COVID VACCINE SENTIMENT ANALYSIS

Sree Divya Sudagoni
Department of Computer Science
California State University, Northridge

***Abstract*—The COVID-19 pandemic has created unprecedented challenges and has caused significant damage to global public health and the economy. The COVID-19 vaccine is the most effective solution to prevent the spread of the virus, but there has been considerable debate on its safety and effectiveness. This research project aims to perform Sentiment Analysis on COVID vaccine-related tweets using Machine Learning algorithms. We focus on extracting data from Twitter, pre-processing it to remove irrelevant information n, and training and testing machine learning models for sentiment classification. Various algorithms such as Naïve Bayes, Support Vector Machines, Random Forest, and fine- tuning pre-trained models will be explored to identify the most effective approach. The performance of the models is evaluated using Accuracy, Precision, Recall, F1-score, and Confusion matrix metrics. This study provides insights into public opinion and sentiment towards the COVID vaccine and is useful for policymakers, healthcare professionals, and researchers.**

***Index Terms*—Keywords—Sentiment Analysis, Twitter, Naïve Bayes, Support Vector Machines, Random Forest.**

I. Introduction

The COVID-19 pandemic, which first emerged in December 2019, has had a profound impact on society, leading to significant loss of life and economic disruption worldwide. As of March 2023, there have been over 480 million confirmed cases and 6.2 million deaths worldwide. The development of effective vaccines has provided hope for a way out of the crisis, but vaccine hesitancy remains a significant concern. A study by the World Health Organization (WHO) found that globally, 27 percentage of people reported being hesitant or unsure about COVID-19 vaccines as of December 2021. Social media platforms like Twitter have become a major source of information for many people, with over 192 million daily active users worldwide. As a result, analyzing tweets related to COVID vaccines provide valuable insights into public sentiment and helps address concerns. Sentiment analysis, a Natural Language Processing technique, can classify text data into positive, negative, or neutral categories based on the text's sentiment. This project aims to perform sentiment analysis on a large corpus of tweets related to COVID vaccines, using Machine Learning algorithms to classify tweets by their sentiment.

II. Related Work

Several studies have stated that effective policies can be made using the data from social media, such as the study by R. Singh., and P. Sharma.[1] The paper explores the different social media platforms, the types of data they generate, and the techniques used for sentiment analysis. The authors review different approaches to sentiment analysis, such as lexicon-based methods, machine learning-based methods, and hybrid methods that combine both approaches. The paper also discusses the challenges associated with sentiment analysis, such as data quality, language ambiguity, and context dependence. The study gives insights into the data.generated by social media platforms, such as text, images, and videos, and the challenges associated with analyzing each type of data.

Yuhao Xie, Tengjiang Wang, Hexuan Zhang, and Tianyuan Yan. 2022. [2] analyzed the rate of increase in vaccines administrated versus Twitter Sentiment Analysis. This study can be used to get insightful information on the number of booster shots required per demand by predicting using sentiment analysis of Twitter comments. The potential of Natural Language Processing and machine learning approaches to analyze social media data and derive insights about public opinion on vaccines is generally highlighted in this research. There are many issues while analyzing sentiments related to the use of abbreviations, emoticons, and hashtags, the presence of noise and sarcasm, and the limited length of tweets. David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018.[3] reported that classification accuracies were below 70 percentage, caused due to the failure in capturing sarcasm and irony, the inability to handle negation and intensification, and the difficulty in detecting the sentiments of neural tweets.

Chandrasekaran R, Mehta V, Valkunde T, Moustakas E.[4] analyzed the topics, sentiments, and changing trends in 13.9 million COVID-19-related tweets using topic modeling and sentiment analysis techniques such as latent Dirichlet allocation (LDA) and VADER sentiment analysis. The authors used traditional LDA to identify the topics and potential keywords related to COVID. Then they used guided LDA to generate a list of topics and anchor words. Further computed a sentiment score for each tweet using the VADER tool to classify the sentiment as positive, negative, or neutral based on the compound score. By researching the tools and techniques utilized in this study, we can apply similar ways to examine the feelings associated with COVID-19 vaccination.

The research Analyzing "Application of Machine Learning Algorithms in Healthcare Sector Review" by K. Shailaja, B. Seetharamulu and M. A. Jabbar [5] published in 2020, aimed to provide a comprehensive review of the current state of machine learning algorithms in the healthcare sector. The paper's discussion on data pre-processing methods is also relevant to our sentiment analysis task, as we need to preprocess the tweet data to extract relevant features and normalize the data before training the machine learning model. F. Jemai, M. Hayouni and S. Baccar. [6] proposed a Survey on Sentiment analysis algorithms and applications. This study explains how lexicon-based approaches use a pre-defined dictionary of words with associated sentiment scores to classify the text.

A well-chosen algorithm can provide accurate and meaningful insights into the data. A popular machine learning technique used for classification tasks is the Naïve Bayes algorithm. A study by F. - J. Yang. [7] describes the implementation of this by using the Gaussian Naïve Bayes approach, which assumes that the features of each class are normally distributed. The implementation is evaluated on several data sets, including the Iris data set, and the Breast Cancer Wisconsin data set. And compared to other classifiers such as Decision trees and K-Nearest neighbors. The results show comparable performance to other classifiers, making it a useful tool for classification tasks.

Another study by Z. Bingzhen, Q. Xiaoming, Y. Hemeng, and Z. Zhubo [8] explores the use of the Random Forest classification model for processing transmission line images. The authors proposed a Random Forest model that utilizes a combination of color and texture features extracted from transmission line images to classify the images into different categories. Explained how the algorithm works, feature selection, and the construction of Decision Trees. By extracting meaningful features from the text data, the accuracy and effectiveness of the sentiment analysis model can also be improved.

The paper by C. Zhang and W. Xu.[9] provides valuable insights into the use of Neural Networks for various applications, including Natural language Processing. It also discusses different neural network architectures, such as recurrent neural networks, which are commonly used for sentiment analysis tasks. This paper gives insights into techniques for optimizing neural network training, which is useful for developing an efficient sentiment analysis model. Another paper by Q. Wang.[10], discussed the mathematical framework of the Support Vector Machine algorithm and its ability to solve classification problems in a high-dimensional feature space. It also discussed different kernel-level functions that can be used to map the input data and provided guidelines for parameter selection and model optimization.

III. Thesis Statement

Thesis statement: This research paper aims to perform Sentiment Analysis on COVID vaccine-related tweets using Machine Learning algorithms, evaluating their effectiveness and providing insights into public opinion and sentiment towards the COVID vaccine. Through the exploration of various machine learning algorithms, including Naïve Bayes, Support Vector Machines, Random Forest, and fine-tuning pre-trained models, the study seeks to identify the most effective approach for sentiment classification. The evaluation of model performance using metrics such as Accuracy, Precision, Recall, F1-score, and Confusion matrix, will provide a comprehensive understanding of the models' capabilities and highlight areas for improvement. This research project is valuable to policymakers, healthcare professionals, and researchers, offering crucial insights for informed decision-making in the context of the COVID-19 pandemic.

IV. Research Methodology

Research methodology: This research study follows a quantitative approach with a focus on sentiment analysis of COVID vaccine-related tweets. The methodology involves several key steps, including data set selection, data pre-processing, algorithm selection, model training and evaluation, and result analysis. To begin, a data set named 'vaccination All Tweets' is chosen for the experiments from Kaggle. This data set consists of 228,207 rows and 16 columns, with each row representing a tweet and each column representing a specific feature of the tweet. The data set is obtained from a reliable source and is relevant to the research question. The data pre-processing phase involves applying various techniques to clean and prepare the data set for sentiment analysis. Text normalization techniques, such as removing stop words, punctuation, URLs, and numeric numbers, are applied. Tokenization, stemming, and lemmatization processes are also utilized to enhance the quality of the data set. Data visualization like generating the word cloud for negative and positive sentiments is performed.

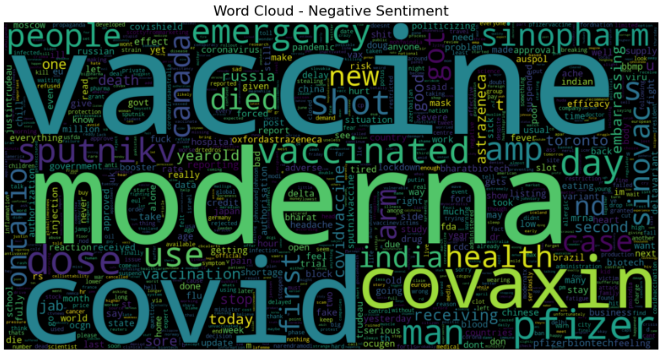

Fig. 1. Word Cloud - Negative Sentiment

For sentiment analysis, several machine learning algorithms are selected, including Support Vector Machines (SVM), Naïve Bayes, Random Forest, and Logistic Regression. These algorithms are chosen based on their effectiveness in sentiment classification tasks and their suitability for the research question. The selected algorithms are trained and evaluated using the pre processed data set. Model training involves splitting the data set into training and testing sets, and then feeding the training set into the chosen algorithms. SVM model is trained using the Radial Basis Function (RBF) which is used to transform the input data. For Naïve Bayes model a Multinomial Naïve Bayes classifier with an alpha value of 0.5 and disabled fitting prior probabilities. For Logistic Regression model a logistic regression classifier with a regularization strength of 1.0 and a maximum of 100 iterations is used for the classification. The trained models are then evaluated using various performance metrics, such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide insights into the effectiveness of each algorithm in classifying sentiments.

This study also includes the analysis and interpretation of the results obtained from the evaluation phase. The findings are discussed in the context of previous studies, highlighting patterns and trends observed in the sentiment analysis of COVID vaccine-related tweets. The prevalence of negative sentiment compared to positive sentiment is noted, aligning with prior research. It is important to acknowledge the limitations of the study. The analysis is limited to a specific time frame and a specific data set, which may not capture the entire spectrum of sentiment on the topic. Future research can expand the analysis to a larger and more diverse data set, as well as explore multilingual sentiment analysis to gain insights into global sentiment patterns. Ethical considerations are taken into account throughout the research process, ensuring the privacy and anonymity of the individuals whose tweets are included in the data set. All the models were developed and trained on Google Colaboratory. The hardware configuration for the experiments is a single CPU core and a GPU run time (Tesla T4) with 16GB memory.

V. Results

A. Support Vector Machine

The SVM model achieved an overall accuracy of 87 percentage in predicting sentiment polarity (positive, negative, or neutral) of the text data. The precision values indicate the model's ability to correctly predict instances of each sentiment class. It achieved a precision of 91 percentage for negative sentiment (class 0), 84 percentage for positive sentiment (class 1), and 93 percentage for neutral sentiment (class 2). The recall values demonstrate the model's effectiveness in capturing instances of each sentiment class. The highest recall was observed for positive sentiment (class 1) at 99 percentage, indicating the model's ability to identify the majority of positive instances. The F1-scores provide a balance between precision and recall, with negative sentiment (class 0) achieving an F1-score of 0.67, positive sentiment (class 1) achieving an F1-score of 0.91, and neutral sentiment (class 2) achieving an F1-score of 0.89. The confusion matrix from Fig. 2. further reveals that the model correctly classified a significant number of instances, with some misclassifications observed across the sentiment classes. Overall, the SVM model demonstrates reasonably good performance in predicting sentiment polarity, with notable accuracy in identifying positive sentiment and neutral sentiment instances.

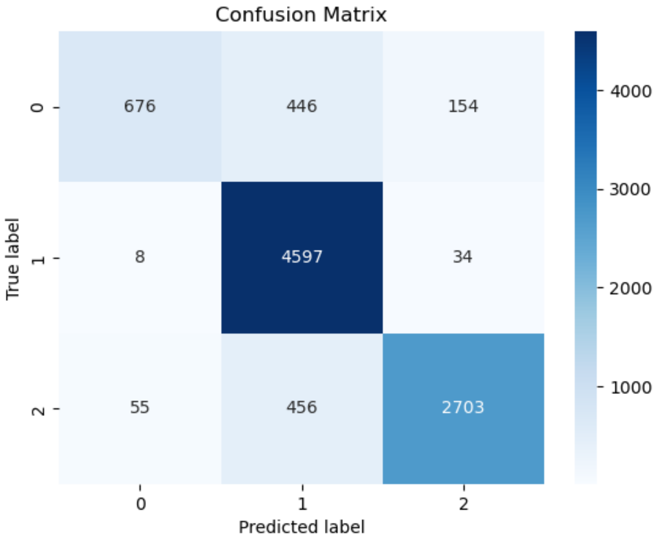

Fig. 2. Confusion Matrix of SVM

B. Naive Bayes

The Naive Bayes algorithm achieved an overall accuracy of 77 percentage in predicting sentiment polarity (positive, negative, or neutral) of the text data. Precision achieved a precision of 51 percentage for negative sentiment (class 0), 94 percentage for positive sentiment (class 1), and 74 percentage for neutral sentiment (class 2). The highest recall was observed for neutral sentiment (class 2) at 88 percentage, indicating the model's ability to identify the majority of neutral instances. The F1-scores provide a balance between precision and recall, with negative sentiment (class 0) achieving an F1-score of 0.59, positive sentiment (class 1) achieving an F1-score of 0.81, and neutral sentiment (class 2) achieving an F1-score of 0.80. The confusion matrix from Fig. 3. shows the distribution of correctly and incorrectly classified instances across the sentiment classes. The Naive Bayes algorithm demonstrates reasonable performance in predicting sentiment polarity, with notable accuracy in identifying positive sentiment and neutral sentiment instances, but relatively lower performance in identifying negative sentiment instances.

C. Random Forest

The Random Forest algorithm achieved an overall accuracy of 88 percentage in predicting sentiment polarity (positive, negative, or neutral) of the text data. Precision achieved a precision of 92 percentage for negative sentiment (class 0), 85 percentage for positive sentiment (class 1), and 92 percentage for neutral sentiment (class 2). The highest recall was observed for positive sentiment (class 1) at 99 percentage, indicating the model's ability to identify the majority of positive instances. The F1-scores provide a balance between precision and recall, with negative sentiment (class 0) achieving an F1-score of

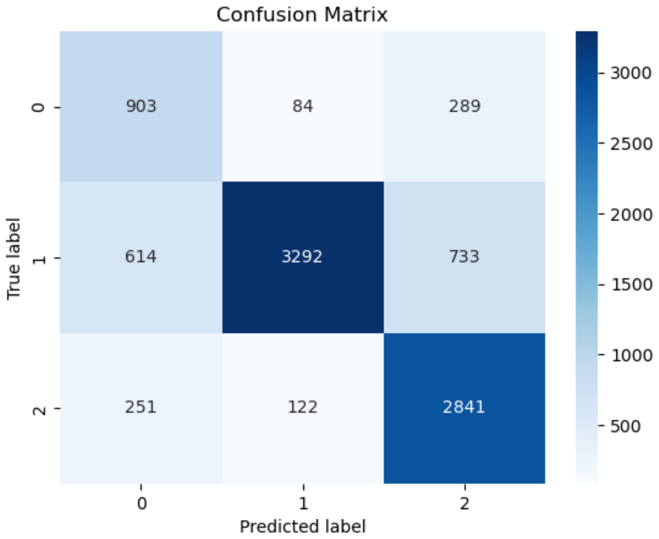

Fig. 3. Confusion Matrix of Naive Bayes

0.70, positive sentiment (class 1) achieving an F1-score of 0.91, and neutral sentiment (class 2) achieving an F1-score of 0.88. The confusion matrix from Fig. 4. illustrates the distribution of correctly and incorrectly classified instances across the sentiment classes. Overall, the Random Forest algorithm demonstrates good performance in predicting sentiment polarity, with notable accuracy in identifying positive sentiment and neutral sentiment instances, while achieving slightly lower precision and recall for negative sentiment instances.

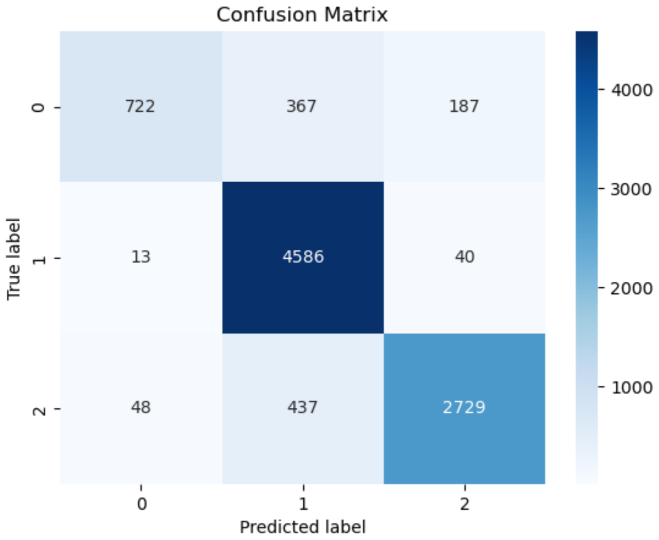

Fig. 4. Confusion Matrix of Random Forest

D. Logistic Regression

The Logistic Regression algorithm achieved an overall accuracy of 88 percentage in predicting sentiment polarity (positive, negative, or neutral) of the text data. Precision achieved a precision of 89 percentage for negative sentiment (class 0), 85 percentage for positive sentiment (class 1), and 93 percentage for neutral sentiment (class 2). The highest recall was observed for positive sentiment (class 1) at 99 percentage, indicating the model's ability to identify the majority of positive instances. The F1-scores provide a balance between precision and recall, with negative sentiment (class 0) achieving an F1-score of 0.67, positive sentiment (class 1) achieving an F1-score of 0.91, and neutral sentiment (class 2) achieving an F1-score of 0.89. The confusion matrix from Fig. 5. illustrates the distribution of correctly and incorrectly classified instances across the sentiment classes. Overall, the Logistic Regression algorithm demonstrates good performance in predicting sentiment polarity, with notable accuracy in identifying positive sentiment and neutral sentiment instances, while achieving slightly lower precision and recall for negative sentiment instances.

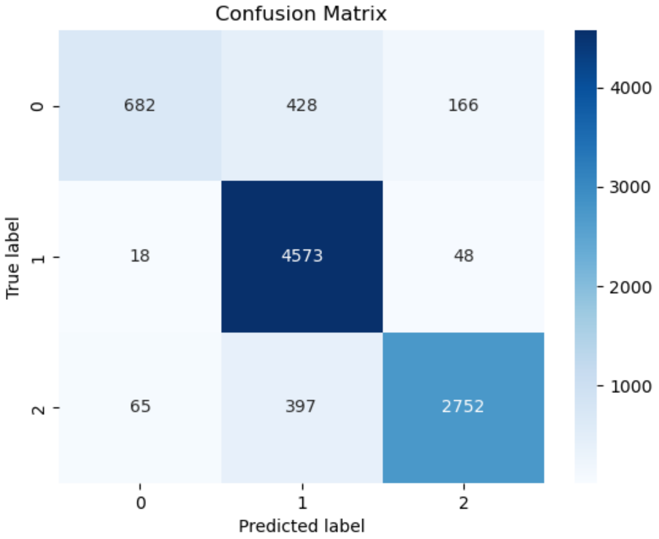

Fig. 5. Confusion Matrix of Logistic Regression

Overall, all the models had good accuracy scores and showed potential in identifying sentiment towards vaccinations on Twitter, but there was a need for improvement in identifying negative sentiment accurately.

VI. Discussion

This section aims to interpret and analyze the findings of the study on sentiment analysis of COVID-19 related tweets. The pre processing steps, including text normalization, removal of stop words, punctuation, URL's, numeric numbers, tokenization, stemming, and lemmatization, were successfully applied to the dataset. In addition to the findings mentioned above, we have included a new column called polarity in the dataset. This was achieved by using the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool from the Natural Language Toolkit (NLTK) library in Python. VADER is a rule based method that uses a lexicon of sentiment-bearing words and a set of rules to determine the sentiment of text. The polarity scores() function is used which returns a dictionary of sentiment scores for the text. We then used the polarity scores to create a sentiment column containing three classes: neutral, positive, and negative. To prepare the data for model training, we converted the sentiment column using LabelEncoder function. To gain more insights into the polarity of the tweets, we plotted a bar plot to show the count of tweets in each polarity class. The plot showed that negative sentiment was the most prevalent in the dataset, followed by neutral and positive sentiment.

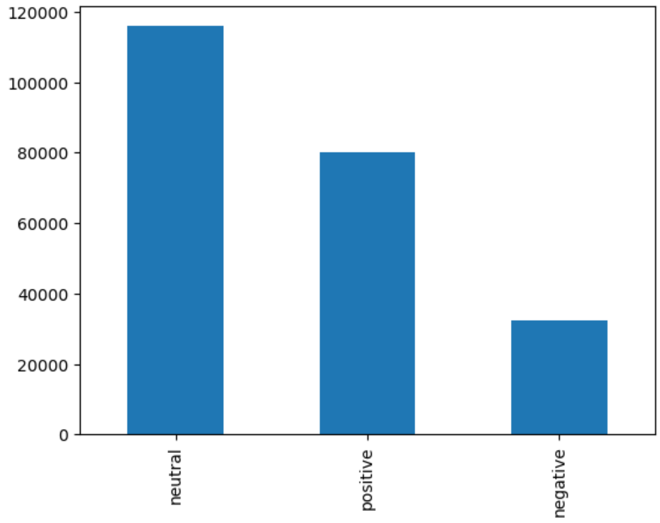

Fig. 6. Bar plot of Sentiment feature

The sentiment analysis was conducted using SVM, Naïve Bayes, Random Forest and Logistic Regression algorithms. The results showed that Random Forest achieved the highest accuracy of 88 percentage followed closely by Logistic Regression and SVM. The accuracies achieved are comparable to those reported in previous studies on sentiment analysis of COVID-19 related tweets. Through the study certain patterns were observed. Negative sentiment was more prevalent compared to positive sentiment in tweets which aligns with previous research. The findings of this study, highlights the potential of machine learning techniques in extracting valuable insights from social media during public health crisis.

The analysis was limited to tweets within a specific time frame, and is limited to a specific dataset. Future research based on the findings of this study can be expansion of the data on a larger and more diverse dataset of the related tweets and multilingual sentiment analysis, analyzing sentiment in different languages can provide insights into global sentiment patterns.

VII. Limitations

It is crucial to recognize the inherent limitations and constraints that may have influenced the findings. The study was conducted on a specific dataset of COVID-19 related tweets within a limited time frame. The sample size, although sufficient for the analysis conducted, may not fully capture the diverse range of sentiments expresses across different regions and languages. The accuracy of sentiment analysis heavily relies on the quality of the data. While efforts were made to preprocess the dataset and remove irrelevant information, some noises or biases may still be present. The reliability of sentiment analysis results can be affected by the presence of sarcasm or contextual nuances that are challenging to capture solely based on textual analysis.

This study focused on four popular classification algorithms: SVM, Naïve Bayes, Random Forest, and Logistic Regression. While these algorithms could potentially yield different outcomes. The choice of algorithms may have impacted the performance of the sentiment analysis. Despite the mentioned limitations, this study provides valuable insights into the sentiment analysis of COVID-19 related tweets.

VIII. Conclusion

The significance of this research lies in its contribution to the understanding of the sentiment dynamics during COVID-19 pandemic. By analyzing a specific dataset of tweets, valuable insights were gained regarding public sentiment. This project aimed to perform sentiment analysis of the tweets using various preprocessing techniques and machine learning algorithms. The application of SVM, Naïve Bayes, Random Forest, and Logistic Regression algorithms yielded promising results, with Random Forest achieving the highest accuracy of 88 percentage. These accuracies demonstrate the effectiveness of machine learning algorithms in sentiment analysis tasks related to COVID-19.

Future research should aim to address the limitations on this study by incorporating larger and more diverse datasets, considering contextual information, and exploring advanced techniques such as sentiment evolution analysis or multilingual analysis. The insights gained from this study have implications for crisis management, mental health support, public engagement strategies. Overall, this study contributes to the growing field of sentiment analysis and its relevance in understanding public sentiment during the COVID-19 pandemic. Further research in this area has the potential to advance or understanding of public sentiment dynamics and promote public well-being during times of crisis.

References

[1] R. Singh and P. Sharma, ”An Overview of Social Media and Sentiment Analysis,” 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-4, doi: 10.1109/ISCON52037.2021.9702359.

[2] Yuhao Xie, Tengjiang Wang, Hexuan Zhang, and Tianyuan Yan. 2022. Analyzing the Rate of Increase in Vaccines Administrated Versus Twitter Sentiment Analysis. In Proceedings of the 2022 International Conference on E-business and Mobile Commerce (ICEMC ’22). Association for Computing Machinery, New York, NY, USA, 77–82.

[3] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. ACM Trans. Manage. Inf. Syst. 9, 2, Article 5 (June 2018), 29 pages.

[4] Chandrasekaran R, Mehta V, Valkunde T, Moustakas E Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study J Med Internet Res 2020;22(10):e22624

[5] K. Shailaja, B. Seetharamulu and M. A. Jabbar, ”Machine Learning in Healthcare: A Review,” 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918..

[6] [6] F. Jemai, M. Hayouni and S. Baccar, ”Sentiment Analysis Using Machine Learning Algorithms,” 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 775-779, doi: 10.1109/IWCMC51323.2021.9498965.

[7] F. -J. Yang, ”An Implementation of Naive Bayes Classifier,” 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 301-306, doi: 10.1109/CSCI46756.2018.00065.

[8] Z. Bingzhen, Q. Xiaoming, Y. Hemeng and Z. Zhubo, ”A Random Forest Classification Model for Transmission Line Image Processing,” 2020 15th International Conference on Computer Science and Education (ICCSE), Delft, Netherlands, 2020, pp. 613-617, doi: 10.1109/ICCSE49874.2020.9201900.

[9] C. Zhang and W. Xu, ”Neural networks: Efficient implementations and applications,” 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, China, 2017, pp. 1029-1032, doi: 10.1109/ASICON.2017.8252654.

[10] Q. Wang, ”Support Vector Machine Algorithm in Machine Learning,” 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2022, pp. 750-756, doi: 10.1109/ICAICA54878.2022.9844516.