How to Cite:
Ponmani, K., & Thangaraj, M. (2022). Clustering based sentiment analysis on Twitter data for COVID-19 vaccines in India. *International Journal of Health Sciences*, *6*(S2), 4732–4748. https://doi.org/10.53730/ijhs.v6nS2.6126

Clustering based sentiment analysis on Twitter data for COVID-19 vaccines in India

Ponmani K
Bharathiar University, Coimbatore, India
Email: ponmanidurai@gmail.com

Thangaraj M
Madurai Kamaraj University, Madurai, India
Email: thangarajmku@yahoo.com

***Abstract*---**Coronavirus is a new and rapidly spreading viral disease. It is essential to have a vaccine in order to reduce the virus's impact. Vaccination-related sentiments can influence an individual's decision to accept the vaccines. Evaluating the sentiments is a time-consuming and challenging process. Sentiment analysis (SA) could have an impact on the vaccination initiatives as well as changes in people's opinions and behaviour around immunizations. Since social media is widely utilized to disseminate information, mining this data is a popular area of study these days. On Twitter, a wide range of opinions about the negative effects of licensed vaccines have been expressed over time. In this research, tweets are gathered, pre-processed to remove extraneous data, and then utilized for sentiments analysis utilizing the Lexicons-based technique and machine learning. After feature extraction, the clustering is performed using MEEM approach. This research proposed a Clustering Based Twitter sentiments analysis of COVID 19 (C-SAT COVID 19) vaccinations in India. An enhanced random forest classifier is implemented in this research to classify the sentiment scores provided by the sentiment analysis. A classification is performed based on the negative, neutral, and positive sentiment analysis to examine people's emotions towards vaccinations accessible in India. The proposed model classifies the sentiment analysis for Covishield as 46.83% positive, 38.42% negative and 14.75% neutral, for Covaxin- 45.27% positive, 40.01% negative, and 14.72% neutral, and finally for Sputnik- 42.08% positive, 42.56% negative, and 15.36% neutral. As a classification result, the proposed model obtained the accuracy of 95.15%, 94.52% sensitivity, and 95.66% specificity, which is better compared to the other existing models.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Ponmani, K.; Email: ponmanidurai@gmail.com*
Manuscript submitted: 27 Feb 2022, Manuscript revised: 18 March 2022, Accepted for publication: 09 April 2022

***Keywords*---**COVID-19 vaccines, improved random forest, machine learning, MEEM, sentiment analysis.

Introduction

Social media is very important in people's lives since it connects us to the entire world. Nowadays, it is impossible for a person to work without access to social media in order to cover all of the updates, news, and events such as coronavirus updates, vaccination updates, and other things. People nowadays rely increasingly on posts and tweets shared on social networking sites such as Instagram, Facebook, and Twitter (Ansari & Khan, 2021). People use social media to share their information, experiences, emotions, and so on. It serves as a communication platform for people who share similar interests to connect with one another. Several tools and strategies are available to research the network that exists in the social media platform, but they are rarely applied in the context of sickness. In a disease context, understanding the link between users provides useful information such as identifying the influential person, the closeness of the network, and so on. Identifying prominent people and other network features aids in the efficient and rapid dissemination of proper information on disease prevention, treatment, and care, as well as the prevention of fake news from being published with little effort (Augustyniak, Szymański, Kajdanowicz, & Tuligłowicz, 2015).

People can express themselves using social media sites such as Facebook, Twitter, and Instagram. Twitter has grown to be one of the very common micro blogging sites for exchanging opinions, information and news. Twitter has been used by researchers to analyze attitudes and sentiments in practically every aspect of life, including politics, religion, the environment, and government policy (Ansari & Khan, 2021) (Augustyniak, Szymański, Kajdanowicz, & Tuligłowicz, 2015).

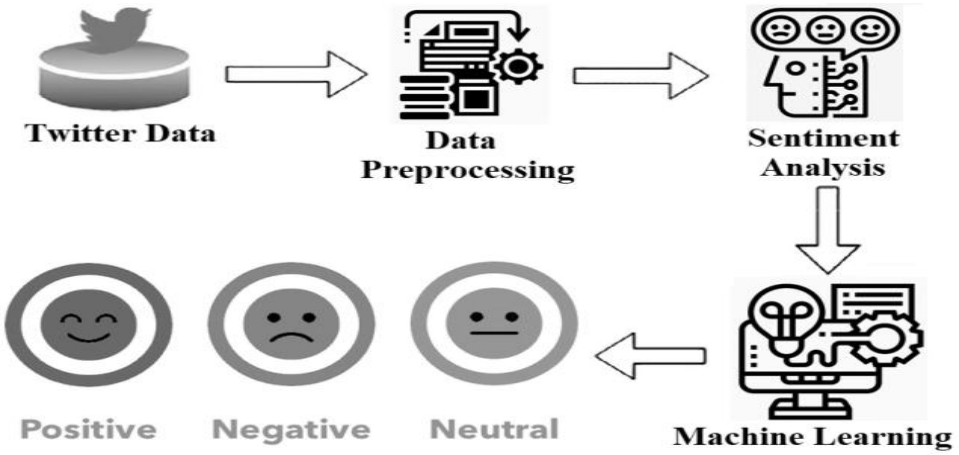

Figure.1. Process of Twitter-based Sentiment Analysis

Sentiment analysis, which detects emotions or views in a big amount of unstructured data, can yield significant insights from social media sites. Analysis

of sentiment was a systematic study of natural language processing (NLP) that encompasses the computational examination of opinions, sentiments, and feelings presented in textual format. SA is divided into three polarity categories: negative, neutral, and positive. Every polarity of a tweet's was fixed by allotting the score ranging from -1 to 1 based on the words utilized, where the negative scores denote the negative attitudes, the positive scores show a positive emotion, and a zero value represents a neutral sentiment (Chaudhri, Saranya, & Dubey, 2021).

Coronavirus disease has had a significant impact on people's daily lives all around the world. Because of the spread of the coronavirus, the world is in an extremely grave situation (Fauzi, 2018). People all throughout the world are dying as a result of the COVID 19 infection. In a short amount of time, social media networks such as Twitter have seen an exponential increase in vaccination-related tweets. Health care preferences are personal and rely on patients' willingness to select the particular therapies. It is critical to perceive the sentiment of individuals within certain demographic or geographic communities in order to develop vaccination distribution programmes. Public health authorities and health care providers must plan the distribution scheme based on the public's attitudes toward vaccination (Hananto, Rahayu, & Hariguna, 2022).

Vaccination-related feelings can influence an individual's decision to receive the vaccine. Measuring these emotions is a time-consuming and difficult undertaking. The World Health Organization has approved vaccinations such as Covishield, Covaxin, and Sputnik in India (WHO). In terms of effectiveness, the general population had conflicting feelings about the permitted immunizations. On Twitter, a wide range of opinions about the negative effects of licensed vaccines have been expressed over time. In this work, tweets are gathered, pre-processed to remove extraneous data, and hence utilized for analysis of sentiment utilizing the Lexicons-based technique and improved random forest approach. This research proposes a Twitter-based SA of COVID-19 vaccinations in India. The rest of the work is presented in the following sections. Section two presents the literature review for the research that covers the analysis of the related works. Section three presents the proposed research model, section four presents the results and analysis of the proposed model, and finally, section five presents the conclusion and future works.

Literature Review

A Hybrid Heterogeneous SVM technique was proposed in (Fauzi, 2018) for analyzing COVID 19 tweets. The Twitter data was analyzed using the R programming language. Based on data such as hash tag keywords connected to COVID 19, an analysis was performed and the data was classed with positive, negative, and neutral sentiment ratings. As a limitation, no sufficient details on numerical results were discussed in this work, resulting in no justification for the results. Based on Twitter tweets about COVID-19 vaccines, a content analysis model was implemented in (Hananto, Rahayu, & Hariguna, 2022) that used the Social Judgment Theory, Elaboration Likelihood Model, and the Extended Parallel Process Method as theoretical models. According to the findings, the Health Information Persuasion Exploration (HIPE) response strategy was established for

addressing dis/misinformation and anti-vaccines messaging. A SA model based on Twitter data was proposed in (Jalil, et al., 2021) using several deep learning and machine learning classifiers. The implemented model evaluates the sentiments of accumulated tweets for classification of sentiment utilizing different classifiers and feature sets. Tweets were divided into three categories based on their sentiment: favourable, negative, and neutral. The early detections of COVID 19 sentiments in tweets provides for a better knowledge and management of the epidemic.

In (Jianqiang, Xiaolin, & Xuejun, 2018), machine learning approaches such as SVM and Naive Bayes classifiers were employed to analyze the sentiment of Twitter accounts. Based on this SA, a comparison of lexicon tools such as Word Sense Disambiguation (WSD), SentiWordNet, and Text Blob was performed in order to find an adaptive lexicon tool. In this study, WSD fared better with tweets. SA and classification of Indian farmers' protests using Twitter data was proposed in (Kaur, Ahsaan, Alankar, & Chang, 2021) using Bag of Words and TF-IDF. For classification and prediction, four conventional machine learning models were used: random forests, SVM, decision trees, and Naive Bayes. The random forest algorithm produced the best results. The constraint was the keywords used to discover material relevant to the purpose. If the keywords were not utilized in the messages, some relevant tweets were most likely overlooked.

Three transfer learning algorithms were implemented in (Kausar, Soosaimanickam, & Nasar, 2021) for analyzing public reaction about HPV vaccine on Twitter. The first involved transferring embeddings and static embeddings from languages model (ELMo) and hence processing them using bidirectional gate recurrent units with attentions, DWE-BiGRU-Att. Fine-tuning the pre-trained model with constrained annotation data, known as fine-tuned generative pre-training (GPT), and BERT-bidirectional encoders representation from transformer were the other strategies. The pre-trained GPT model was used to build the fine-tuned GPT model. The BERT method was used to build the fine-tuned BERT model. These methodologies frequently fail to take into account common-sense understanding for public opinion on common health problems. This issue should be addressed by knowledge augmented ensemble learning model on social media. A SA technique was used to present an analysis of the public's opinions to present vaccinations drives around the world via COVID 19 vaccine-related tweets (Narasamma & Sreedevi, 2021). The Naive Bayesian approach was utilized for SA. Overall, the tone of the Tweets was largely unfavourable, and a significant vaccinations trend could be noticed in global health perspective, as indicated by an examination of the significance of comprehensive research and science in vaccinations. In this paper, there is a lack of detail when describing the performance model.

The word embeddings method based on unsupervised learning on big Twitter corpora was proposed in (Neogi, Garg, Mishra, & Dwivedi, 2021). This strategy made advantage of latent contextual semantics link and statistical co-occurrences characteristic between words in tweet. To develop a sentiment features collection of tweets, the word embeddings integrated n-grams with words sentiment polarity score feature. For training and predicting sentiment categorization labels, the feature set was merged into a deep CNN. The Catboost Recurrent Neural

Framework with the error pruning technique was implemented in (Paliwal, Parveen, Afshar Alam, & Ahmed, 2022) to assess SA utilizing tweets from Twitter data based on the opinions of user. The tweets regarding COVID 19 were taken as the dataset that was used to classify the sentiment as negative, positive or neutral. The layer of this CRNF model was used to classify sentiments using aspects words. The results may have been focused on SA if they had been examined solely on standard classification. A sentiment analysis based on the random

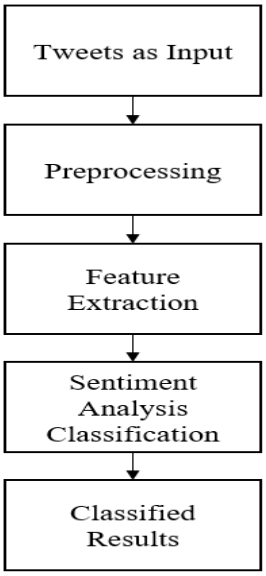

Figure.2. General Process for SA Classification

forest method was proposed in (Ramamoorthy, Karmegam, & Mappillairaju, 2021). The fundamental issues of sentiment analysis and sentiment polarity classifications were addressed. In (Robnik-Šikonja, 2004), a similar sentiment analysis task using a random forest classifier was implemented. In this work, the COVID 19 vaccine sentiment analysis was conducted using Twitter data. The Vader and Text Blob lexicon approaches were used to analyze sentiment. A linear regression classification method was proposed in (Scannell, et al., 2021) to evaluate Twitter tweets on the COVID-19 vaccine and determine positive or negative sentiment in the text. It would be preferable if the processed data were processed and normalized to avoid producing a too wide emotion, resulting in a greater level of accuracy. One significant disadvantage was the lack of data.

Proposed Methodology

In this research, a sentiment analysis model was proposed to evaluate tweets from Twitter on the threeCOVID19 vaccines (Covishield, Covaxin, and Sputnik) and indicate the negativity or positivity of sentiments in the texts using NLP and an improved random forests classification algorithm. To develop the model, initially, tweet data was received from Twitter using the Tweepy library using Twitter's normal search, and the accumulated information was saved in CSV data

format. Because the tweet comprises hyperlinks, special characters, emoji, retweets, and stickers, this data must be pre-processed. NLP is used to pre-process data and prepare it for the implementation of a supervised classification algorithm. Data was tokenized after the special characters have been removed. The data was lemmatized and normalized before further processing. After the data has been pre-processed with NLP, subjectivity and polarity are computed. The polarity data was classified using an improved random forest classifier.

Data Collection

The tweets were collected using Twitter's Python library "Tweepy." It allows for the retrieval of relevant data by searching for trends, hash tags, keywords, or geo-locations. The accessible Twitter messages or text ("tweets") with minimum one of the following key terms were collected: "Covishield", "Sputnik", "Covaxin", "Corona Virus", "Vaccine" and "COVID-19". For this study, keywords in English were utilized for extracting tweets instead of hash tags since keywords techniques were more comprehensive since, they include hash tags tweet as well. The tweet time, text and date of posting, and user location were all extracted. The tweets were collected from the 24th to the 30th of June 2021. 15,843 tweets for Indian locations were acquired from Twitter, pre-processed, and hence utilized for performing SA. Duplicates in Twitter data may occur when the similar tweets were acquired again through API or if numerous users retweet/post the same text. This study makes use of 8460 tweets after removing duplicate rows and irrelevant data.

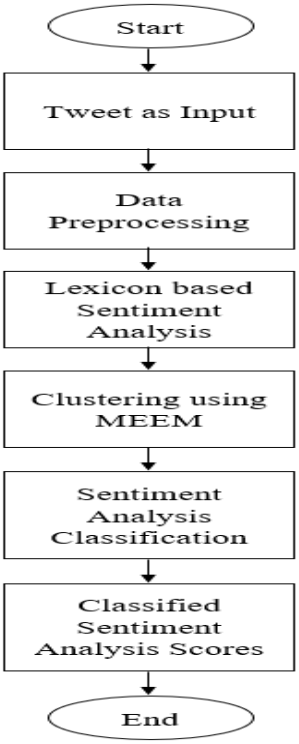

Figure.3. Workflow of C-SAT COVID 19 Model

Pre-Processing Stage

Normally, Twitter data is unstructured and comprises extraneous symbols that are not required for the experiment of SA. The raw data must be pre-processed in order to be suitable for the use of proposed improved random forest approach (Shamrat, et al., 2021). Pre-processing is a main step before conducting SA to clean the original tweets by eliminating noise and undesired features. The tweets were pre-processed using tokenization, stop words removal, and texts normalization.

Stop word removal: Stop words are normal words that appear in phrases and provide weight to the sentences, like "of," "the," and "for." These words serve as a connection between phrases and ensure that they were correct by grammar. Stop words are filter out words that are used prior to processing natural languages input. This is a common pre-processing strategy. A set of stop words listed in the Python nltk module was utilized in this procedure, and a custom function was constructed to replace contraction. A spell check was performed to correct misspelt words in order to reduce complexity.

Tokenization is a way for dividing the sentences, phrases, paragraph, or entire texts into small bits. Individual terms or phrases can be acquired in this manner, and all these terms individually was called as a token. It comprises the following:

- Eliminate unnecessary terms: Remove non-ASCII terms and unnecessary words from the tweet. All the symbols were replaced by empty space.
- Eliminate any hash tags or mentions: Removing the data of hash tags or mentions would aid in noise reduction.
- Eliminate Email and URL: Using regular expressions eliminate emails and URLs.
- Eliminate duplicates that offer no value.
- All text in tweets is changed to lowercase.

Text normalizations: Texts should be normalized prior it can be processed further since it enhances text matching. Normalization, in general, means that words are treated equally and that processing can proceed uniformly. The two activities performed to normalize the text are stemming and lemmatization:

Stemming is the process of removing affixes from the words to create a word stem. In stemming, the end of the word is typically chopped off, and it functions better majorly. Words such as vaccinating, vaccinate, and vaccinated, for example, are derived from the term "vaccine." Porter stemmer is the most commonly used technique since it is very rapid.

Lemmatization: The purpose was similar to stemming, however stemming sometimes causes the word's meaning to be lost. It minimizes the count of words in the texts and hence improves the analysis. Lemmatization refers to doing things correctly with morphological and vocabulary words analysis. It returns the dictionary or vocabulary form of the words, which was called as a lemma. It is a real language term restored by lemmatization, whereas stem may or may not be a real language words restored by stemming. Words such as virus and coronavirus,

for example, were changed into their citations form "COVID." Normalization procedures include stemming and lemmatization. The lemmatization strategy was utilized to normalize in this work since it is an appropriate approach for vocabulary and morphological word analysis (Singh, 2021).

Feature Extraction

Vectorization is the process of translating tokens to numbers, and it was a necessary step because the algorithms require data in numerical instead of texts form (Villavicencio, Macrohon, Inbaraj, Jeng, & Hsieh, 2021). Each tweet is presented as the vector in the space term, with preprocessing stage's specific words serving as its features. The value of the feature vector is decided by some term weighting procedure. Inverse Document Frequency (IDF) and Term Frequency (TF), and the integration of both, TD-IDF, are the most used term weighting methods. TF-IDF was the mathematical statistic approach used to find how important the words are to the corpus, as well as to represent how relative a phrase was in the specific texts. The word's frequency in the tweet is compared to the frequency of the word in the full set of tweets in this procedure.

Term Frequency assigns weights by assuming that each term contributes in proportion to the number of times it appears in the document. Some prominent TF variants are Raw TF, Binary TF, and Logarithmic TF. Every tweet was presented as the binary vectors using Binary TF. The term that appears in the tweet would be assigned a value of 1 in the vectors; or, the terms that does not exist in the tweet would be assigned a value of 0. This weighting of terms does not take into account the count of occurrences of the word, just 0/1 values.

Unlike Binary TF, the Raw TF approach takes into account the count of term occurrences. The number of times a phrase appears in a tweet determines its worth. Meanwhile, Logarithmic TF takes into account the count of occurrences of the phrase. The distinction is that Logarithmic TF believes that the relevance of a term in a tweet does not rise according to how often it occurs in the tweet. Using Logarithmic TF, the weight of terms t in tweet d could be computed as follows:

$$\mathrm{TF}(t, d) = 1 + \log(f_{t,d}) \qquad (1)$$

Where, $f_{t,d}$ was the count of times the term t appears in tweet d.

Meanwhile, IDF is a global word weighting calculated based on the phrase's distribution in the data set. This weighting of terms would provide larger values to the rare terms that presents in only a few tweets. The weights of term t calculated with IDF are as follows:

$$\mathrm{IDF}(t) = 1 + \log\left(\frac{N_d}{df_t}\right) \qquad (2)$$

Where N_d was the count of tweets in the data set and df_t was the count of tweets in the data set that contain the term t.

The most commonly used phrase weighting was TF-IDF. It was the product of IDF and TF. The following is a count of the weight integration of terms t in tweet d.

$$TF \cdot IDF(t, d) = TF(t, d) \cdot IDF(t) \quad (3)$$

$TF(t, d)$ represents the TF values of terms t in tweet d, while IDF(t) represents the IDF values of terms t.

Clustering using Modified Efficient Expectation Maximization

MEEM is technique that is familiar with K-means for the way it substitutes among an Expectation step (E-step), which is equal to movement, and a Maximization step (M-step), which is similar to recompilation of the model's limitations. The centroids are located at the edges of K-means. The conditional parameters were recalculated throughout the maximizing step. 'd' is the amount of Twitter data provided in accordance with a probability distribution.
MEEM Algorithm:

1. Input: Tweet Data
2. Output: C (c_0, c_1, c_2, c_3)
3. i←1
4. $\alpha_k, N, \beta_{nk}, f_{mk}, c$←Initialize ($\alpha_k, N, \beta_{nk}, f_{mk}, c$)
5. Convergence ←False
6. While Convergence == False do
7. i←i+1
8. Log Probability
9. $\alpha_k = \frac{\sum_{n=1}^{N} \beta_{nk}}{N}$
10. Expectation Step
11. $\beta_{nk} = \frac{\alpha_k (\prod t_m \in d_n f_{mk})(\prod t_m \notin d_n (1 - f_{mk}))}{\sum_{k=1}^{k} \alpha_k (\prod t_m \in d_n f_{mk})(\prod t_m \notin d_n (1 - f_{mk}))}$
12. Maximization Step
13. Update the parameter based on β_{nk}
14. $f_{mk} = \frac{\sum_{m=1}^{N} \beta_{nk} I(t_m \in d_n)}{\sum_{n=1}^{N} \beta_{nk}}$
15. If $|\alpha_k, \beta_{nk}, f_{mk}| < \varepsilon$ then
16. Convergence ← True
17. End While
18. return $\alpha_k, \beta_{nk}, f_{mk}$

Here, c indicates clusters, α_k indicates log probability, β_{nk} indicates probability estimation of every tweet, and f_{mk} indicates the updation of parameters based on β_{nk}.

The E-step is in charge of estimating the parameters of each cluster's probability distribution. This algorithm terminates when the distribution variables have reached convergence or when the maximum count of iterations has been reached. It accepts the tweets as input and returns the number of clusters found in the tweets. The algorithm described above initializes all of the parameters. If the convergence is not true, find the log probability of the situation. Increasing the value; calculating the probability of each tweet (E-Step); and updating the

parameters (M-Step) depending on the results of E-Step. If the convergence is successful, all of the parameters are returned. Due to the speed with which the algorithm operates when working with only two mixture components, the E-Step and M-Step processes could be iterated repetitively till the model parameters do not modify by a specified ε, where ε is as small as 0.0001.

Improved Random Forest

Random forest is an algorithm for ensemble learning. The algorithm's core concept is that creating a tiny decision tree with few features is a computationally efficient procedure. An enhanced random forest classifier accomplishes classification with the fewest trees possible. The proposed approach removes some unimportant features iteratively. The theoretical upper limits on the count of trees added to the forest to assure increase in classification accuracy was derived according to the count of unimportant and important attributes. The method converges on a small but significant collection of features (Zhang, Fan, Peng, Rao, & Cong, 2020).

In the proposed improved random forest, the count of trees in the random forests grows iteratively. A construction pass is the name given to each of these iterations. The procedure begins with a small number of trees. The list of vital and unimportant features was then updated at each construction pass using the four methods outlined below. First, the weights of various features were determined, followed by their ranking based on those weights. The threshold weight was then computed. The characteristics with weights less than the threshold are then eliminated. Following that, from the sets of remaining traits, some were designated as 'important' based on the criteria. The remaining features are labelled as 'unimportant.' It is worth noting that once a feature is designated as significant during the constructions pass, it would remain so until the end and would not be eliminated in later passes (Scannell, et al., 2021).

Assume that the starting feature vector $\mathcal{F}_0(\cdot)$. An entropy-based assessment for splitting of node was utilized while building a random tree. Consider node i in the trees. Let $p(c)$ be the class label c's probability at this node. Thus, that node's entropy was,

$$E = \sum_{\forall c} p(c) \ln \frac{1}{p(c)} \quad (4)$$

To separate this node, a set ($\forall$) of f features were chosen at random from $\mathcal{F}_0(\cdot)$ with no replacement. Tree T's normalized weight is

$$\gamma^T = \frac{1/\delta^T}{\max_T(1/\delta^T)} \quad (5)$$

A higher γ^T value suggests that tree T makes less classification errors. As a result, the attributes utilized to separate the nodes of the tree T were more differential. The global weights of feature j were calculated utilizing the local weights of the features and the tree's weights:

$$w(j) = \frac{\sum_{\forall T} w^T(j)\gamma^T}{\max_j \sum_{\forall T} w^T(j)\gamma^T} \quad (6)$$

An attribute with a greater weight value is more relevant for classification. The features were ordered according to the global weights w(j) for determining the unimportant and important attributes. There is no way of knowing how many attributes were important (class discriminative). As a result, an approach was employed to identify the relevant attributes. At first, the top u_0 features on the ranked list were labelled as 'important,' while the rest were labelled as 'unimportant.'

Let Γ (Γ_0) represent the bags of significant feature and Γ' (Γ_0') represent the bags of irrelevant feature. These feature bags were updated with each build pass. First, the initial forests Θ_0 with B_0 trees and feature vectors $\digamma_0(\cdot)$ are created. Following the features ranking by (6), the top u_0 count of features are designated as important. The significant features are maintained in bag Γ_0, while the remaining features are kept in bag Γ_0'. Let σ_0 and u_0 be the standard deviation and mean of feature weight in Γ_0', respectively. The initial subsets of feature from Γ_0' whose weights were lower than (u_0-$2\sigma_0$) is then R_0. (If no such features exist, R_0 comprises the features with the small global weights in Γ_0'). The attributes of R_0 from Γ_0' are deleted. $\digamma_1(\cdot) = \digamma_0(\cdot) - R_0$ is the feature vector with a decreased set of features. The Δu and Δv are then computed using (7). Let v and u represent the count of unimportant and important attributes, individually.

$$\Delta u = \#\Gamma_{n+1} - \#\Gamma_n, \Delta v = \#\Gamma'_{n+1} - \#\Gamma'_n \quad (7)$$

Based on these variables, the count of trees to be integrated (ΔB) to the forests was calculated using (8).

$$|v\Delta B| < |lq_u\Delta u + lq_v\Delta v| \rightarrow |\Delta B| < \left|\frac{l(q_u\Delta u + q_v\Delta v)}{v}\right| \quad (8)$$

Then a forest Θ_1 is built from scratch using $B_0 + \Delta B$ and feature vectors $\digamma_1(\cdot)$. The features are rated first at any pass n (following the expansion of the forests Θ_n with B_n tree and the reduction of the feature set $\digamma_n(\cdot)$). The set of new significant characteristics A_n and the sets of attributes to be minimized R_n are then discovered. The feature vector with decreased features is obtained $\digamma_{n+1}(\cdot) = \digamma_n(\cdot) - R_n$, and the feature bags are updated using (9).

$$\Gamma_{n+1} = \Gamma_n - A_n, \Gamma'_{n+1} = \Gamma'_n - R_n - A_n \quad (9)$$

Following that, Δu and Δv are determined from (7) and ΔB using (8). Finally, a forest $\digamma_{n+1}(\cdot)$ with $B_n + \Delta B$ trees and feature vectors $\digamma_{n+1}(\cdot)$ was constructed. Random forest is a generic formulation. The converged random forest classifier is used to classify unlabelled test data. The test data was included at the root nodes of every tree in the converged forests for classification. The tree weights were obtained by (5). Weighted voting from the tree was utilized to determine the class labels of the testing data as it reaches the leaf nodes of the tree. Let B^* be the number of trees in the converged forest, and $\gamma^*(T)$ be the weights of trees T in the converged forests. If $p_T^*(c)$ represents the class label probability anticipated by

trees T for the test data of input, hence the original class labels (c^*) for the test data of input was presented by:

$$c^* = arg \max_{c} \sum_{T=1}^{B^*} \gamma^*(T) p_T^*(c) \quad (10)$$

As a result, the final experimental analysis utilizing the proposed model is described in the next section.

Experimental Analysis

The tweets gathered for this analysis is from the public domain called Twitter, and there is no interaction with users. Unspecified user data has been excluded from the research findings. 15,843 tweets were collected for this analysis and after removing duplicate rows and irrelevant data, 8460 tweets were used in this research for evaluation. Sentiment analysis is divided into three polarity categories: negative, neutral, and positive. Each tweet's polarity is established by allotting the number from -1 to 1 according to the words utilized, where the negative scores indicate the negative attitude, the positive scores show the positive sentiments, and a zero value represents the neutral sentiments. A subjectivity score is awarded to every tweet according to whether it has an objective or subjective meaning; the subjectivity score range is likewise from 0 to 1, with a value near 0 representing objective and a value near 1 representing subjective. From the subjectivity and polarity data, the median, mean, maximum mean, and minimum mean were determined for every vaccination. The maximum average polarity was determined once every ten tweets.

$$mean\ \bar{x} = \frac{\sum x}{n} \quad (11)$$

$$median = \frac{n+1}{2} \quad (12)$$

$$Min.Average = \frac{(n-1)min+max}{n} \quad (13)$$

$$Max.Average = \frac{min+(n-1)max}{n} \quad (14)$$

By applying equations (11) to (14), the mean, median, average minimum (Min.Avg), and average maximum (Max.Avg) for the three vaccinations are determined according to the tweet subjectivity and polarity. The results of the calculations are displayed in the tables below.

Tables 1 and 2 demonstrate the text polarity and subjectivity scores for the three vaccinations, Covishield, Covaxin, and Sputnik. Table 3 shows the quantity of tweets assessed for sentiment analysis from the collected tweets used in this study. Figure 4 shows the sentiment analysis graph in terms of negative, positive, and neutral emotions, as well as the overall number of tweets. The greatest average is derived from the polarity of ten tweets.

Table.1. Computation of Polarity for three Vaccines based on Tweets

Vaccine	Polarity			
	Mean	Max.Avg	Min.Avg	Median

Covishield	0.133416	1.0	-1.0	0.0
Covaxin	0.115412	1.0	-1.0	0.0
Sputnik	0.079158	1.0	-1.0	0.0

Table 2. Computation of Subjectivity for three Vaccines based on Tweets

Vaccine	Subjectivity			
	Mean	Max.Avg	Min.Avg	Median
Covishield	0.34101	1.0	-1.0	0.3343
Covaxin	0.349514	1.0	-1.0	0.376
Sputnik	0.305892	1.0	-1.0	0.310

Table.3. Sentiment Analysis based on Number of Tweets

Sentiment	Number of Tweets
Negative	3384
Neutral	1269
Positive	3807

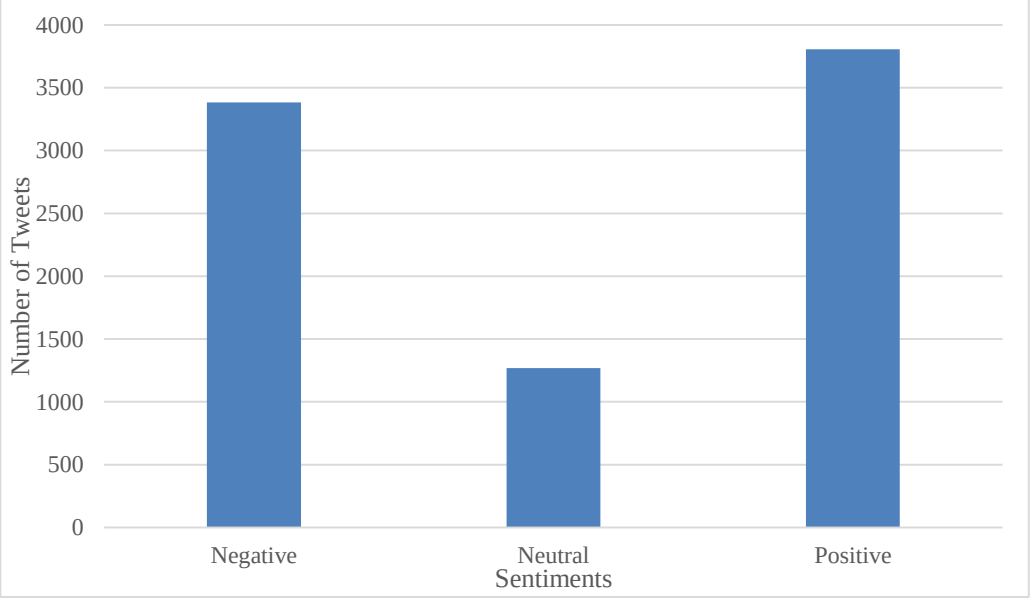

Figure.4. Comparison of Sentiment Analysis based on Number of Tweets

The textual data is transformed into polarity scores. Because the improved random forest classification approach cannot process text data, this score was employed in the classification process. In addition, polarity reveals the sentiment or emotions behind the text data. In the proposed model, a supervised improved random forest classification method was used. This approach divides the polarity scores into three categories: negative, positive, and neutral. Classification was performed on the data of each vaccine under consideration. Table 4 shows the classification's final outcome.

The twitter data shows that there are negative, positive, and neutral attitudes concerning the three distinct vaccines. Figure 5 depicts a comparison of classified sentiment scores for vaccines. All three vaccines were compared using this comparison figure and the sentiment scores provided by the suggested model in terms of negative, neutral, and positive. Based on the sentiment scores, the majority of respondents had minimally positive and maximally negative feelings towards the Sputnik vaccination. Based on the tweets, Covishield and Covaxin obtained sentiments that were comparable. One of the main reasons for the least performance on sputnik was it is not available for free and it was introduced in India after the Covishield and Covaxin vaccinations. All the government agencies mostly providing Covishield and Covaxin vaccinations in India for the general public.

Table.4. Sentiment Classification for three Vaccines based on Tweets

Vaccine	Negative	Neutral	Positive
Covishield	38.42	14.75	46.83
Covaxin	40.01	14.72	45.27
Sputnik	42.56	15.36	42.08

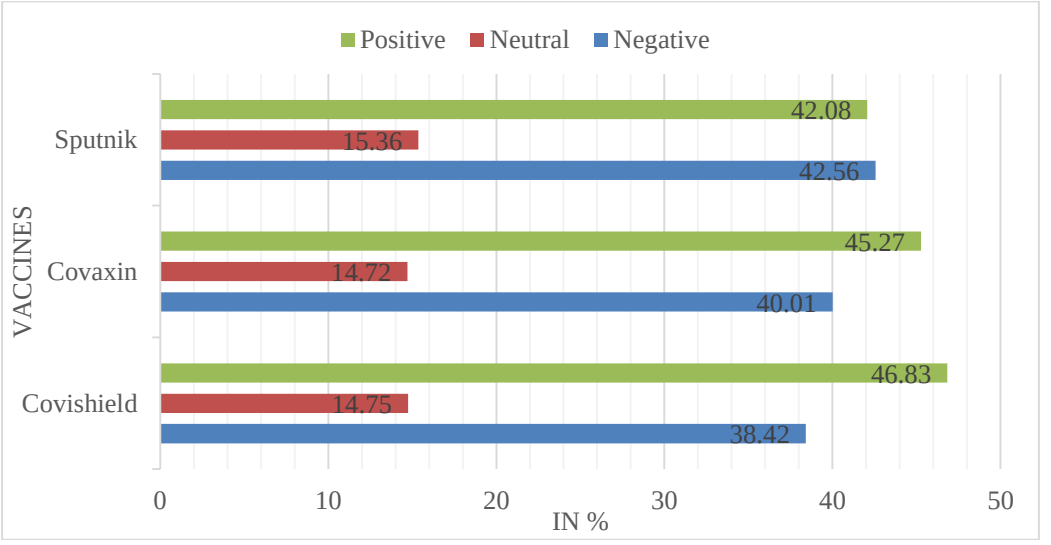

Figure 5. Comparison of Classified Sentiment Scores for Vaccines

Table 5. Performance Analysis Comparison based on Sentiment Classification

Models	Accuracy	Sensitivity	Specificity
Decision Tree	92.95	90.87	93.02
Naïve Bayes	93.33	92.70	93.95
SVM	94.19	93.21	94.79
Random Forest	94.67	93.98	95.02
C-SAT COVID 19	95.15	94.52	95.66

As shown in table 5, the proposed improved random forest classifies the twitter analysis based on the sentiment scores and this performance analysis of the proposed model was compared with other existing models like random forests, SVM, decision trees, and Naive Bayes for the validation of results. The proposed model obtained the accuracy of 95.15%, 94.52% sensitivity, and 95.66% specificity, which is better compared to the other existing models as shown in figure 6.

From the results, it is clear that the tweets about Covishield, Covaxin, and Sputnik were evaluated. It was found that Covishield had more positive sentiments than Covaxin and Sputnik. The following are the primary causes of these results:

- Many Indian scientists have expressed concerns about Covaxin's approval, alleging that the vaccine was approved without appropriate efficacy data. As a result of such assertions, the Indian population is skeptical of this immunization.
- According to claims, Covaxin trial volunteers in Bhopal were deceived and vaccinated without their knowledge. The participants also claimed that they were not informed about the potential adverse effects, and that such reports have resulted in such negative attitudes and emotions as wrath and contempt in Covaxin tweets.

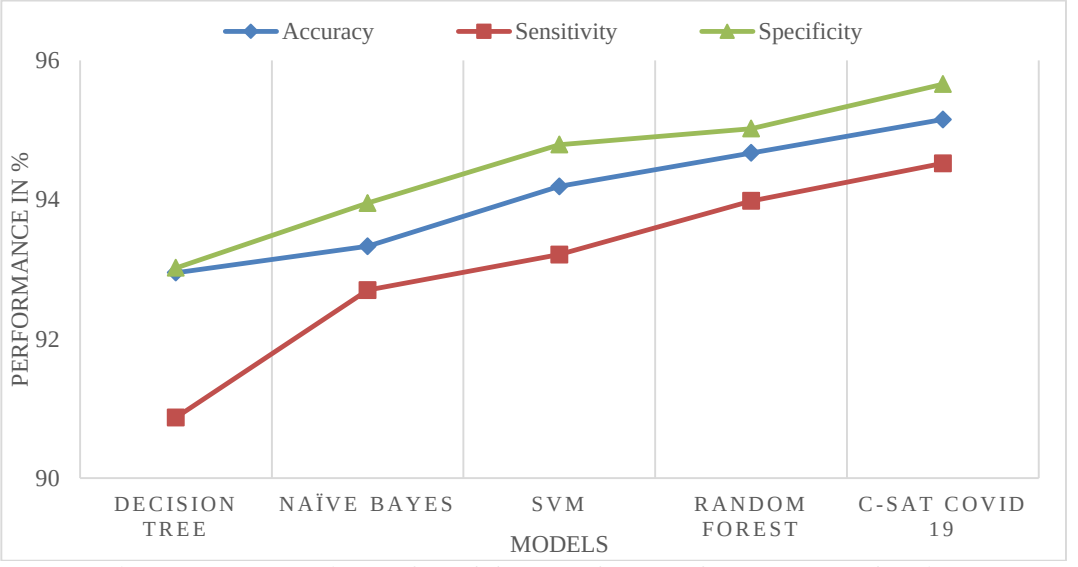

Figure.6. Comparison of Models Based on Performance Evaluation

Conclusion

A sentiment analysis classification model was proposed in this research to analyze Twitter tweets related to the COVID 19 vaccines (Covishield, Covaxin, and Sputnik) and show the negativity or positivity of sentiments in the text using NLP and an improved random forest classification algorithm. To acquire data, the Twitter API was used to collect tweets on vaccines in India A sentiment analysis

classification model was proposed in this research to analyze Twitter tweets related to the COVID 19 vaccines (Covishield, Covaxin, and Sputnik) and show the negativity or positivity of sentiments in the text using NLP and an improved random forest classification algorithm. To acquire data, the Twitter API was used to collect tweets on vaccines in India. Based on the sentiment analysis score, this proposed model performed data pre-processing, feature extraction, clustering and classification. The proposed model classifies the sentiment analysis for Covishield as 46.83% positive, 38.42% negative and 14.75% neutral, for Covaxin- 45.27% positive, 40.01% negative, and 14.72% neutral, and finally for Sputnik- 42.08% positive, 42.56% negative, and 15.36% neutral. The proposed model C-SAT COVID 19 obtained the accuracy of 95.15%, 94.52% sensitivity, and 95.66% specificity, which is better compared to the other existing models. During the analysis period, public opinions of vaccines varied, with initial debates focusing on public worries regarding the creation of vaccines that would later be distributed to the public, followed by discussions on side effects produced by certain types of vaccines. Overall, there has been a vast variance in user reactions to information recently, with some users showing a desire in engaging in positive behaviours. The number of people receiving vaccinations is increasing, and 54.4 percent of the population in India is fully vaccinated. In the future, a global sentiment analysis based on vaccination can be undertaken to examine the sentiment of different countries around the world.

References

Ansari, M. T., & Khan, N. A. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. *Electronic Journal of General Medicine, 18*.

Augustyniak, Ł., Szymański, P., Kajdanowicz, T., & Tuligłowicz, W. (2015). Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy, 18*, 4.

Chaudhri, A. A., Saranya, S. S., & Dubey, S. (2021). Implementation paper on analyzing COVID-19 vaccines on twitter dataset using tweepy and text blob. *Annals of the Romanian Society for Cell Biology*, 8393–8396.

Fauzi, M. A. (2018). Random Forest Approach fo Sentiment Analysis in Indonesian. *Indonesian Journal of Electrical Engineering and Computer Science, 12*, 46–50.

Hananto, A. R., Rahayu, S. A., & Hariguna, T. (2022). COVID-19 Vaccination: A Retrospective Observation and Sentiment Analysis of the Twitter Social Media Platform in Indonesia. *International Journal of Informatics and Information Systems, 5*, 56–68.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications, 23*, 11.

Jalil, Z., Abbasi, A., Javed, A. R., Khan, M. B., Hasanat, M. H., Malik, K. M., et al. (2021). COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques. *Frontiers in Public Health, 9*.

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access, 6*, 23253–23260.

Kaur, H., Ahsaan, S. U., Alankar, B., & Chang, V. (2021). A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. *Information Systems Frontiers, 23*, 1417–1429.

Kausar, M. A., Soosaimanickam, A., & Nasar, M. (2021). Public sentiment analysis on Twitter data during COVID-19 outbreak. *Int. J. Adv. Comput. Sci. Appl, 12*, 415–422.

Narasamma, V. L., & Sreedevi, M. (2021). Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework. *International Journal of Advanced Computer Science and Applications*, 440–447.

Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights, 1*, 100019.

Paliwal, S., Parveen, S., Afshar Alam, M., & Ahmed, J. (2022). Sentiment Analysis of COVID-19 Vaccine Rollout in India. In *ICT Systems and Sustainability* (pp. 21–33). Springer.

Ramamoorthy, T., Karmegam, D., & Mappillairaju, B. (2021). Use of social media data for disease based social network analysis and network modeling: A Systematic Review. *Informatics for Health and Social Care, 46*, 443–454.

Robnik-Šikonja, M. (2004). Improving random forests. *European conference on machine learning*, (pp. 359–370).

Scannell, D., Desens, L., Guadagno, M., Tra, Y., Acker, E., Sheridan, K., et al. (2021). COVID-19 vaccine discourse on Twitter: A content analysis of persuasion techniques, sentiment and mis/disinformation. *Journal of health communication, 26*, 443–459.

Shamrat, F. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., et al. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indones. J. Electr. Eng. Comput. Sci, 23*.

Singh, G. (2021). Sentiment Analysis of Code-Mixed Social Media Text (Hinglish). *arXiv preprint arXiv:2102.12149*.

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*, 204.

Zhang, L., Fan, H., Peng, C., Rao, G., & Cong, Q. (2020). Sentiment analysis methods for hpv vaccines related tweets based on transfer learning. *Healthcare*, *8*, p. 307.