30, 2021
Revised Jun 8, 2021
Accepted Jun 17, 2021

Keywords:

COVID 19
KNN
NLP
Sentiment analysis
Vaccines

ABSTRACT

The pandemic has taken the world by storm. Almost the entire world went into lockdown to save the people from the deadly COVID-19. Scientists around the around have come up with several vaccines for the virus. Among them, Pfizer, Moderna, and AstraZeneca have become quite famous. General people however have been expressing their feelings about the safety and effectiveness of the vaccines on social media like Twitter. In this study, such tweets are being extracted from Twitter using a Twitter API authentication token. The raw tweets are stored and processed using NLP. The processed data is then classified using a supervised KNN classification algorithm. The algorithm classifies the data into three classes, positive, negative, and neutral. These classes refer to the sentiment of the general people whose Tweets are extracted for analysis. From the analysis it is seen that Pfizer shows 47.29% positive, 37.5% negative, and 15.21% neutral, Moderna shows 46.16% positive, 40.71% negative, and 13.13% neutral, AstraZeneca shows 40.08% positive, 40.06% negative, and 13.86% neutral sentiment.

Corresponding Author:

F. M. Javed Mehedi Shamrat
Department of Software Engineering
Daffodil International University
102/1, Sukrabad, Mirpur Road, Dhaka 1207, Bangladesh
Email: javedmehedicom@gmail.com

1. INTRODUCTION

The planet faced a major coronavirus epidemic by the end of 2019. The virus spreads quickly across different media. Maximum countries lock their citizens to deter the transmission of this lethal virus. People shifted exorbitantly in their minds. Social networking connectivity has been a common location for people to express their emotions. Some 20 firms attempted to create the vaccine. Vaccinations, including Pfizer, AstraZeneca, and Moderna have been endorsed by the world health organization (WHO) [1]. The licensed vaccinations received a mixed assessment of the general public's based on effectiveness. Over time, an extensive amount of sentiments have been shared about the side effects of the approved vaccines on Twitter.

During the Covid-19 timeframe, numerous volumes of research were carried out based on group feelings. And Twitter has become a popular source for data collection for conducting various researches [2], [3]. Jia Xue *et al.* in [4] used multiple hashtags to extract data from Twitter. Using the LDA machine learning algorithm on this data, sentiment analysis is done. It is found that fear is significant in the discussion about covid-19. In the paper [5], Twitter data is extracted manually by data crawling using Twitter API access token with “Vaccine” and “COVID-19” as keywords. Naïve Bayes algorithm is used for sentiment analysis and is found that the majority of the tweets have a negative sentiment. The main purpose of the paper [6] is to extract Twitter data using the keywords “COVID”. The data is preprocessed using NLP. Sentiment classification is done on the data using RNN. The authors in the paper [7] extracted raw tweets from Twitter using a list of keywords. To determine the appropriate keywords a census of two cycles was done. To preprocess the raw data, NLP was used. Topic modeling was used to determine the semantic structure of the processed tweet using an unsupervised LDA algorithm. To determine the positive, negative, and neutral sentiment behind the tweets, valence aware dictionary and sentiment reasoner (VADER) is implemented.

In this research, an analysis based on public sentiments about approved covid-19 vaccines based on Twitter data is showed using natural language processing and supervised KNN classification algorithm. The raw data used in the process is the Tweets extracted from Twitter related to the Covid-19 vaccines, Pfizer, Moderna, and AstraZeneca. These tweets are preprocessed using NLP. The processed text data are then converted to polarity and subjectivity for classification as machine learning algorithms [8] cannot classify text data. For classification, KNN classification is used for sentiment analysis. The algorithm classifies the data into three classes, positive, negative, and neutral. These classes determine the sentiment of the tweets about the three vaccines.

The following section follows the same structure. This section contains the most recent research and studies in this field. The analysis approach for modeling the whole system is defined in Section 2. The third section examines the outcomes of the scheme that has been implemented. Section 4 concludes with a hypothesis, shortcomings, and research ideas.

2. METHODOLOGY

In this paper, a system is proposed to analyze Twitter tweets about the three COVID-19 vaccines (Pfizer, Moderna, and AstraZeneca) and show the positivity or negativity of sentiment in the text using natural language processing (NLP) and a supervised machine learning classification algorithm. To implement the system, first tweet data is fetched from Twitter by Twitter standard search using Tweepy library, and the retrieved data is saved in CSV file format. This data needs preprocessing as it contains special characters, hyperlinks, retweets, emoji, and stickers. Natural language processing is used to preprocess the data and to make it suitable to implement a supervised classification algorithm [9]. After removing the special characters, data is tokenized. For further processing, normalization and lemmatization of the data are done. After the data is preprocessed using NLP, polarity, and subjectivity are calculated. A supervised KNN classifier is used for the classification of the polarity data. Finally, data visualization is done on the classified data and further analyzed for comparison. The complete system diagram is illustrated in Figure 1.

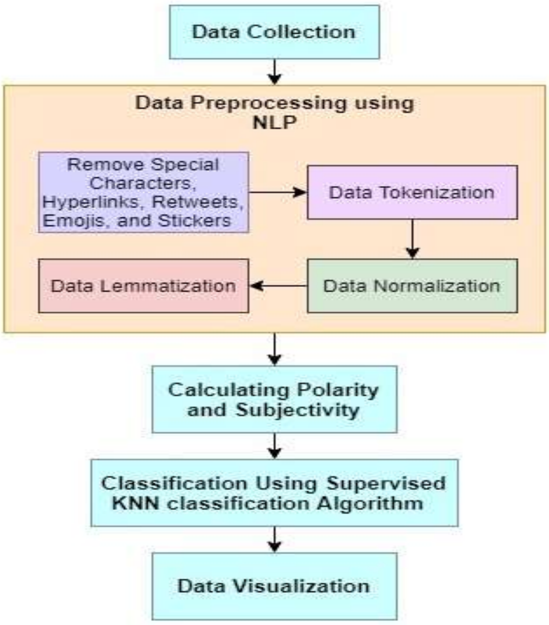
Data Collection
Data Preprocessing using NLP
Remove Special Characters, Hyperlinks, Retweets, Emojis, and Stickers
Data Tokenization
Data Normalization
Data Lemmatization
Calculating Polarity and Subjectivity
Classification Using Supervised KNN classification Algorithm
Data Visualization

Figure 1. Proposed system diagram

2.1. Data collection

The implementation worked for the proposed system is done on text data from Twitter tweets specifically the ones that are related to COVID-19 vaccines. To extract tweets [10], a Twitter developer account is mandatory. There Twitter keys/API credentials will be stored in variables and then create the authentication object. Finally, the access token and access token secret are set and is authenticated to Twitter. Figure 2 shows the flow diagram of the process.

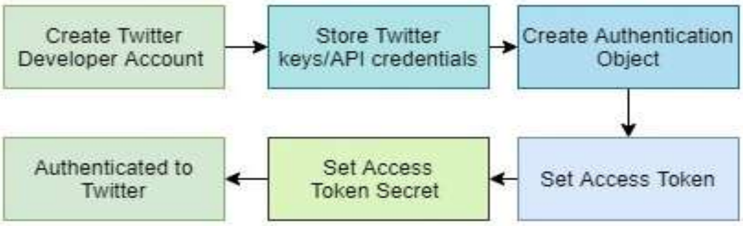
Create Twitter Developer Account
Store Twitter keys/API credentials
Create Authentication Object
Set Access Token
Set Access Token Secret
Authenticated to Twitter

Figure 2. Twitter authentication

To extract the tweets of the three COVID-19 vaccines from Twitter, tweets using vaccine hashtags, i.e. #Pfizer, #Moderna, and #AstraZeneca are targeted. Ten thousand (10,000) tweets for each hashtag are fetched. From the tweets, only the text data are extracted and stored in CSV format as the dataset.

2.2. Preprocessing dataset

To make data suitable for the application of machine learning algorithms, raw data has to go through the preprocessing stage. Natural Language Processing [11] is used in this system for data preprocessing. In this stage, firstly the text data is converted to lower case. From this form, all stop words are removed and contractions are replaced. A list of stop words is defined in the python nltk library that is used in this process and to replace contraction a custom function is created to complete the task. To avoid complexity, a spelling check is done to fix misspelled words. One of the most important steps in preprocessing in this work is to replace emoji with the expression they represent in plain English for instance :) / :-) with the English text “smiley”. Next, the special characters, URLs, and HTML tags are removed from the text. Finally, tokenization [12], normalization, and lemmatization are done on the text data before moving to Object Identification. Figure 3 demonstrates the flow of the process.

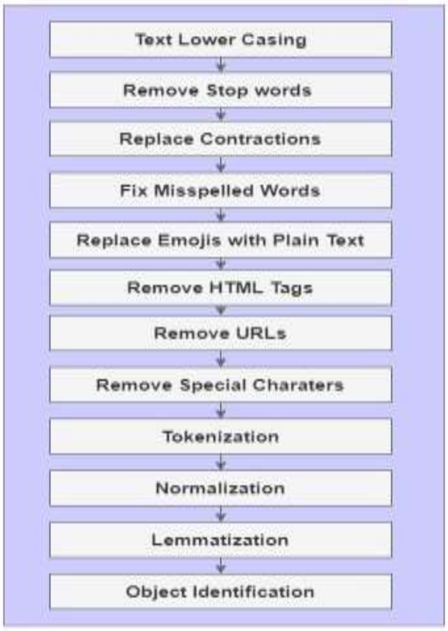

Figure 3. Data preprocessing

Tokenization, normalization, and lemmatization are three major functions in natural language processing for preprocessing text before classification.

1) Tokenization: In NLP, tokenization refers to splitting a text document into small units. Each unit is called a token. In this work, each word is converted into a token [13].
2) Normalization: Text normalization is to convert any unusual text into its standard form. At times, people write a word in an unusual form to express themselves [14]. This text needs to be converted into its correct form and correct spelling.
3) Lemmatization: A word can have different forms based on it tense, gender and comparison adjective, the base form of each word is called lemma and the process of converting any word to its base form is referred to as lemmatization.
4) Object Identification: The final step of preprocessing is object identification. This function fetches each data column and checks if it is blank [15]. If the column is blank, it set the value 0 and else sets value 1 and stored it in a new identification column.

2.3. Calculating polarity and subjectivity

Basically, sentiment analysis depends on polarity and subjectivity. Subjectivity contains facts, opinions and desires. Polarity contains feelings and emotions. To analyze the sentiment [16], polarity and subjectivity of text have to be calculated. For this python library call TextBlob. To process NLP tasks such as Sentiment Analysis, the TextBlob python library provides an API.

From the polarity and subjectivity data, mean, median, average minimum, average maximum is calculated for each vaccine. Maximum average polarity is calculated per 10 tweets. The equation used in the calculations is illustrated.

$$\text{mean}, \bar{x} = \frac{\sum x}{n} \tag{1}$$

$$\text{median} = \frac{n+1}{2} \tag{2}$$

$$\text{average Minimum } = \frac{(n-1)\min + \max}{n} \tag{3}$$

$$\text{average maximum } = \frac{\min + (n-1)\max}{n} \tag{4}$$

2.4. Data classification using KNN

Data classification is done to the polarity score. If a tweet has a polarity score greater than zero (Polarity>0) then it is a positive tweet. If the polarity score is less than zero (Polarity<0), it is a negative tweet. If the polarity score is equal to zero (Tweet Polarity==0), it is neutral.

To obtain the classification result, a supervised k-nearest neighbor (KNN) classification algorithm [17] is used. KNN uses feature similarity where it assigns a data point based on how close it is to its neighbor. The algorithm for the KNN that is shown in algorithm 1 is used for the classification of the data.

```
Algorithm 1: K-nearest neighbor classification algorithm
Step 1: Load dataset
Step 2: Select the value of K
Step 3: Calculate the distance between each data point using Euclidean distance
Step 4: Sort data point according to the distance calculated
Step 5: Select the top K row
Step 6: Assign data point on the most frequent class
Step 7: END
```

To calculate the distance of the data point in the KNN algorithm Euclidean distance [18] is calculated using (5).

$$d\,(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{5}$$

The KNN algorithm uses tweet polarity for classification. It classifies the data into a positive negative and neural class. The classification result is stored and analyzed.

3. RESULTS AND DISCUSSION

Visualization of processed Tweet is done using word cloud. word cloud shows the most occurred word in the data and is categorized by different sizes for difference score [19], [20]. Figure 4, shows the word cloud of Pfizer, Moderna, and AstraZeneca tweet data.

Using (1) to (4), mean, median, average minimum (amin), average maximum (amax) is calculated for the three vaccines based on the tweet polarity and subjectivity. The result of the calculation is shown in Tables 1 and 2 respectively.

The text polarity scores and subjectivity scores for the three vaccines, Pfizer, Moderna, and AstraZeneca are plotted in bar diagrams for visualization. At the same time, the scatter plot is also demonstrated for a better understanding of the frequency of the scores [21], [22]. Figures 5 and 6 demonstrate the bar diagram of polarity and subjectivity respectively of the three mentioned vaccines. Figure 7 shows the scatter plot of the scores for the same.

The maximum average is calculated from the polarity of 10 tweets [23]. The obtained values are stored and the visualization of the result is shown in the following Figure 8. This diagram shows the results for the Pfizer, Moderna, and AstraZeneca vaccines separately.

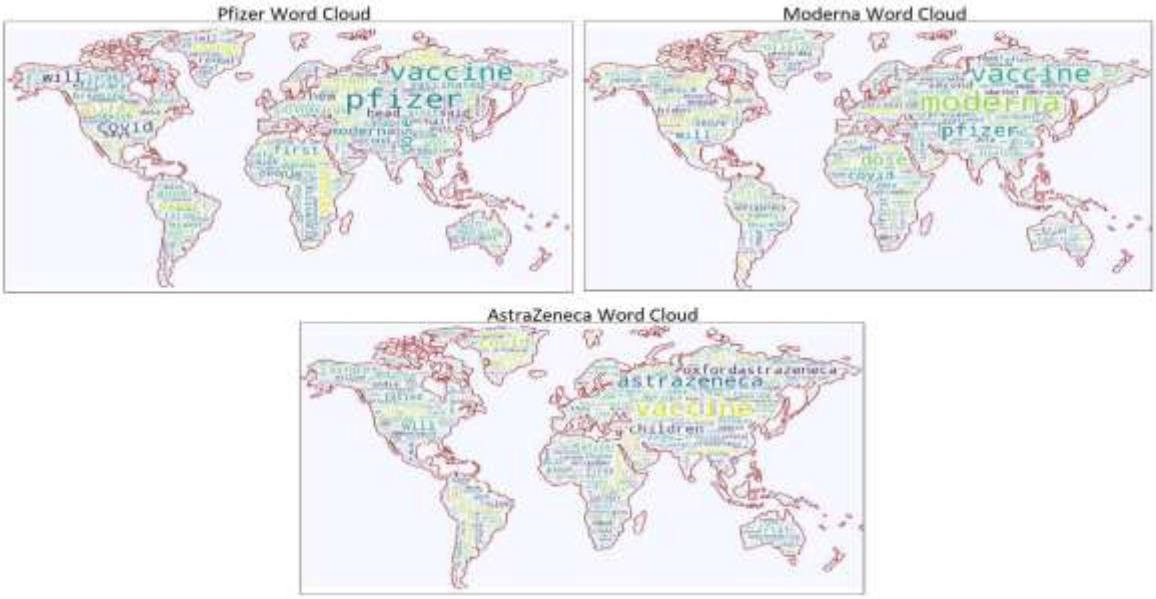

Figure 4. Word cloud visualization of the three vaccines tweet data

Table 1. Polarity calculation of the three vaccine tweets

Name	Polarity			
	mean	amax	amin	median
Pfizer	0.133202	1.0	-1.0	0.0
Moderna	0.114308	1.0	-1.0	0.0
AstraZeneca	0.078346	1.0	-1.0	0.0

Table 2. Subjectivity calculation of the three vaccine tweets

Name	Subjectivity			
	mean	amax	amin	median
Pfizer	0.34011	1.0	0.0	0.333333
Moderna	0.348156	1.0	0.0	0.375
AstraZeneca	0.304983	1.0	0.0	0.3

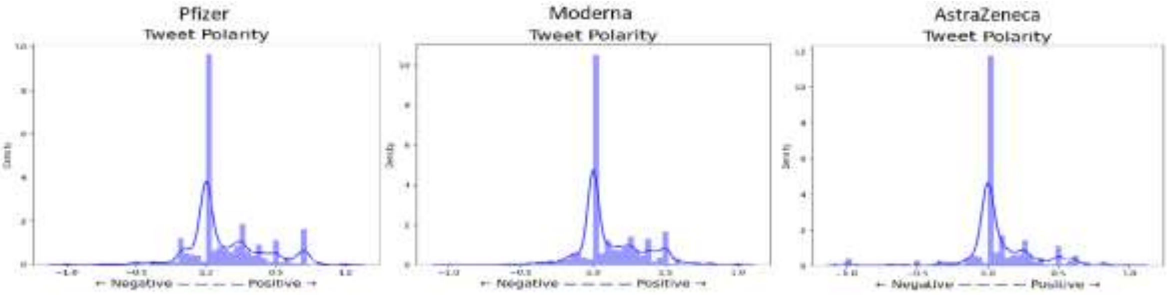

Figure 5. Tweet Polarity bar diagram of the three vaccines

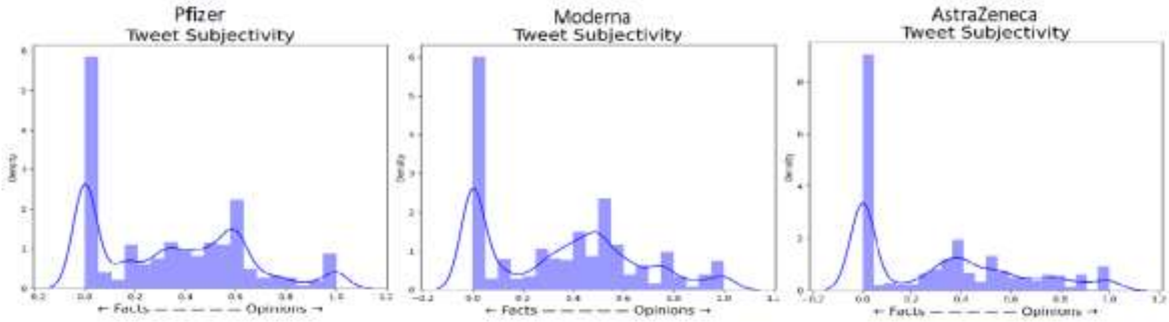

Figure 6. Tweet Subjectivity bar plots of the three vaccines

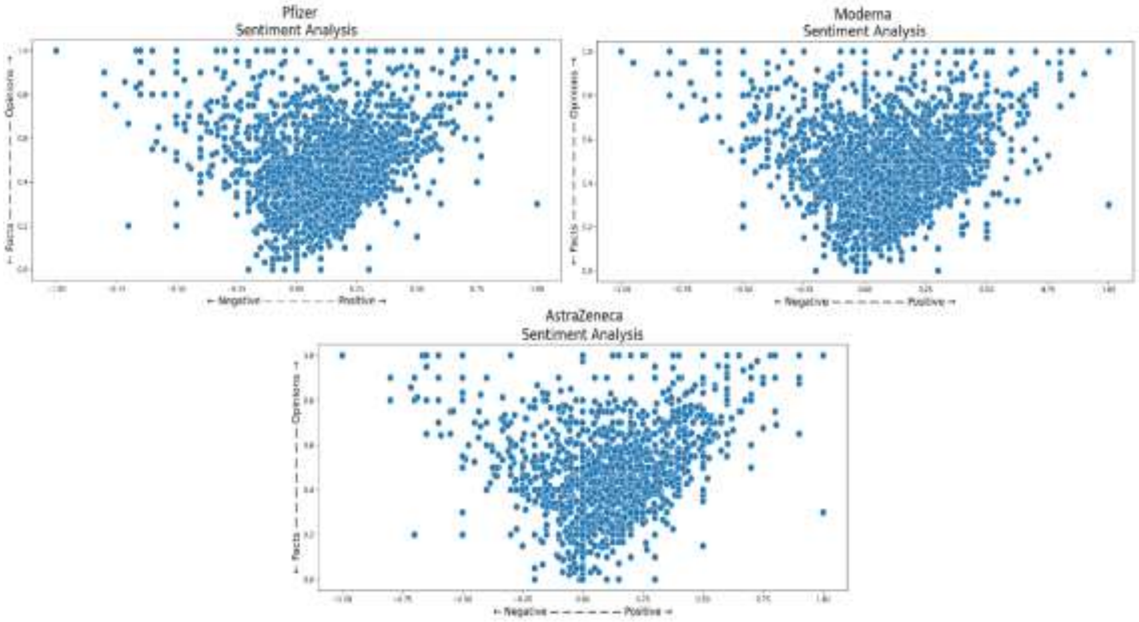

Figure 7. Polarity and subjectivity scatter plots of the three vaccines

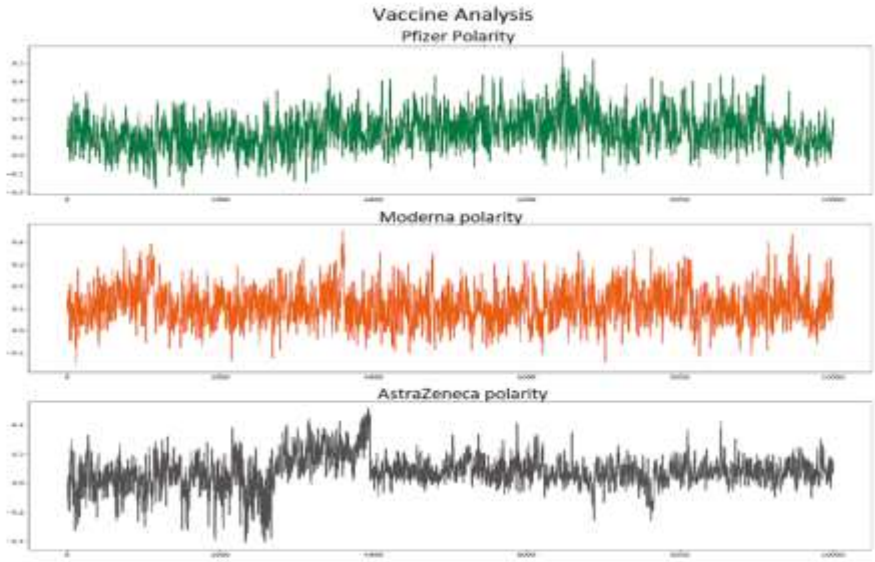

Figure 8. Maximum average tweet polarity per 10 tweets

The text data are converted into polarity scores. This score is used in the classification process as the KNN classification algorithm cannot process text data [24], [25]. Furthermore, polarity shows the emotion or sentiment behind text data.

In the proposed system, a supervised KNN classification algorithm is implemented. This algorithm classifies the polarity score into three classes, positive, negative, and neutral. Classification is done on the data of all three vaccines in the discussion. The final result of the classification is stated in Table 3.

Table 3. Classification of sentiment based on tweets about the three vaccines

Name	Positive	Negative	Neutral
Pfizer	47.29	37.5	15.21
Moderna	46.16	40.71	13.13
AstraZeneca	40.08	40.06	13.86

It can be seen that there are positive, negative, and neutral sentiments in the tweet data about the three different vaccines. For better understanding, visualization of the numbers is shown in Figure 9. Here we can see compared to Pfizer and Moderna vaccine, general people have much less positive sentiment towards AstraZeneca vaccine and higher negative sentiment as well.

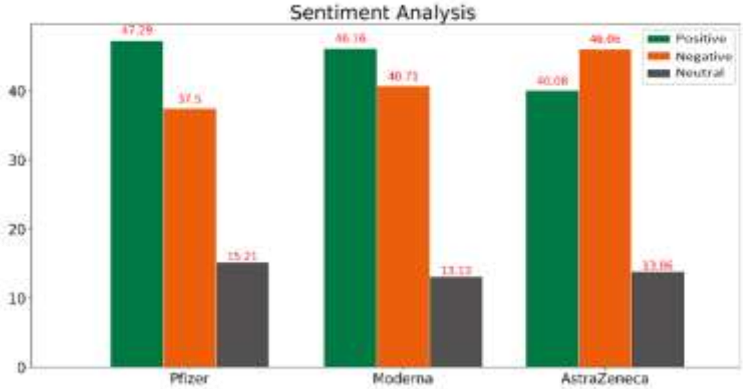

Figure 9. Sentiment comparison about Pfizer, Moderna, and astrazeneca vaccines

4. CONCLUSION

This study illustrates general peoples' sentiment towards the Pfizer, Moderna, and AstraZeneca vaccines made to fight COVIS-19. During the pandemic, people under lockdown have been expressing their feeling on social media like Twitter about COVID-19 and its vaccines. Therefore Twitter has become an important source of information. Extracting such tweets, authors analyzed the sentiments of general people towards the vaccines. Using NLP to preprocess the raw tweets and KNN Classification Algorithm to classify the processed data, it is seen that general people have higher positive sentiment towards Pfizer and Moderna vaccine with the rate of 47.29 and 46.16 respectively compare to AstraZeneca vaccine with a rate of 40.08. This analysis can help the authority interact with the people and provide them the vaccine they trust and peacefully control the pandemic.

REFERENCES

[1] Le, B, Nguen H, "Twitter Sentiment Analysis Using Machine Learning Techniques," *Advanced Computational Methods for Knowledge Engineerin*g, pp. 279-289, AISC, vol. 358, doi: 10.1007/978-3-319-17996-4_25.

[2] Agarwal B, Nayak R, Mittal N, Patnaik S, "Deep Learning-Based Approaches for Sentiment Analysis," Part of the Algorithms for Intelligent Systems book series [AIS], doi: 10.3390/electronics9030483.

[3] Sarlan A, Nadam C, Basri, S, "Twitter sentiment analysis," *Proceedings of the 6th International Conference on Information Technology and Multimedia, IEEE*, pp. 212-216, 2014, doi: 10.1109/ICIMU.2014.7066632.

[4] Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, Zhu T, "Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach," *J Med Internet Res.*, vol. 22, no. 11, e20550, 2020, url: https://www.jmir.org/2020/11/e20550, doi: 10.2196/20550.

[5] Pristiyono, Mulkan Ritonga, Muhammad Ali Al Ihsan, Agus Anjar, Fauziah Hanum Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *Annual Conference on Computer Science and Engineering Technology (AC2SET) 2020*, vol. 1088, 012045, 2021, doi:10.1088/1757-899X/1088/1/012045.

[6] László Nemes and Attila Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1-15, doi:10.1080/24751839.2020.1790793.

[7] Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, Hon SD, Park J, and Dang P, "Lipsky MS Social Network Analysis of COVID-19 Sentiments," *Application of Artificial Intelligence J Med Internet Res*., vol. 22, no. 8, e22590, 2020, doi: 10.2196/22590 PMID: 32750001 PMCID: 7438102.

[8] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bangluru, India, pp. 1-7, 2020, doi: 10.1109/INOCON50539.2020.9298026.

[9] P. Ghosh *et al*., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in *IEEE Access*, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[10] F. M. Javed Mehedi Shamrat, Zarrin Tasnim, A. K. M Sazzadur Rahman, Naimul Islam Nobel, Syed Akhter Hossain, "An Effective Implementation of Web Crawling Technology to Retrieve Data from the World Wide Web (www)," *International Journal of Scientific & Technology Research*, vol. 9, no. 1, ISSN: 2277-8616, pp. 1252-1256, January 2020, doi: 10.1109/ACCESS.2021.3053759.

[11] Pang B, and Lee, L, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, "In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Barcelona, Spain, July 2004, p. 271, doi: 10.3115/1218955.1218990.

[12] F. M. Javed Mehedi Shamrat, Z. Tasnim, P. Ghosh, A. Majumder and M. Z. Hasan, "Personalization of Job Circular Announcement to Applicants Using Decision Tree Classification Algorithm," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bangluru, India, pp. 1-5, 2020, doi: 10.1109/INOCON50539.2020.9298253.

[13] Hussain Ghulam, Feng Zeng, Wenjia Li, Yutong Xiao, "Deep Learning-Based Sentiment Analysis for Roman Urdu Text," *Procedia Computer Science*, vol. 147, pp. 131-135, 2019, ISSN 1877-0509, doi: 10.1016/j.procs.2019.01.202.

[14] Mehta R. P, Sanghvi M. A and Shah D. K, Singh A, "Sentiment Analysis of Tweets Using Supervised Learning Algorithms," In: Luhach A., Kosa J., Poonia R., Gao XZ., Singh D. (eds) *First International Conference on Sustainable Technologies for Computational Intelligence*, Advances in Intelligent Systems and Computing, vol. 1045, Springer, Singapore, 2020, doi: 10.1007/978-981-15-0029-9_26.

[15] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu, "Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary," in *IEEE Access*, vol. 7, pp. 43749-43762, 2019, doi: 10.1109/ACCESS.2019.2907772.

[16] Sun X, and He, J, "A novel approach to generate a large scale of supervised data for short text sentiment analysis," *Multimed Tools Appl*, vol. 79, pp. 5439-5459, 2020, doi: 10.1007/s11042-018-5748-4.

[17] Emadi, M, Rahgozar, M, "Twitter sentiment analysis using fuzzy integral classifier fusion," *J Inform Sci 2020*, vol. 46, no. 2, pp. 226-242, 2020, doi: 10.1177/0165551519828627.

[18] Díaz-Faes, A. A, Bowman, T. D, Costas, R, "Towards a second generation of 'social media metrics': characterizing Twitter communities of attention around science," *PLoS One 2019*, vol. 14, no. 5, e0216408, 2019.

[19] Agarwal, A, Xie, B, Vovsha, I, Rambow, O, and Passonneau, R. J, "Sentiment analysis of twitter data," In *Proceedings of the workshop on language in social media (LSM 2011)*, pp. 30-38, June 2011,

[20] Kouloumpis, E, Wilson, T, and Moore, J, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 538-541, 2011, Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14185

[21] M.Rambocas, and J. Gama, "MarketingResearch:TheRoleof SentimentAnalysis," The 5th SNA-KDD Workshop'11. Universityof Porto, p. 489, 2013.

[22] Hassan, S.-U, *et al*., "Sentiment analysis of tweets through Altmetrics: A machine learning approach," *Journal of Information Science*, 2020, doi: 10.1177/0165551520930917.

[23] A. H. Huang, D. C. Yen, and X. Zhang, "Exploring the effects of emoticons," *Information & Management*, vol. 45, no. 7, pp. 466-473, 2008, doi: 10.1016/j.im.2008.07.001.

[24] D. Boyd, S. Golder and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," *2010 43rd Hawaii International Conference on System Sciences*, pp. 1-10, 2010, doi: 10.1109/HICSS.2010.412.

[25] Ji, X, Chun, S. A, Wei, Z, and James Geller, "Twitter sentiment classification for measuring public health concerns," *Soc. Netw. Anal. Min.* vol. 5, no. 13, 2015, doi: 10.1007/s13278-015-0253-5.