ANALYSING SENTIMENTS IN COVID-19 VACCINE TWEETS USING VARIOUS ML METHODS

Dr. Sumit Chopra[1], Er. Mukesh [2], Er. Gagandeep Singh Bains[3], Er. Mamta Bansal[4]

[1]*Associate Professor , GNA University Phagwara*

[2]*Student, GNA University Phagwara*

[3]*Assistant Professor , GNA University Phagwara*

[4]*Assistant Professor, GNA University Phagwara*

Abstract

Public opinion expressed on social media have a very significant impact on how people react to large-scale public health initiatives such as vaccination campaigns. During the COVID-19 pandemic, Twitter became a major platform where people openly shared opinions and experiences related to vaccines. This study analyses the sentiments in tweets collected from December 2020 to November 2021 related to COVID-19 to understand the changing opinions among public discussions on vaccination. Building on advanced Machine Learning methods, we gathered sizable amount of tweets dataset that consist of different phases within pandemic course as well as vaccination campaigns. Sentiment analysis in combination with multiple machine learning classifiers was used to determine the sentiments associated with tweets and performance of these models was evaluated using accuracy, precision, recall and F1-score. According to experimental results, the Decision Tree classifier performed more evenly across sentiment classes, especially when it came to differentiating between neutral and negative sentiments with overall accuracy (89.81%). This research paper has both theoretical importance and practical implications for the makers and health communication experts trying to improve vaccination acceptance strategies. With the world still fighting against COVID-19, understanding what people are saying about vaccines is important for improving vaccination campaign for COVID-19.

Keywords: Vaccine, COVID-19, Twitter, Sentiment Analysis, NLP, Emotion Detection, social media, Public Sentiment

1. INTRODUCTION

Currently, the rapid growth of social media has changed the way people express opinions, emotions and beliefs on public issues. Twitter has become a popular platform for real-time discussions. During the COVID-19 pandemic, Twitter became a critical medium to talk about health policies, preventive measures and most importantly, vaccination efforts. The large volume of tweets shared during this period provides a valuable opportunity to study public sentiment and understand people's response to

major health interventions.

In recent times, sentiment analysis has become a leading area for research interest (Manguri et al., 2020) and (Nezhad , 2022). It has powerful and careful evaluation of opinions, feelings and sentiments towards bodies such as brands, services issues events topics along with some special properties. The spectacular powers of Machine Learning have left their impacts in different fields and with outstanding effects. Natural Language Processing (NLP) is another such domain where Machine Learning has proved that it can surely achieve significant progress on performing complicated tasks (Khan, et al. , 2020) and (Braig et al., 2021).

In the last few years sentiment analysis has come up as a powerful and useful research tool not only for big organizations but also for small ones. In healthcare domain patient sentiment holds much importance. Every interaction, whether positive or negative, creates a response for the patient. Therefore, sentiment analysis assists people to find the patient's emotion for the healthcare organizations. Fig. 1 and Fig. 2 shows the sample of positive and negative tweets analyzed respectively.

Sentiments analysis of patients can give healthcare professionals information which will be helpful in eliminating the gap that exists between patients and medical institutions. This useful tool allows them to raise the patient`s satisfaction level and the world` s organization`s results. Sentiment analysis fetches data ranging from different sites through the World Wide Web (WWW) as well as social media sites such that we can acquire information which can give us public sentiments throughout that field (Garcia et al. , 2021) and (Villavicencio et al. , 2021)

```
work skilled nursing facility got first vaccine far less painful flu vaccin

200321 752308 vaccinations new daily record daily vaccination numbers rolling 7 day average

ukgovernment cant even vaccinate properly ethically get right kinds wrong

got first dose less waiting time airport vaccination center pain side effects

nas_k27 second dose due end next month well far pfizerbiontech
```

Fig. 1 Sample of Positive tweets analyzed

```
work skilled nursing facility got first vaccine far less painful flu vaccin

200321 752308 vaccinations new daily record daily vaccination numbers rolling 7 day average

ukgovernment cant even vaccinate properly ethically get right kinds wrong

got first dose less waiting time airport vaccination center pain side effects

nas_k27 second dose due end next month well far pfizerbiontech
```

Fig. 2 Sample of Negative tweets analyzed

This research paper is structured to guide the reader through the comprehensive process of analyzing sentiment in Twitter data, with each section contributing to the overall understanding and findings of the study. We begin with an induction to the significance of sentiment analysis in social media, particularly Twitter (Alshamrani, et al., 2020), which serves as a popular platform where millions of users express their opinions on a wide array of topics daily. Recognizing the value of this data, we first conducted a detailed review of existing literature on sentiment analysis of Twitter sentiment. This review allowed us to explore various methodologies previously employed, from traditional approaches to advanced machine learning and natural language processing (NLP) approaches (Mahalakshmi et al., 2024), (Madanian et al. , 2021) and (Shamrat et al. , 2021). Through this exploration, we identified key trends and gaps in the research, which helped us refine our study's focus. Our research aims to compare different sentiment analysis techniques and evaluate the performance of various classifiers on a large Twitter dataset. Data retrieval played a crucial role in this study, as the accuracy and relevance of our findings heavily rely on the quality of the data collected. We used different methods to gather a comprehensive and representative sample of tweets, ensuring that our dataset was diverse enough to capture the broad spectrum of public sentiment on the topics of interest. The data was then preprocessed to remove noise, such as irrelevant content, which could potentially distort the analysis (Sharma, S. et al, 2020) and (Alabid et al. , 2021). This preprocessing step included tokenization, stemming, and the removing stop words, preparing the data for more accurate sentiment analysis. In handling the large Twitter dataset, we utilized a combination of powerful tools and techniques, including open-source libraries and custom algorithms designed to manage and analyze large volumes of data efficiently. The results of our analysis are presented in the 'Results and Discussions' section, where we examine the outcomes of applying various classifiers to the data. This section provides a detailed comparison of the performance of different sentiment analysis techniques, highlighting their strengths and limitations in identifying sentiment in tweets. Finally, the 'Conclusion' section summarizes our findings, discussing the implications of our research and suggesting potential directions for future studies. This study not only contributes to the ongoing research on sentiment analysis in social media but also provides practical insights into the effectiveness of different approaches in analyzing large datasets, which could be used by researchers and practitioners in fields such as marketing, political science, and public opinion research.

1.1 KEY CONTRIBUTIONS

- The key contributions of this study are summarized as follows:
- A comprehensive sentiment analysis of COVID-19 vaccine–related Twitter data that covers several stages of the vaccination rollout.
- A systematic comparison of five classical machine learning classifiers using a single

experimental framework.

- An empirical analysis demonstrates the effectiveness of Decision Tree-based models in handling sentiment imbalance.
- Useful insights into classification issues that can inform future sentiment analysis research and public health communication strategies.

2. RELATED WORKS

Advanced Data Mining includes Sentiment Analysis which is constantly in development. This interesting topic of research is about automatic text analysis, dividing the large webwork between views and feelings. Another thing worth noting is that, through this massive work, the focus has been on examining the medical posts in the blogs to get additional knowledge on this subject. Fundamentally, the older studies related to sentiment analysis also encourage other researchers for more studies in this very exciting area.

Several studies have explored techniques of sentiment analysis related to COVID-19 and Twitter data, each contributing unique approaches to understanding public sentiment. Chakraborty et al. (2020) checked tweets related to COVID-19 and the World Health Organization (WHO), focusing on 23,000 retweets posted. Their study aimed to predict sentiments through retweets and utilized tools like Word Clouds and word frequency analysis, though it fell short of extracting useful keywords and phrases for deeper insights. (Chen et al.,2020) introduced a novel technique leveraging Twitter's embedded emotions, employing a two-step process where they trained a model to learn dual embeddings for positive and negative sentiments. This method, using a long-short-term memory (LSTM) network, effectively captured emotions conveyed by emoticons, outperforming traditional sentiment analysis methods. (Reddy et al. ,2018) adopted a different approach by classifying tweets into positive and negative sentiments through structured preprocessing and the use of CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), ANN (Artificial Neural Network) models. They introduced new concepts in building sentiment lexicons tailored for sentiment analysis, providing insights into sentiments within the Twitter domain. Together, these studies illustrate the evolving methodologies in sentiment analysis, specifically in the context of COVID-19.

In related research (Hasan et al. ,2018) developed a method that merges sentiment analysis with machine learning techniques such as Support Vector Machines (SVMs) and Naive Bayes to investigate political sentiments. Their approach provided valuable insights into the connection between sentiment and political opinions. (Sanders et al., 2021) conducted a large-scale analysis of over a million COVID-19-related tweets, focusing on public sentiment toward mask usage. Utilizing advanced Natural Language Processing (NLP) tools, their study revealed significant patterns in public opinion, offering a robust framework for understanding societal responses to health measures.

Similarly, (Neethu and Rajasree et al.,2013) explored sentiment analysis on Twitter, employing machine learning techniques to classify tweets into negative, positive and neutral categories. Their research emphasized the importance of feature selection and model accuracy, particularly in the context of analyzing COVID-19 vaccination sentiments. Collectively, these studies underscore the crcuial role of sentiment analysis in understanding public opinion during significant global events, providing methodologies that can be adapted to various domains.

(Gupta et al., 2017) explored Twitter sentiment analysis using various machine learning algorithms in Python, emphasizing the importance of preprocessing and feature selection in enhancing sentiment classification accuracy. This work is highly relevant for analyzing COVID-19 vaccination tweets, as the methodologies can be adapted to track public sentiment trends during the pandemic. Similarly, (Dubey et al. ,2020) conducted sentiment analysis on tweets related with COVID-19, revealing shifts in public emotions corresponding to key events, highlighting the social media role in shaping public opinion during crises [8]. Boon-Itt and Skunkan et al., 2020) further contributed by applying sentiment analysis and topic modeling to Twitter data, examining public perceptions of the COVID-19 pandemic. Their study demonstrated the effectiveness of combining these techniques to gain comprehensive insights into public discourse, particularly for understanding the complex sentiments surrounding COVID-19 vaccinations. Together, these studies underscore sentiment analysis role in monitoring public opinion during significant global events.

(Naseem et al.,2021) introduced COVIDSenti, a comprehensive Twitter dataset designed to enhance COVID-19 related tweets sentiment analysis. This large-scale dataset provides a benchmark for evaluating sentiment analysis models, crucial for understanding public sentiment throughout the pandemic, particularly regarding vaccine attitudes. Building on this foundation, (Alanezi & Hewahi et al., 2020) analyzed tweets from the pandemic period to assess emotional responses, utilizing machine learning techniques to classify sentiments. Their research sheds light on how public sentiment evolved, especially as vaccination efforts began. Complementing these studies, (Jabalameli, et al. ,2022) performed a spatial and sentiment analysis to examine regional variations in public opinion during the early vaccination stages. By integrating geographic and sentiment data, their work underscores the influence of spatial factors on sentiment, offering insights into regional disparities in vaccine uptake and public opinion. Together, these studies provide a nuanced understanding of sentiment trends and regional differences during the pandemic.

(Barkur & Kamath et al. ,2020) investigated the sentiment of the Indian public in response to the nationwide COVID-19 lockdown, revealing the emotional and psychological effects of lockdown measures. Their findings are instrumental in understanding how similar interventions might influence public sentiment during vaccine rollouts, highlighting the importance of timely sentiment analysis for

shaping public health responses . In a different context, (Heras-Pedrosa et al. 2020) conducted sentiment and emotion analysis in Spain, focusing on the emotional drivers behind public sentiment and their impact on digital discourse. Their research offers insights into the emotional factors influencing attitudes toward COVID-19 vaccination within different cultural settings. Complementing these studies, (Yeasmin et al.,2022) developed predictive models to analyse and forecast user sentiment about the pandemic, providing a dynamic view of public opinion over time. This approach is crucial for anticipating shifts in sentiment, particularly during significant events like vaccine introductions, and informing effective public health strategies.

The social media platform has become a place for discussions and updates about the COVID-19 vaccination, with different voices sharing their different ideas and feelings. These different social media platforms help in the exchange of views about the vaccines during the pandemic related works. The amount of data produced in these online platforms helped the computer scientists and researchers, providing a very good chance to study people opinion for COVID-19 vaccines. This analysis of emotions helps to get important insights. This comprehensive study examines the responses by COVID-19 vaccines using sentiment analysis. This article begins this important journey by using tweets related to COVID-19 vaccines to find public response on the vaccine campaigns. This research includes experimental studies for analysis of emotions and brings insights on vaccine related data. Essentially, this research uses social media users to find the different emotions, demographics and thoughts that are related to COVID-19 vaccines.

3. DATA RETRIEVAL AND METHODOLOGY

3.1 OVERVIEW AND DATA COLLECTION

The various steps used in the proposed work are summarized in Fig. 3 which provides a visual representation of our experiment's architectural layout. The Twitter Application Programming Interface (API) was utilized to generate tweets that were accessible to the public for our study. Preprocessing step is used for dataset filtering, converting emoticons into text representation, and removing retweets and URLs. In order to improve data quality, stop words are removed and tokenization, stemming, and other methods are used for additional filtering. The computation of word frequency was part of the dataset exploration process.

Following the preprocessing stage, we moved on to the data's sentiment analysis. To achieve this, we used 'TextBlob' along with a variety of techniques. The `TextBlob` library is a user-friendly and powerful tool for natural language processing (NLP). Built on Pattern libraries and NLTK, it simplifies complex tasks like sentiment analysis, part-of-speech tagging, noun phrase extraction, and translation. Its intuitive interface allows researchers to quickly analyze text, determine sentiment, and understand grammatical structures with minimal code. `TextBlob` also supports seamless translation

using the Google Translate API, making it ideal for multilingual data analysis. Overall, `TextBlob` offers a comprehensive solution that streamlines text processing, making it accessible and efficient for both researchers and developers. Vectorization was applied to the data by means of 'CountVectorizers'. 'CountVectorizer' is a fundamental tool in natural language processing (NLP) used to convert a collection of text documents into a matrix of token counts. It transforms each document into a vector where each element represents the frequency of a word (or token) in the document, capturing the raw occurrence of terms without considering their order. This technique is widely used in text classification, clustering, and information retrieval.
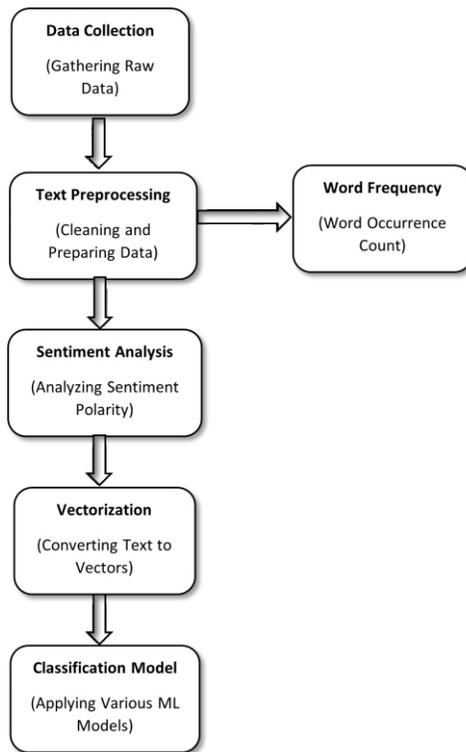

Fig. 3. Experimental Design

By converting textual data into numerical form, 'CountVectorizer' enables machine learning algorithms to process and analyze textual information, making it an essential preprocessing step in many Natural Language Processing tasks. In the last stage of our study, we created and assessed various models to divide the COVID-19 Twitter dataset into three groups: neutral, negative, and positive. Several machine learning techniques, such as LinearSVC, Logistic Regression, Multinomial Naive Bayes, Random Forest, and Decision Tree, were used in the construction of these models.

3.2. TEXT PREPROCESSING

In sentiment analysis, text preprocessing is a critical step that plays a vital role in ensuring accurate

and reliable outcomes. The raw text data gathered from various sources, such as social media, reviews, or feedback forms, is often cluttered with irrelevant information like URLs, retweets, and unnecessary punctuation. These elements can interfere with the sentiment analysis process, leading to misleading results. By meticulously cleaning the data, we can focus on the actual content and emotions expressed in the text. For instance, emojis, which are frequently used to convey emotions, are converted into corresponding words so that the sentiment behind them can be properly captured by the machine learning algorithms. This transformation process not only makes the data more structured but also ensures that the algorithms are analyzing the sentiment in a meaningful way.

Tokenization is another crucial aspect of text preprocessing, where the text is broken down into individual words or tokens. This allows the algorithms to understand the text at a granular level. Following this, stop word removal is performed to strip away common words that do not contribute to the sentiment, such as "and," "the," or "is." This leaves only the most relevant words, making it easier to detect the sentiment being expressed. Stemming further refines the data by reducing words to their base forms, ensuring that different variations of the same word are treated as one. The end result is a well-organized dataset, free from noise and redundancy, that can be effectively utilized in sentiment analysis. This thorough preprocessing ensures that the insights drawn from the analysis are both accurate and actionable, allowing businesses to understand and respond to customer sentiments with greater precision. Fig. 4 give a visual representation of the various techniques used for Preprocessing of tweets.
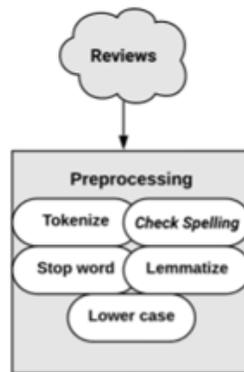

Fig. 4 Preprocessing of Tweets

3.2.1. RETWEETS

In Twitter, a retweet may also be considered as releasing a tweet from an outburst of the brain of another user. If everyone uses the same kinds of words in their tweets, it can harm the natural patterns of language. This makes it harder for computer models to learn and work correctly. We need more

diverse tweets to help these models improve. On the other hand, this method states that its use cases include the cutting of re-tweets which helps in the generation of pure data.

A retweet on Twitter is when a user shares a tweet that has been shared by another user. This could lead to redundant data. These same tweets have a chance to change the initial word frequency value, interfere with our models' smooth operation, and increase the experiment's processing requirements. Therefore, this method requires that retweets be filtered in order to maintain the integrity of our data.

3.2.2. URLS AND PUNCTUATION

This is the next phase of our analysis that will require careful extraction of URLs from the data set. These can introduce noise to the data. Their presence in this case can lead to overestimation. Consequently, a key step in maintaining data integrity and obtaining accurate results is getting rid of URLs. Python's regular expression library was then used to systematically identify and remove URLS as well as other unwanted symbols such as hashtags, question marks, punctuation marks etc. This cleanup procedure is necessary before any analysis could be done on the text data. By doing this we create a good base for generating valuable insights from these tweets.

3.2.3. TOKENIZATION AND NORMALIZATION

In the next phase of our analysis, the raw text in each tweet was broken down into individual words by a process called tokenization. As such, techniques such as text normalization were concurrently applied to enhance efficiency and reduce computational overhead. Normalization makes words simpler. It changes words like "loved", "loving", and "love" to a basic form like "love". This helps computers understand words better. This included converting everything into lowercase thus standardizing word forms and eliminating unnecessary variations. By combining tokenization with normalization, the dataset was prepared for further analysis while at the same time optimizing processing time. This was followed by allocating each of these resulting tokens to one of three sentiment polarities; positive, negative or neutral. Because words are categorized based on their emotional implications which help to determine the overall sentiment presented within a tweet. In this way, sentiment analysis has been built up through a detailed procedure that provides a formal pattern of recorded data for future work with textuality.

3.2.4. STOP WORDS

In our Twitter sentiment analysis, one of the main things we did was to remove stop words once we had identified relevant terms in the tweets. Stop words are common words like 'to,' 'was,' 'wasn't' and 'weren't', which are necessary for constructing sentences but basically do not have much meaning in the context of sentiment analysis. Such words can introduce noise that may mask more important words bearing the true sentiment of a tweet. This enabled us to narrow down our focus by filtering out these stop words, thus enabling an analysis based on those words that reveal emotions or opinions as

shared on Twitter. To achieve this, we utilized 'stopword' removal feature from 'sklearn' library, specifically using tools such as 'CountVectorizer' or 'TfidfVectorizer'. 'sklearn'is a powerful Python library for machine learning, including sentiment analysis. It provides tools for text preprocessing (like tokenization, stop word removal, and feature extraction), model training (using algorithms like Naive Bayes, Support Vector Machines, or Random Forest), and evaluation. 'sklearn' provides a built-in list of common English stopwords that can be easily removed from our text data. Our data has been cleaner and more focused because of this preprocessing step, thereby enabling the sentiment analysis model to better understand and interpret what people truly feel towards their tweets better leading to more meaningful insights with high accuracy.

3.2.5. STEMMING

Stemming is a process of reducing words to their root or stem form as a text preprocessing technique. For instance, "looked, looking and looks" can be reduced to look as a stem. We employed the 'PorterStemmer' algorithm to reduce word forms in our study. The 'PorterStemmer' is one of the most widely used algorithms for stemming English words. It works on a set of rules that remove suffixes common with English language words. It cannot be considered perfect, but it approximates word stems well enough thus allowing us to reduce the dimensionality of our text data while keeping critical semantic information unharmed for sentiment analysis.

3.3. DAILY SENTIMENT ANALYSIS

An important area of Natural Language Processing (NLP) is text sentiment analysis. It transforms unstructured data into a structured range of statistical opinions, creating a tool that can be trained to extract insightful information from written content. This simple method creates an emotion score by counting the number of positive, negative, and neutral words in each page. Positive words add one point to the score in this numerical process, whereas negative words have the opposite effect and reduce one point. Neutral words have no effect on the sentiment score and provide the foundation for the numerical depiction of textual feelings. Text sentiment analysis uncovers the underlying emotions in words in this way.

$$\text{Sentiment} = (P - N) / (P + N - O \qquad\qquad (1)$$

$$\text{Sentiment} = (P - N) / (P + N) \qquad\qquad (2)$$

$$\text{Sentiment} = \log(P + 0.5) - \log(N + 0.5) \qquad\qquad (3)$$

For Equation (1), the absolute proportional difference bounds between 0 and 1, Relative Proportional Difference bounds between 0 to 1 in case of Equation (2) and Range of Logit scale bounds from $-\infty$ to $+\infty$ as represented in case of Equation (3). In sentiment analysis, we give the words distinct roles: 'P' stands for positivity, 'N' for negativity, and 'O' for neutral, reflecting the range of human emotions. With these designations, the numerical representation of the polarity of human emotions in our

sentiment scoring method stays between a score of -1 and +1. Using a combination of sentiment computation techniques, including the features of TextBlob and machine learning, this research study seeks to uncover the range of human sentiment hidden in the dataset.

3.3.1. TEXTBLOB

In our analytical study, we used the Python module TextBlob() to determine the range of sentiment seen in individual COVID-19 vaccine-related tweets. With the use of this tool, we were able to generate TextBlob objects, each of which may receive polarity values from tweets. We were able to determine the emotional subtext of each tweet by looking through the properties of these items, particularly the polarity feature. With the use of pandas dataframe, we gathered a collection of tweets, each with a unique sentiment score. Positive values in sentiment indicate positivity, and negative values indicate the tones of negativity. In this sentiment analysis study, neutrality among the range of emotions was indicated by polarity values near zero.

3.4. VECTORIZATIONS

In sentiment analysis, vectorization is the process of converting words into numerical vectors, enabling machine learning algorithms to process and analyze text data. This transformation is essential as it translates the unstructured language into a structured format that algorithms can understand. Techniques like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used for this purpose (Bania , 2020). By representing words as vectors, these methods help the model recognize patterns and associations between words and sentiments, allowing for more accurate sentiment classification and insight extraction from the text data. The Sentiment Distribution and Sentiment Count of Vaccination Tweets from the time period of December 2020 to November 2021 have been represented in the terms of graphs as shown in Fig. 5 and Fig. 6 .
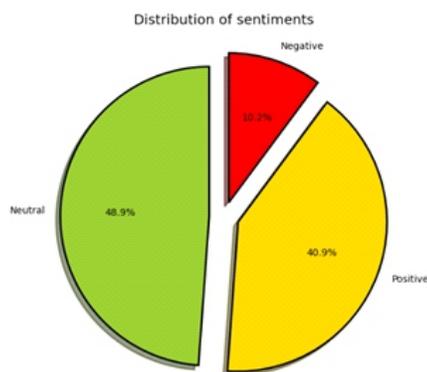

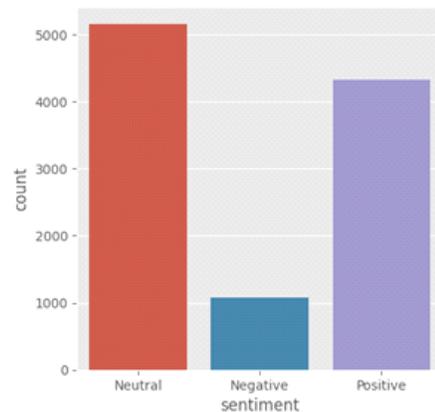Fig. 5. Sentiment Distribution of Vaccination Tweets (December 2020 - November 2021)

Fig. 6 Sentiment Counts of Vaccination Tweets December 2020 - November 2021

3.4.1. COUNT VECTORIZER

A collection of texts can be converted into an organized vector format with the help of the effective system 'CountVectorizer'. 'CountVectorizer' is a tool in sentiment analysis that transforms text data into a matrix of word counts. Each word in the text is represented by a column, and each row corresponds to a document, like a tweet. This numerical representation helps machine learning models identify patterns and associations in word usage related to sentiment. This facilitates the creation of the matrix form, which is used to quantify and record the text's tokens as counts (Sivanantham et al., 2021). The textual data first goes through a preprocessing step that uses tokenization to divide the data before this process (Kausar et al., 2021). In the resulting matrix, every column contains a distinct word. This word count example demonstrates how data is represented as a matrix:

df = ['We', 'red', 'radius', 'offers', 'honor', 'story', 'open', 'COVID′, 'vaccination'].

3.5. CLASSIFICATION MODELS

Within the field of predictive modeling, classification models serve as effective instruments for classifying input data, hence providing a sense of order and structure to the available information (Vijay et al. , 2020) and (Alanezi et al., 2020). In order to perform this classification, we need a training dataset; a reliable training dataset is one that contains a sufficient number of examples for each class label. In order to assess the performance of these models, we split our data into two subsets: a training set that makes up 80% of the data, and a testing set that makes up the remaining 20%. This allowed us to calculate the model's efficiency. Metrics such as accuracy, precision, recall, and F1 score are used in the testing of mathematical problems. The accuracy of a model is determined by dividing the total number of observations by the ratio of correctly predicted results as represented in Equation 4. On the other hand, precision, which is defined as the ratio of accurately predicted positive instances to all expected positive and same is represented in the Equation 5. It indicates how accurate the model is in predicting positive outcomes. Recall as expressed in Equation 6, which is often referred to as sensitivity, measures the model's capacity to recognize real positive examples throughout the entire class. At last, the F1 score provides a weighted average that discusses the overall performance of the model by working as a harmonic combination of Precision and Recall. The same has been represented in the Equation 7.

$$\text{Accuracy.} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FP}$$

$$\text{F1 Score} = \frac{2 \; X \; (Recall \; X \; Precision)}{Recall+Precision}$$

The outcomes of the model evaluation process are displayed using acronyms that carry specific meanings. True Positives, or TP for short, indicate the times when our model has correctly predicted positive results. On the other hand, TN, or True Negatives, stands for those instances in which the model correctly identified and precisely predicted the negative values. False Positives, or FPs, are the opposite of these real outcomes and occur when our model predicts positive values wrongly. False Negatives, or FN for short, are the times where our model erred and predicted negative results (Ahmed et al., 2021) and (Boon-Itt, et al., 2020) . The concept of model performance is defined by these elements used collectively.

3.5.1. RANDOM FOREST

Random Forest (RF) helps in handling both regression and classification problems. Regression and classification are supervised machine learning techniques used for predicting values. Regression predicts a continuous numerical value, like house prices based on features, while classification predicts a category or label, such as categorizing emails as spam or not spam. The key difference lies in the nature of the output: numerical for regression, categorical for classification. It combines a collection of individuals, termed as decision trees, to make highly accurate predictions. This whole approach is further done by random feature selection (Li, Alam, & Melnokov et al., 2021). The equation of Random Forest used for data analysis and prediction is shown in Eq. 8 and Eq.9 respectively.

$$p = \text{mode}\{T_1(y), T_2(y), ..., T_m(y)\}$$

$$p = \text{mode}\{\Sigma T_m(y)\}$$

According to Khan et al. (2021), the final prediction 'p' in Random Forest is the result of a collective choice made by the number of decision trees, each represented by the letters $T_1(y)$, $T_2(y)$, $T_3(y)$, and up to $T_m(y)$. The classification technique was created by combining these decision trees. We chose a random forest with a count of 100 trees for our experiment. Each of these trees has its maximum depth of 5 cuts, creating a balance between complexity and performance. The correctness of our model was ensured by leaving all other hyperparameters like min_samples_split, min_samples_leaf, max_features, and criterion in the sklearn library at their default values.

3.5.2. DECISION TREE

A decision-making and categorization method with a branching tree-like form is called a decision tree. Every branch that emerges from a node in this structure discusses possible values, and every line in the tree corresponds to an important attribute test. As we've implemented our Decision Tree model, we've come across a crucial line of code that drives this strong algorithm within the sklearn framework: the definition of the minimal sample split.

Decision Tree': {'min_samples_split': [1, 3, 5]}.

3.5.3. LINEARSVC

As a support vector classification technique, linear SVC (Linear Support Vector Classification) is known for its effectiveness in both sparse and dense inputs. 'LinearSVC' is a supervised machine learning algorithm used for classification tasks. It seeks to find the optimal hyperplane that separates data points into different classes. Unlike its non-linear counterpart, 'LinearSVC' assumes a linear decision boundary. Users are now able to accurately fine-tune their models. The convergence speed of linear SVC makes it a suitable option for efficient models, particularly when applied to big datasets. We employed the sklearn library's default hyperparameter values for our research as a systematic methodology to guarantee the accuracy and dependability of our experiments. While Linear SVC doesn't have a direct equation like linear regression, it aims to maximize the margin between data points of different classes. The decision function is essentially a linear combination of features as represented in Eq. 10.

$$f(x) = w * x + b \qquad\qquad (10)$$

where 'f(x)' is the decision function output, 'w' is the weight vector, 'x' is the feature vector, 'b' is the bias term. The goal is to find w and b that correctly classify data points while maximizing the margin.

3.5.4. LOGISTIC REGRESSION

A powerful approach for determining the response variable that has a linear equation linked to exploratory factors is logistic regression. Logistic regression is a statistical method used in sentiment analysis to classify text as positive or negative as shown in Eq.11. It estimates the probability of a text belonging to a positive sentiment class based on features like word occurrences and their associated weights, providing a probabilistic measure of sentiment. The equation becomes nonlinear, resulting in the logistic function or S-shaped logistical curve. The addition of nonlinearity introduces an additional level of complexity, enabling Logistic Regression to accurately carry out the classification process.

$$I = \log_b(p / 1 - p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \qquad\qquad (11)$$

A logistic regression equation in sentiment analysis combines coefficients representing the impact of different words or phrases on the likelihood of a text expressing a positive sentiment. The intercept coefficient (β_0) represents the baseline probability of a positive sentiment, while coefficients for specific words (β_1, β_2, ..., β_n) quantify their contribution to shifting this probability. Together, these coefficients form a model that predicts the sentiment polarity of a given text.

3.5.5. MULTINOMIAL NAÏVE BAYES

The Naïve Bayes approach, which operates on the concept of Bayes' theorem, is a classification model

in general. It makes the assumption that each variable is distinct from the others and independent. It functions on the naive assumption that features belonging to one class have no bearing on those of other classes. A Naïve Bayes classifier, which serves as the basis for Bayesian network models, reflects this simplicity. The name "Probabilistic Classifier" refers to this model, which is based on the independent distribution probabilities found in Bayes' theorem. This methodology uses the capabilities of this classification model to ensure the integrity of our analysis as given in the Eq. 12.

$$P(C|D) = (P(D|C) * P(C)) / P(D) \qquad (12)$$

Where P(C|D) represents the probability of a text belonging to a sentiment class (C, such as positive or negative) given its content (D). P(D|C) calculates the likelihood of observing the specific words in a text given its sentiment class. P(C) is the prior probability of a sentiment class occurring independently of the text, and P(D) is the probability of observing the given text, which is often treated as a constant. The "naive" part comes from the assumption of independence between words, which simplifies the calculations.

4. RESULTS AND DISCUSSION

4.1 RESULTS

Various classification models, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB) classifiers, were used to do sentiment analysis of tweets related to the COVID vaccination. A diverse range of tweets expressing views and opinions regarding the COVID-19 vaccine are included in the dataset.

4.1.1 DATA PREPROCESSING

Tokenization, stop word removal, and stemming were among the preprocessing processes performed on the dataset prior to the model training process. This made it easier to make sure the text data was formatted properly for the machine learning models.

4.1.2 MODEL TRAINING

With the goal of identifying differences in the opinions stated in the tweets, each classification model was trained on a subset, or training data, of the dataset. To maximize and enhance each model's performance, hyperparameter tuning and feature extraction were the two main components of the training process.

4.1.3 EVALUATION METRICS

A number of general sentiment analysis criteria, such as accuracy, precision, recall, and F1 score, were used to assess these models' performance. These metrics were assessed in order to give a thorough knowledge of the models' performance for each sentiment class (positive, negative, and neutral).

4.1.4 MODEL PERFORMANCE

Each categorization model performs differently, which naturally reflects the difficulties in extracting feelings from unstructured data. While some models did well in capturing negative attitudes, others demonstrated superior accuracy in classifying positive thoughts. The models with random forests, decision trees, and logistic regression performed well and had higher accuracy on a variety of requirements. The results obtained from the sentiment analysis using various machine learning models provide valuable insights into the performance of these classifiers in identifying sentiments within the dataset. The accuracy, recall, and precision metrics of each model were evaluated to determine their effectiveness in classifying sentiments as positive, negative, or neutral.

Word clouds were generated to visualize the most frequently used words in the tweets, categorized by sentiment. The word cloud for negative tweets as shown in Fig. 5 highlighted terms such as "arm," "little," "death," and "side effects" as the most common words. For neutral tweets words like "dose," "safety," "Pfizer vaccine," and "observe" were predominant as shown in Fig. 6. In contrast, the word cloud for positive tweets as shown in Fig. 7 showed that words such as "happy," "vaccinated," "first," "Covid-19," and "Pfizer-BioNTech" were frequently used.

Fig. 7 Wordcloud for Negative tweets   Fig. 8 Wordcloud for Neutral Tweets

Fig. 9 Wordcloud for Positive tweets

The analysis as summarized in Table 1revealed that the Decision Tree (DT) model outperformed other classifiers, achieving the highest accuracy of 89.81%. This indicates that the DT model was the most effective in accurately categorizing the sentiments expressed in the tweets. The Random Forest (RF) model and Logistic Regression (LR) followed, with accuracies of 84.73% and 84.64%, respectively. These models also demonstrated strong performance but were slightly less effective than DT in terms of accuracy. Despite the good accuracy shown by Support Vector Machine (SVM) and Naive Bayes (NB), these models exhibited lower precision and recall values, particularly for negative tweets. This suggests that they were biased towards positive sentiments and, therefore, less reliable for classifying negative sentiments. As a result, these models were considered less suitable for sentiment classification in this context.

Table 1. Results of Evaluation metric performance of different classifiers on dataset

Evaluation Metric	Classification Model	Sentiment Score			Accuracy
		Negative	Neutral	Positive	
Precision	LR	0.86	0.79	0.94	84.64%
	SVM	0.89	0.70	0.92	76.58%
	RF	0.90	0.80	0.93	84.73%
	DT	0.78	0.88	0.94	89.81%
	NB	0.38	0.84	0.73	75.68%
Recall	LR	0.32	0.99	0.82	84.64%
	SVM	0.04	0.99	0.69	76.58%
	RF	0.27	1.00	0.82	84.73%
	DT	0.57	0.77	0.88	89.81%
	NB	0.23	0.77	0.88	75.68%
Recall	LR	0.46	0.88	0.87	84.64%
	SVM	0.07	0.82	0.79	76.58%
	RF	0.41	0.88	0.87	84.73%
	DT	0.65	0.93	0.92	89.81%
	NB	0.29	0.81	0.80	75.68%

From the clustered column chart shown in the Fig.10, it's evident that Decision Tree (DT) consistently performs well across all metrics and sentiment categories, with particularly high scores in Neutral and Positive sentiment classifications. This model achieves the highest overall accuracy, making it a robust choice for sentiment analysis. Random Forest (RF) also shows strong performance, particularly in the Precision metric, with balanced scores across sentiments. However, its Recall is notably lower for the Negative sentiment. On the other hand, Naive Bayes (NB) performs poorly in the

Negative sentiment classification across all metrics, indicating its limitation in correctly identifying negative sentiments. SVM, despite a high Neutral recall, struggles with the Negative sentiment, leading to lower overall effectiveness compared to other models. Logistic Regression (LR) provides a balanced performance but doesn't excel in any metric, making it a reliable yet unremarkable option for sentiment classification.
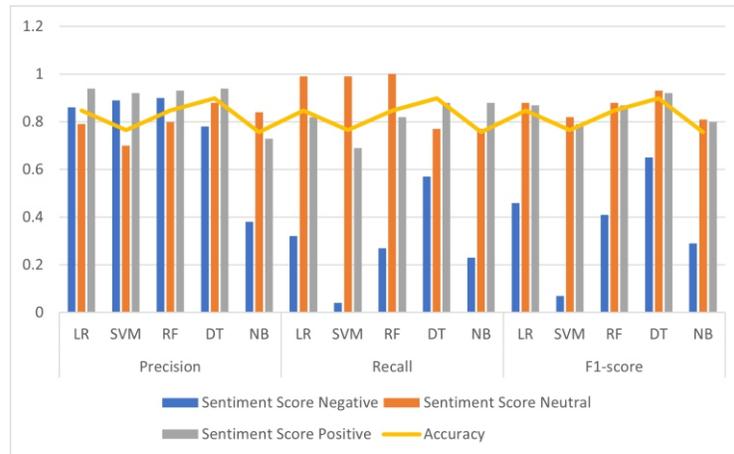

Fig. 10. Clustered column chart for Evaluation metric performance of different classifiers on dataset

The confusion matrix results further supported these findings by illustrating the relationship between actual and predicted sentiments. It was observed that 48.9% of the tweets were classified as neutral, 40.9% as positive, and only 10.2% as negative. This distribution reflects the overall sentiment trends within the dataset and highlights the effectiveness of the DT model in accurately identifying these sentiments.

In conclusion, the Decision Tree model was selected as the most suitable classifier due to its superior accuracy, precision, and recall in categorizing tweets into positive, negative, and neutral sentiments. Other models, while showing good overall accuracy, were less reliable in distinguishing between different sentiment classes, particularly for negative sentiments.
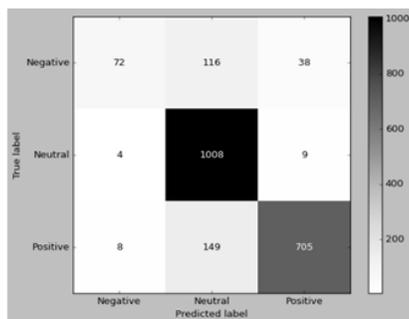

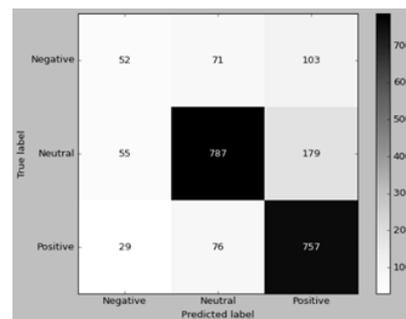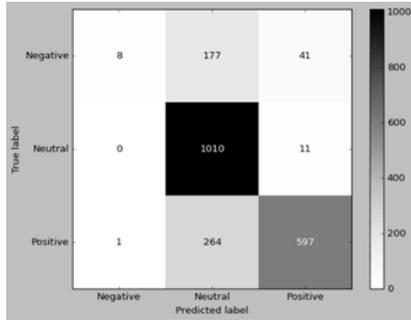Fig. 11 Confusion matrix for LR

Fig. 12 Confusion matrix for NB

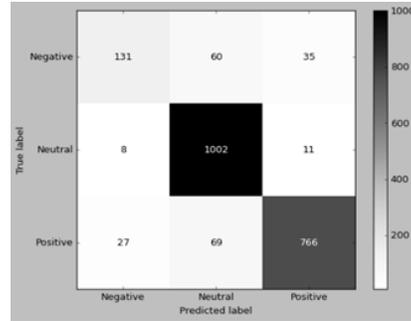Fig. 13 Confusion matrix for SVM
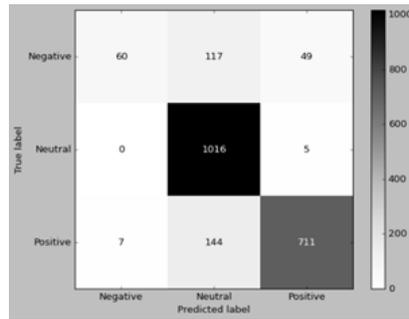

Fig. 14 Confusion matrix for DT

Fig. 15 Confusion matrix for RF

5. CONCLUSION AND FUTURE SCOPE

In summary, the usefulness of various machine learning techniques in understanding public sentiment was the primary focus of the sentiment analysis of COVID-19 vaccine tweets using the different classification models. The Random Forest and SVM models performed quite well in tweet classification, demonstrating their potential for use in various real-world sentiment analysis applications. According to our research, the overall performance of these various models could be enhanced by additional study and experimentation following hyperparameter tuning and collective methods.

The findings from this study provide crucial insights into public sentiment toward COVID-19 vaccination, offering a foundation for future research and public health initiatives. Understanding these sentiments allows policymakers and health officials to craft more effective, empathetic communication strategies that address public concerns, reduce vaccine hesitancy, and build trust. By combining sentiment analysis with vaccination rate data, researchers can explore how public attitudes directly impact the success of vaccination programs and broader public health outcomes. This approach can reveal patterns indicating how emotions influence people's willingness to get vaccinated, leading to more targeted and responsive public health campaigns. Ultimately, this study

demonstrates the value of sentiment analysis as a tool to enhance public health efforts, ensuring that vaccination strategies are not only scientifically robust but also resonate with the public's feelings and needs.

REFERENCES

Chakraborty, Indranil, and Prasenjit Maity. "COVID-19 outbreak: Migration, effects on society, global environment and prevention." Science of the total environment 728 (2020): 138882.

Chu, Derek K., et al. "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis." The lancet 395.10242 (2020): 1973-1987.

Reddy, D. Sushma, and P. Rama Chandra Prasad. "Prediction of vegetation dynamics using NDVI time series data and LSTM." Modeling Earth Systems and Environment 4 (2018): 409-419.

Hasan, Ali, et al. "Machine learning-based sentiment analysis for twitter accounts." Mathematical and computational applications 23.1 (2018): 11.

Sanders, Abraham C., et al. "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse." AMIA Summits on Translational Science Proceedings 2021 (2021): 555.

Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT) (pp. 1-5). IEEE.

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., & Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 29-34.

Dubey, A. D. (2020). Twitter sentiment analysis during COVID-19 outbreak. Available at SSRN 3572023.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 6(4), e21978.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE transactions on computational social systems, 8(4), 1003-1015.

Alanezi, M. A., & Hewahi, N. M. (2020, October). Tweets sentiment analysis during COVID-19 pandemic. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-6). IEEE.

Jabalameli, S., Xu, Y., & Shetty, S. (2022). Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination. International Journal of Disaster Risk Reduction, 80, 103204.

Barkur, G., & Kamath, G. B. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry, 51, 102089.

de Las Heras-Pedrosa, C., Sánchez-Núñez, P., & Peláez, J. I. (2020). Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. International journal of environmental research and public health, 17(15), 5542.

Yeasmin, N., Mahbub, N. I., Baowaly, M. K., Singh, B. C., Alom, Z., Aung, Z., & Azim, M. A. (2022). Analysis and prediction of user sentiment on COVID-19 pandemic using tweets. Big Data and Cognitive Computing, 6(2), 65.

Sivanantham, K. (2021). Sentiment analysis on social media for emotional prediction during COVID-19 pandemic using efficient machine learning approach. Computational intelligence and healthcare informatics, 215-233.

Kausar, M. A., Soosaimanickam, A., & Nasar, M. (2021). Public sentiment analysis on Twitter data during COVID-19 outbreak. International Journal of Advanced Computer Science and Applications, 12(2).

Vijay, T., Chawla, A., Dhanka, B., & Karmakar, P. (2020, December). Sentiment analysis on covid-19 twitter data. In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (pp. 1-7). IEEE.

Alanezi, M. A., & Hewahi, N. M. (2020, October). Tweets sentiment analysis during COVID-19 pandemic. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-6). IEEE.

Ahmed, M. S., Aurpa, T. T., & Anwar, M. M. (2021). Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic. Plos one, 16(8), e0253300.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 6(4), e21978.

Alshamrani, S., Abusnaina, A., Abuhamad, M., Lee, A., Nyang, D., & Mohaisen, D. (2020). An analysis of users engagement on twitter during the covid-19 pandemic: Topical trends and sentiments. In Computational Data and Social Networks: 9th International Conference, CSoNet 2020, Dallas, TX, USA, December 11–13, 2020, Proceedings 9 (pp. 73-86). Springer International Publishing.

Mahalakshmi, V., Shenbagavalli, P., Raguvaran, S., Rajakumareswaran, V., & Sivaraman, E. (2024).

Twitter sentiment analysis using conditional generative adversarial network. International Journal of Cognitive Computing in Engineering, 5, 161-169.

Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 23(1), 463-470.

Bania, R. K. (2020). COVID-19 public tweets sentiment analysis using TF-IDF and inductive learning models. INFOCOMP Journal of Computer Science, 19(2), 23-41.

Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research, 54-65.

Nezhad, Z. B., & Deihimi, M. A. (2022). Twitter sentiment analysis from Iran about COVID 19 vaccine. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 16(1), 102367.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied soft computing, 101, 107057.

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. Information, 12(5), 204.

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, A. (2020). Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. J. Crit. Rev, 7(9), 2761-2774 .

Braig, N., Benz, A., Voth, S., Breitenbach, J., & Buettner, R. (2023). Machine learning techniques for sentiment analysis of COVID-19-related twitter data. IEEE Access, 11, 14778-14803.

Sharma, S., & Sharma, A. (2020, November). Twitter sentiment analysis during unlock period of COVID-19. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 221-224). IEEE.

Alabid, N. N., & Katheeth, Z. D. (2021). Sentiment analysis of twitter posts related to the covid-19 vaccines. Indonesian Journal of Electrical Engineering and Computer Science, 24(3), 1727-1734.