Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

An Adaptive Ensemble Machine Learning Classifier for Sentiment Analysis on Twitter

Shakirah Mohd Sofi[1,2,*], Ali Selamat[2,3,4], Zatul Alwani Shaffiei[2]

[1] Jabatan Komputeran, Fakulti Multimedia Kreatif & Komputeran, Universiti Islam Selangor (UIS), 43000 Bandar Seri Putra, Kajang, Selangor, Malaysia
[2] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia
[3] School of Computing, Faculty of Engineering, & Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Skudai 81310, Johor Bahru, Malaysia
[4] Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic

ARTICLE INFO

Article history:
Received 22 March 2025
Received in revised form 15 August 2025
Accepted 20 October 2025
Available online 15 January 2026

Keywords:
Sentiment analysis; ensemble classifier; feature extraction; LDA; TF-IDF; SVM; NB; RF; DT; voting; bagging; XGBoost and stacking

ABSTRACT

Social media platforms serve as ubiquitous channels for individuals to connect, communicate, and share information in real-time across the globe. The exponential growth of social media platforms, particularly Twitter, has led to a significant increase in textual data, shaping social discourse and sentiment analysis. This abundance of data presents challenges and opportunities for understanding the dynamics of social media interactions and sentiment expression. Sentiment analysis faces challenges: sparse data limits understanding, while topic coherence and interpretability demand improvement for clearer insights. The primary goal of this paper is to improve the accuracy and effectiveness of sentiment analysis through the application of advanced techniques and classifiers. Traditional machine learning techniques often struggle to effectively capture the nuanced sentiment expressed in tweets. To address this issue, we propose a novel ensemble learning framework that dynamically adapts to the evolving characteristics of Twitter data. We experiment with baseline classifiers such as Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB) on Twitter data. Our approach combines these weak learners through ensemble methods like Voting, Bagging, XGBoost, and stacking, incorporating a meta-learner to optimize prediction performance. The experimental findings demonstrate that our innovative ensemble classifier achieves a remarkable accuracy rate, significantly surpassing that of individual classifiers. This paper contributes to the advancement of sentiment analysis techniques tailored for social media data, offering insights into the potential of adaptive ensemble learning in addressing the unique challenges posed by Twitter sentiment analysis.

[*] *Corresponding author.*
E-mail address: syakirah@uis.edu.my

https://doi.org/10.37934/ard.136.1.340357

1. Introduction

In today's digital age, social media platforms have become essential instruments for modern communication, enabling worldwide networking, information exchange, and real-time conversation [3]. Among these platforms, Twitter stands out as an effective platform for the rapid spread of news, ideas, and conversations on a wide range of issues. The rapid rise of Twitter and other social media platforms has altered not just the communication landscape, but also the contours of public debate and interpersonal interactions. Traditional research methods and analytical tools are often insufficient to handle the complexity and scale of Twitter data, necessitating the development of innovative approaches and techniques. Machine learning algorithms, natural language processing (NLP) approaches, and social network analysis methodologies have emerged as essential tools for extracting information from Twitter data and understanding the complexities of social interactions.

Twitter's unique characteristics, such as its microblogging format within the 140-character limit, real-time updates, and vast user base, have made it a rich source of data for researchers and practitioners across various domains. Researchers have increasingly turned to Twitter data to study a wide range of topics, including politics, public health, marketing, and sentiment. The brevity and immediacy of tweets offer insights into public sentiment, emerging trends, and social phenomena, making Twitter a valuable resource for understanding contemporary society analysis [15].

In recent years, sentiment analysis, a subfield of natural language processing (NLP), has gained popularity as a method of extracting insights from textual data, such as tweets, to better comprehend public opinion, emotion, and sentiment on different topics. The integration of sentiment analysis and Twitter data has opened up new paths for investigating a wide range of phenomena, including political discourse [14], consumer behaviour [22], public health trends [16,23] and social movements [9,27]. Researchers are increasingly using powerful machine learning algorithms, deep learning techniques, and computational linguistics methodologies to assess tweet sentiment and extract useful information.

The informal nature of tweets, characterized by abbreviations, slang, and grammatical variations, poses challenges for effective sentiment analysis. Therefore, preprocessing steps are essential to clean and standardize the text data. These steps typically include tokenization to break down tweets into individual words or tokens, removal of noise such as special characters, punctuation, and URLs, as well as normalization of text through techniques like lowercasing and stemming [20]. Additionally, handling specific tweet elements such as hashtags, mentions, and emoticons requires specialized processing to retain their contextual relevance. Following preprocessing, features extraction becomes crucial for representing tweet content in a format suitable for machine learning algorithms. Through meticulous preprocessing and feature extraction, researchers can effectively harness Twitter data for sentiment analysis, enabling insights into public opinion and sentiment trends.

The primary objective of this paper is to demonstrate the effectiveness of ensemble learning, particularly the stacking method, in enhancing sentiment analysis performance on Twitter data about COVID-19. By integrating multiple base learners with a meta-learner, we aim to achieve superior predictive accuracy, robustness to noise, and generalization capability compared to individual models. Additionally, we seek to investigate the complementary strengths of SVM, NB, DT, and RF in capturing the subtle aspects of tweet sentiment regarding COVID-19, leveraging their collective wisdom through ensemble learning techniques such as voting, bagging, boosting, and stacking.

The key contributions of this paper can be summarized as follows:

i. Proposing a novel ensemble framework that integrates voting, bagging, and boosting with stacking classifier techniques, combining diverse base learners tailored for tweet sentiment analysis.
ii. Conducting extensive experiments to assess the performance of the ensemble approach against individual models and traditional ensemble methods, thereby demonstrating its superiority in accuracy, robustness, and scalability.

2. Related Works

Ankit and Saleena Nabizath [2] proposed a framework utilizing base classification techniques with Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB). They introduced an ensemble classifier that combines these base learning classifiers using a voting method. Their framework leverages the Bag-of-Words technique for feature extraction, converting training tweets into numeric representations. The study classifies tweets from four different Twitter sentiment analysis datasets: the Stanford Sentiment 140 corpus, Health Care Reform (HCR), CrowdFlower, and Kaggle. This approach is particularly useful for companies to monitor consumer opinions about their products and for consumers to make informed decisions based on public opinions. The results of their research indicate that the proposed ensemble classifier outperforms standalone classifiers as well as the popular majority voting ensemble classifier. Although the primary focus of the proposed work is sentiment analysis of Twitter data, its methodology implies the potential for extension to analyse data from various other social network platforms.

Ahlam Alrehili and Kholood Albalawi [1] used an ensemble machine learning method to develop a sentiment classification model for categorizing customer reviews as positive or negative. The information was sourced from Amazon and contains reviews for electrical items such as the Kindle, Fire TV Stick, tablet, and laptop. The ensemble machine learning approach used five classifiers: Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), Bagging and Boosting. The ensemble classifier determined the category of reviews by majority voting. They ran situations through unigram, bigram, and trigram models, both with and without stop word removal. The Random Forest approach got the maximum accuracy, 89.87%, while employing unigrams with stop word removal. However, the voting algorithm performed the best in all other circumstances.

Shervin Minae *et al.*, [18] introduced a model that utilizes an ensemble of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for classifying positive and negative sentiment analysis reviews from IMDB and Stanford Sentiment Treebank2 (SST2) datasets. The model incorporates GloVe word embeddings into both the CNN and LSTM architectures for prediction. Experimental studies revealed a performance improvement with the ensemble model (CNN+GloVe, LSTM+GloVe) compared to individual LSTM and CNN models.

Vipin Jain and Kanchan Lata Kashyap [11] applied a number of NLP techniques, including text pre-processing, tokenization, data labelling, and feature extraction. They used a stack-based ensemble machine learning model to identify feelings as positive, negative, or neutral when assessing Indians' perspectives on the COVID-19 vaccination. The methodology employed Bi-Gram and TF-IDF techniques to extract features and generate word vector representations. The 5-fold cross-validation approach was used to train and evaluate seven machine learning models: Linear Regression, Decision Tree, Random Forest, AdaBoost, GradientBoost, XGB, and stack ensemble. The XGB classifier was used as the meta-classifier in this study. The suggested ensemble model outperformed all other classifiers in terms of classification accuracy, scoring 97.2%. Nevertheless, this study exclusively examines tweets portraying positive, negative, or neutral attitudes towards the Covid-19 vaccine, omitting any expressions of sadness, fear, or happiness. It underscores the importance of considering social media opinions as significant concerns for government entities and policymakers, prompting them to take proactive measures to address these apprehensions.

Sunitha *et al.*, [25] proposed methods for analyzing real-time tweets related to coronavirus. The study collected approximately 3,100 tweets from individuals in India and Europe between March 23, 2020, and November 1, 2021, sourced from GitHub.com. Data preprocessing included topic modelling using the Latent Dirichlet Allocation (LDA) technique for quantitatively analyzing the topics in the generated dataset. Following this, feature extraction was performed using Term Frequency-Inverse Document Frequency (TF-IDF), GloVe, pre-trained Word2Vec, and fastText embeddings. Twitter sentiment analysis was then conducted using an ensemble classifier that combines a Gated Recurrent Unit (GRU) and a Capsule Neural Network (CapsNet). This classifier was used to categorize user sentiments into anger, sadness, joy, and fear. The experimental results demonstrated that the proposed model achieved prediction accuracies of 97.28% for Indian users and 95.20% for European users. However, the computational complexity of the proposed method is high, especially when performing experiments with large feature lengths from GloVe, Word2Vec, and fastText embeddings. Consequently, it is essential to select relevant features to improve prediction accuracy while minimizing computational complexity.

SreeJagadeesh Malla and Alphonse P.J.A. [17] introduced an ensemble deep learning model for categorizing informative tweets. These tweets were obtained from the organizers of the WNUT 2020 Shared Task 2, which included a comprehensive Twitter corpus related to the COVID-19 pandemic. The objective of the project is to identify significant tweets during the pandemic and provide crucial information to government authorities, medical organizations, and victim services. The model leverages state-of-the-art feature extraction and sentiment analysis techniques, utilizing the TextBlob method for its advanced capabilities. The study implements a Majority Voting technique-based Ensemble Deep Learning (MVEDL) model to effectively identify COVID-19-related informative tweets. This ensemble model integrates leading-edge deep learning architectures, including RoBERTa, BERTweet, and CT-BERT, to ensure optimal performance. The ensemble model achieved an impressive accuracy of 91.75% in identifying informative English tweets about the ongoing coronavirus pandemic. However, the project faces challenges such as substantial memory requirements for corpus training and high time complexity when compared to traditional machine learning models.

This research, introduced by Anu Priya and Abhinav Kumar [21], presents a deep ensemble-based approach for detecting fake news related to COVID-19. Extensive experiments were conducted using character and word n-gram TF-IDF features in conjunction with the proposed deep ensemble model, which includes Dense Neural Network (DNN) and Convolutional Neural Network (CNN) classifiers. Additionally, eight different conventional machine learning models were implemented: (i) Support Vector Machine (SVM), (ii) Random Forest (RF), (iii) Logistic Regression (LR), (iv) K-Nearest Neighbour (KNN), (v) Naive Bayes (NB), (vi) Gradient Boosting (GB), (vii) Decision Tree (DT), and (viii) AdaBoost. Various combinations of n-gram features were used to perform the experiments. The results indicate that the ensemble model combining SVM, DNN, and CNN with character features achieved the best performance, with a weighted precision, recall, and F1-score of 0.97. While deep learning models are generally more effective for large datasets, this research utilized a relatively small dataset. However, it suggests a new approach to deep learning with features extraction suitable for analyzing text sentiments at the word level.

Rezaul Haque *et al.*, [7] developed a supervised deep learning classifier using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks to perform multi-class sentiment analysis (SA) on a dataset consisting of 42,036 Bengali social media comments. After preprocessing the data, the study employed feature extraction techniques including TF-IDF, count vectors, and word embeddings. Baseline models such as Multinomial Naive Bayes (MNB), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Decision Tree (DT), Random Forest (RF), and Support Vector Classifier (SVC) were used to classify Bengali social media comments into four categories. Several commonly used deep learning (DL) models, including LSTM, Bidirectional LSTM (Bi-LSTM), Bidirectional GRU (Bi-GRU), and Convolutional LSTM (C-LSTM), were experimentally evaluated to identify the most effective approach for multi-class sentiment analysis (SA). The proposed ensemble model, CLSTM, outperformed existing DL and baseline models across categories including acceptable, political, religious, and sexual, without experiencing overfitting. The suggested CLSTM architecture significantly enhanced sentiment analysis performance, achieving an accuracy rate of 85.8%. Additionally, they developed a web application using the Flask framework capable of implementing both their proposed model and the most effective baseline model for discerning real-life sentiments within social media comments.

Wahyu Fadli Satrya *et al.*, [24] conducted a study aiming to analyse Indonesian public sentiment towards the chief of the Indonesian National Police using Twitter data collected from August 20th to August 30th, 2022. The research explores Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF) as word embedding techniques for feature extraction, converting text into numerical representations. Ensemble learning, which combines multiple models to enhance stability and prediction performance, is employed. Two types of ensemble learning, Bagging-based and Boosting-based, are investigated, with Random Forest and Balanced Random Forest used for Bagging-based ensemble learning, and XGBoost selected for Boosting-based ensemble learning. The proposed ensemble models are integrated into a multi-level ensemble model, which combines Multinomial Naive Bayes with oversampling techniques to address imbalanced data. Experimental results demonstrate that the proposed ensemble models outperform individual models in terms of accuracy, precision, and F1-scores.

Rania Kora and Ammar Mohammed [13] proposed a meta-ensemble deep learning approach aimed at creating a powerful learner through the combination of multiple weaker learners. The approach that integrates baseline deep learning models like Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Gated Recurrent Units (GRU), with shallow meta-classifiers including Support Vector Machines (SVM), Gradient Boosting (GB), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR) as top-level meta-learners. To evaluate its performance, experiments were conducted on six sentiment benchmark datasets comprising 10,000 annotated tweets in English, Arabic, and various dialects. Results indicate that the meta-ensemble deep learning technique outperforms baseline deep learning models across all six benchmark datasets. In summary, their proposed ensemble methodology employs parallel ensemble approaches, where baseline learners are concurrently created without data dependency, and fusion methods rely on meta-learning techniques. However, challenges arise in selecting the appropriate number of models to ensure accurate predictions across diverse datasets, and computational time complexity increases with data volume.

3. Methodology

In this paper, we propose an advanced ensemble learning approach for sentiment analysis of tweets by leveraging the strengths of stacking, voting, bagging, and boosting techniques. Our primary method, the stacking ensemble classifier, combines multiple base learners, such as Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and Random Forests (RF), to create a powerful predictive model. The stacking method involves training these diverse base learners and using their predictions as inputs for a meta-learner, which in our case is the efficient and high-performing algorithm.

This section presents an in-depth overview of the methodologies employed in the sentiment analysis (SA) of tweets conducted in this study. The experimental techniques were executed utilizing hardware resources comprising an Intel i5 processor with 16GB of RAM and an Intel GPU. All experimental procedures were conducted within the Google Colaboratory environment, leveraging the capabilities of Python 3. Data processing and mathematical operations were facilitated using the Pandas and NumPy libraries, renowned for their efficiency in handling structured data and numerical computations, respectively. Machine learning classifiers were constructed using the Scikit-learn (Sklearn) libraries, which offer a comprehensive suite of tools for various machine learning tasks. Performance evaluation metrics were computed using the sklearn.metrics module, specifically through the utilization of classification reports and accuracy scores. These metrics provide a detailed breakdown of the classifier's performance, encompassing precision, recall, F1-score, and overall accuracy, thereby offering comprehensive insights into the effectiveness of the classification models.

Figure 1 depicts the proposed an advanced ensemble learning approach for this empirical study. The approach consists of six stages: 1) data preparation, 2) data pre-processing, 3) feature extraction and selection, 4) machine learning baseline, 5) stacking ensemble classifier, and 6) performance evaluation.

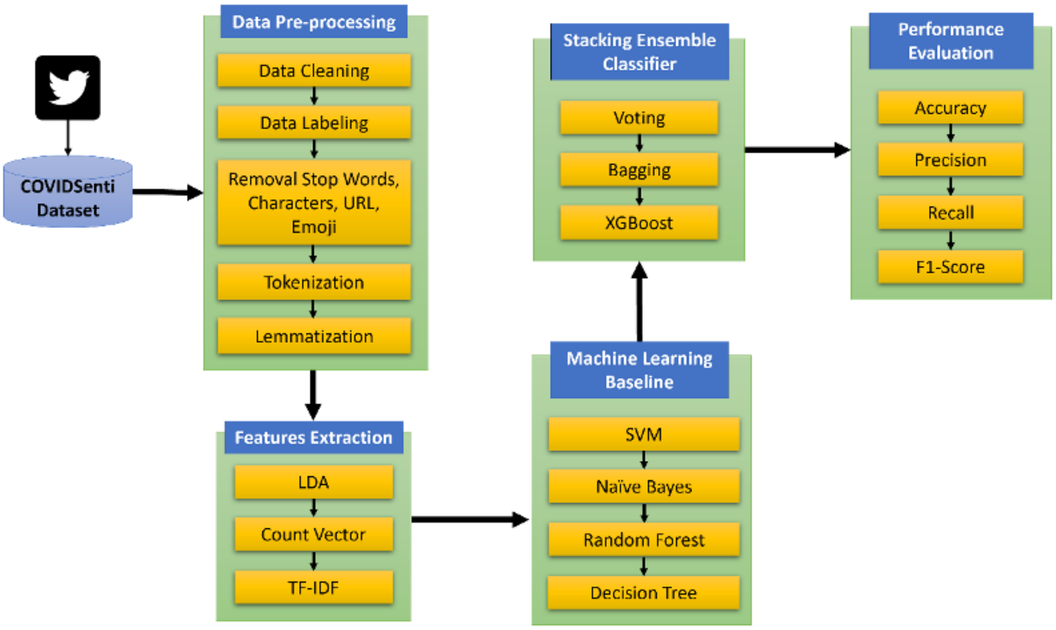

Fig. 1. Overview of the proposed approach

3.1 Dataset Preparation

In the initial phase, we obtained a dataset publicly accessible on GitHub from the CovidSenti repository. This dataset comprises 2.1 million sentiment-related tweets documented between February and March 2020 [19]. Our analysis exclusively focused on English-language tweets. Through the application of specific keywords, we meticulously filtered the dataset to ensure its relevance to Covid-19 and associated themes. The CovidSenti dataset consists of 90,000 cleaned records, which were subsequently divided into three equal subsets based on sentiment polarity: 6,280 positive, 16,335 negative, and 67,385 neutral comment reviews extracted from tweets.

A word clouds generated from the Covid-19 sentiment dataset is divided according to positive polarity sentiment (Figure 2), negative polarity sentiment (Figure 3), and neutral polarity sentiment

(Figure 4). Additionally, the distribution of polarity is illustrated in a pie chart, where positive tweets represent 7%, negative tweets represent 18%, and neutral tweets represent 75% (Figure 5).

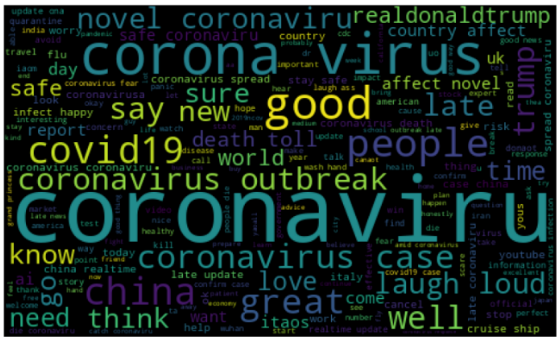

Fig. 2. Positive tweets

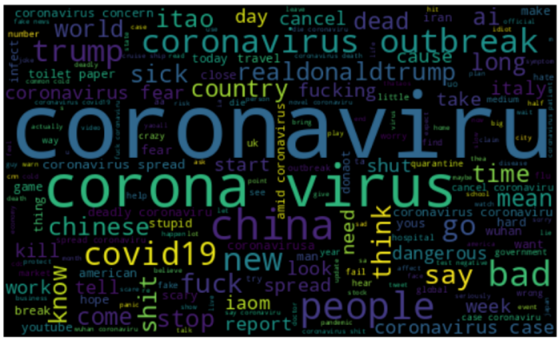

Fig. 3. Negative tweets

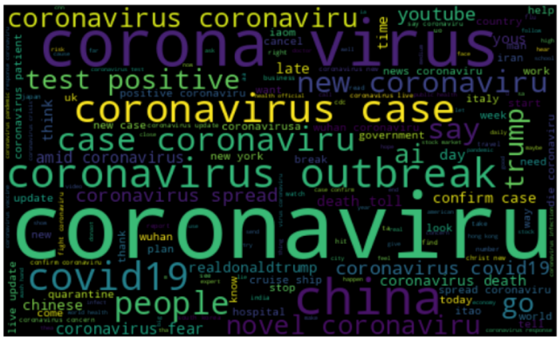

Fig. 4. Negative tweets

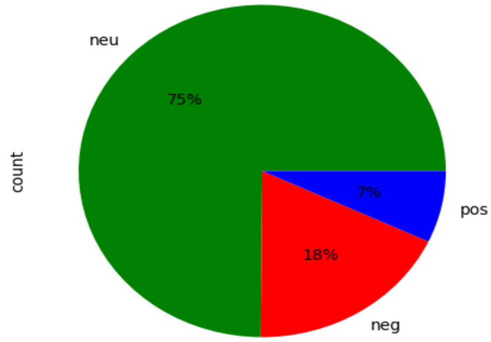

Fig. 5. Dataset distribution-based sentiment polarity

3.2 Data Pre-processing

Information gathered from social media platforms is often noisy, unstructured, informal, and diverse. The initial stage of sentiment analysis involves data pre-processing, which aims to enhance the textual content's meaningfulness. This is achieved through the following sequential strategies:

a) *Data cleaning:* Data cleaning encompasses a range of tasks aimed at improving the quality and reliability of a dataset for analysis. It involves identifying and rectifying errors, inconsistencies, and inaccuracies that may be present in the data. The steps typically include addressing missing values, removing duplicate records, standardizing variable formats, handling outliers, and normalizing data features. By performing these tasks, data cleaning ensures that the dataset is well-prepared for subsequent analysis and modeling, leading to more accurate and reliable insights.

b) *Data labeling:* Each tweet reviews in the dataset is labeled with a sentiment category: positive, negative and neutral. Then using automated tool each review is assigned a sentiment label: 1 for positive, -1 for negative, and 0 for neutral. These labeled data points serve as the training examples for building a sentiment analysis model.

c) *Removal Patterns: R*emoval patterns are commonly applied to reduce noise in textual data. This involves removing stop words and unnecessary characters that do not contribute significantly to the interpretation of the sentiment aspect of a phrase. The process typically includes converting uppercase letters to lowercase and then eliminating special characters, emojis, hyperlinks, hashtags (e.g., #StaySafe, #COVID-19), stop words (such as "for," "the," and "is"), and URLs from the dataset tweets.

d) *Tokenization:* The pre-processing pipelines involving the conversion of raw text into tokens before transforming it into vectors. Tokenization breaks down the raw text into individual words and phrases, known as tokens. This process is crucial for understanding the meaning of the text by analyzing the sequence of words. Example:
Raw Text: "The brown cat jumps over the lazy dog."
Tokens: ["The", "cat", "jumps", "over", "the", "lazy", "dog", "."]

e) *Lemmatization:* In the final step of text pre-processing, lemmatization is applied to return words to their base or root form by utilizing vocabulary and morphological analysis of sentences. Lemmatization ensures that different inflected forms of a word are mapped to a single root word, thereby reducing redundancy and standardizing the text data. (e.g., "running" to "run" or "jumping" to "jump").

3.3 Features Extraction

a) *Top Modeling with Latent Dirichlet Allocation (LDA*): LDA, or Latent Dirichlet Allocation, serves as a generative probabilistic model utilized to uncover the latent topics within a corpus of documents or text (Blei et al., 2003). It operates under the assumption that each document comprises a blend of various topics, and each individual word in a document is associated with one of these topics. In the realm of text analysis, LDA finds its utility in the automated categorization of documents into topics and the comprehension of the predominant themes embedded within a collection of documents.

b) *Count Vector:* Process to convert text data into numerical representations. It involves representing text documents as vectors where each feature (dimension) corresponds to a unique word in the vocabulary, and the value of each feature represents the frequency of that word in the document. Count Vectorization enables text data to be processed and analyzed using machine learning techniques, as it transforms the textual information into a format that machine learning algorithms can understand and operate on.

c) *Term Frequency-Inverse Document Frequency (TF-IDF)*: is a widely used technique in natural language processing (NLP) to represent the importance of a word in a document relative to a collection of documents. It combines two components: term frequency (TF) and inverse document frequency (IDF).

i. **Term Frequency (TF),** where evaluated as mention in Eq. 1 as follow:

$$TF_{t,d} = \frac{Number\ of\ times\ term\ t\ appears\ in\ tweet\ d}{Total\ number\ of\ terms\ in\ tweet\ d} \tag{1}$$

given by *d* number of tweet, *t* number of terms.

ii. **Inverse Document Frequency (IDF)**, where evaluated as mention in Eq. 2 as follow:

$$IDF_t = log\left(\frac{Total\ number\ of\ tweet}{Number\ of\ tweet\ containing\ term}\right) \quad (2)$$

given by *t* number of terms.

iii. TF-IDF Score, it is calculated by multiplying the term frequency (TF) of the term in the document by the inverse document frequency (IDF) of the term as shown in Eq. 3:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (3)$$

given by *d* number of tweet, *t* number of terms.

3.4 Machine Learning Baseline

a) *Support Vector Machines (SVM)*: Support Vector Machines (SVM) continues to be highly regarded in machine learning for its adeptness in classification tasks, owing to its effectiveness in handling high-dimensional data and capturing non-linear relationships. In recent years, SVM has found extensive application in diverse fields such as natural language processing (NLP) for sentiment analysis and document categorization, bioinformatics for protein structure prediction and disease diagnosis, image recognition for object classification, finance for financial forecasting and fraud detection, medical diagnosis for patient outcome prediction, and remote sensing for land cover classification. For instance, research by Sanjey Dey *et al*., [6] highlights SVM's efficacy in predicting feedback for Amazon book products. Their study considers factors like quality, content relevance, and timing of reviews. They demonstrate SVM's ability to analyze different aspects of customer feedback and its impact on product rankings, particularly emphasizing the importance of historical positive reviews in shaping consumer perceptions on e-commerce platforms. Similarly, Nurul Huda Zainuddin *et al*., [28] showcase SVM's success in sentiment classification analysis on hate crimes tweets. They propose a hybrid ensemble hybrid sentiment classification approach for Twitter, which includes a feature selection method. Their work underscores SVM's relevance and effectiveness in sentiment analysis, particularly at the document level, highlighting its application in addressing real-world issues.

b) *Naive Bayes (NB)*: Naive Bayes (NB) is a popular probabilistic classifier based on Bayes' theorem with a strong assumption of feature independence. Despite its simplicity, NB has found widespread application in various fields due to its efficiency, scalability, and surprisingly good performance in many real-world scenarios. Its ability to handle high-dimensional data with relatively few training examples makes it a versatile and practical choice for many machine learning applications, especially when computational resources are limited or when dealing with noisy and sparse data. Jackins *et al*., [10] embarked on an empirical investigation aimed at categorizing diverse illness datasets. To accomplish this task, they employed Naive Bayes and Random Forest classifier models trained on three distinct illness datasets: diabetes, coronary heart disease, and cancer. This work holds promise for advancing healthcare by facilitating early detection and intervention strategies for these medical conditions.

c) *Decision Trees (DT)*: A decision tree is a common machine learning technique that may be used for classification and regression. It's a tree-like structure in which each core node represents a "decision" based on a characteristic and each leaf node reflects the result or prediction. Decision trees can handle numerical and categorical data and capture complicated feature connections. Decision trees are widely utilized in a variety of disciplines, including consumer segmentation, risk assessment, and medical diagnosis, to provide useful insights into data patterns and facilitate informed decision-making. Kaveh Khalili-Damghani *et al.*, [12] used clustering techniques, data mining and DT analysis to anticipate newbie client behaviour from existing customers and identify potentially profitable ones in the insurance and telecoms industries. A novel feature selection technique identifies the most significant consumer traits, improving forecast reliability and accuracy. Dharmendra Dangi and his colleagues [5] offered one of five strategies for estimating the effect of the epidemic: using machine learning classifiers, particularly decision trees. Data collected before and after the pandemic or lockdown is fed into the algorithm to determine its performance metrics. Incorporating sentiment data from social media is crucial for making informed decisions, particularly considering the potential negative influence on national economies.

d) *Random Forests (RF):* Random Forests (RF) are a powerful ensemble learning method based on decision trees. RF builds multiple decision trees during training and combines their predictions through a process called "bagging" (bootstrap aggregating) to improve predictive accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the training data and a random subset of the features, resulting in a diverse set of trees. The final prediction of the RF is typically the mode (classification) or the average (regression) of the predictions made by individual trees. Random Forests are known for their robustness, scalability, and high accuracy, making them suitable for a wide range of applications. They have been successfully applied in fields such as finance for credit scoring and fraud detection [8], healthcare for disease diagnosis and prognosis [10], accident prediction [29] and risk assessment [4], and marketing for customer segmentation [26]. Moreover, Random Forests are effective in handling high-dimensional data, noisy data, and missing values, making them a popular choice for real-world machine learning tasks.

3.5 Ensemble Classifier

An ensemble classifier is a machine learning approach that aggregates the predictions of numerous independent classifiers to increase classification performance and accuracy. Instead of depending on a single model, ensemble approaches use the wisdom of crowds to aggregate predictions from numerous models, typically producing better results than any particular model alone. There are several types of ensemble classifiers, including:

a) *Voting*: Voting combine predictions from various classifiers, such as decision trees, support vector machines, and logistic regression, to determine the class with the most votes (mode) or calculate the average probability across all classifiers.

b) *Bagging*: Bagging is the process of training several instances of the same classifier on various subsets of training data, generally using bootstrapping (sampling with replacement). The final forecast is calculated by averaging or voting on the projections of these distinct models.

c) *Boosting*: Boosting algorithms iteratively train weak learners (classifiers that perform slightly better than random guessing) and give more weight to misclassified instances in subsequent iterations. Examples include AdaBoost (Adaptive Boosting) and Gradient Boosting.

d) *Stacking*: Stacking is the process of training numerous base classifiers before combining their predictions with a meta-classifier. The meta-classifier is trained with the basic classifier's outputs (predictions) as features.

In this study, the improved performance is attributed to the ensemble classifier's capacity to merge diverse algorithms alongside the individual predictions of Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). By integrating these predictions, the ensemble classifier strengthens its predictive capability. Additionally, the ensemble classifier harnesses the adaptive technique, which iteratively adjusts the weights of weaker classifiers to develop a more resilient classifier. Approaches such as voting, bagging, boosting, and stacking ensemble are crucial in enhancing the overall effectiveness of machine learning classifiers. They enable the ensemble model to leverage the strengths of each individual classifier while mitigating their weaknesses, ultimately leading to improved performance across various tasks, including sentiment analysis on tweets.

3.6 Performance Metrics

a) Accuracy measures the overall correctness of the model and is calculated as the ratio of correctly predicted instances to the total instances. As show in Eq.4 below:

$$Accuracy = \frac{TP+TN}{TP + FP + FN + TN} \tag{4}$$

given by,

TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

b) Precision (also called Positive Predictive Value) measures the proportion of positive predictions that are actually correct. As formulated in Eq. 5:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

c) Recall (also called Sensitivity or True Positive Rate) measures the proportion of actual positives that are correctly identified by the model. Please refer Eq. 6:

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

d) F1-Score the mean of Precision and Recall, providing a balance between the two. It is especially useful when you need a single metric to evaluate the performance of a model with imbalanced classes. Where evaluated as mention in Eq. 7 as follow:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

4. Experimental Result and Discussion

4.1 Performance Evaluation

Several experimental setups were conducted to determine the optimal model for achieving the best prediction results. In the initial experiment, individual machine learning baseline models including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB) were employed. These models were evaluated using TF-IDF feature extraction without any ensemble classifier intervention. Table 1 provides a comparison of the performance of the proposed machine learning baseline models with TF-IDF features in classifying the sentiment of tweets. Results from Table 1, along with observations from Figure 6, demonstrate that the Gaussian Naïve Bayes (NB) classifier and Support Vector Machine (SVM) both achieved an impressive accuracy of 86% during classification. However, in this scenario, the Gaussian NB classifier outperformed the SVM classifier in terms of precision, recall, and F1-score. This indicates that the Gaussian NB classifier demonstrated greater effectiveness in correctly identifying actual positives, leading to its superior performance compared to the SVM classifier.

Table 1
Performance comparison of machine learning baseline and TF-IDF classifier

Classifiers	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1 Score (%)*
TF-IDF+Gaussian NB Classifier	86	86	86	86
TF-IDF+RF Classifier	85	81	64	70
TF-IDF+DT Classifier	82	70	67	68
TF-IDF+SVM Classifier	86	80	67	72

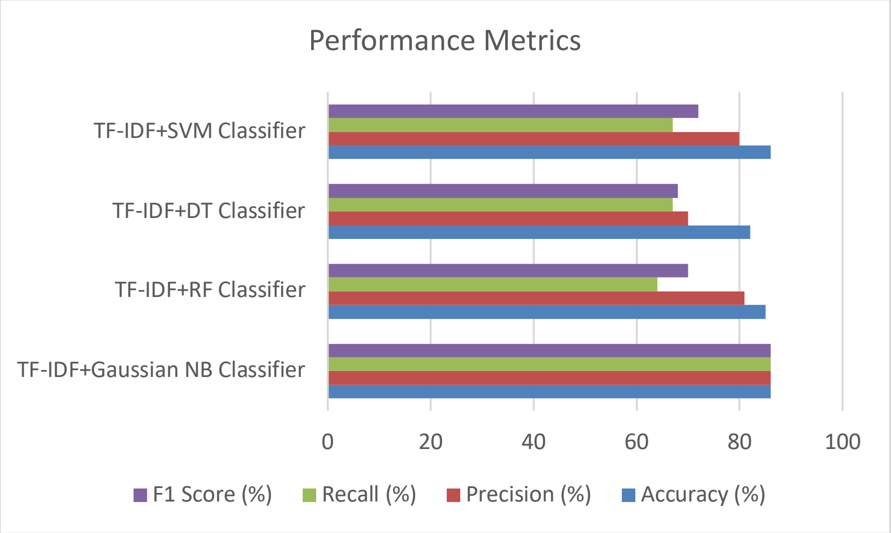

Fig. 6. Performance comparison of machine learning baseline and TF-IDF classifier

Table 2
Performance comparison of machine learning baseline with TD-IDF and ensemble classifier

Classifiers	Features	Ensemble Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Gaussian NB+Voting Classifier	TF-IDF	Hard Voting	86	86	86	86
Gaussian NB+Bagging Classifier		Bagging	86	86	86	86
Gaussian NB+XGBoost Classifier		Stacking	90	90	90	90
RF+Voting Classifier	TF-IDF	Hard Voting	90	90	90	90
RF+Bagging Classifier		Bagging	90	90	90	90
RF+XGBoost Classifier		Stacking	90	90	90	90
DT+Voting Classifier	TF-IDF	Hard Voting	90	90	90	90
DT+Bagging Classifier		Bagging	89	89	89	89
DT+XGBoost Classifier		Stacking	85	85	85	85
SVM+Voting Classifier	TF-IDF	Hard Voting	88	88	88	88
SVM+Bagging Classifier		Bagging	88	88	88	88
SVM+XGBoost Classifier		Stacking	90	90	90	90

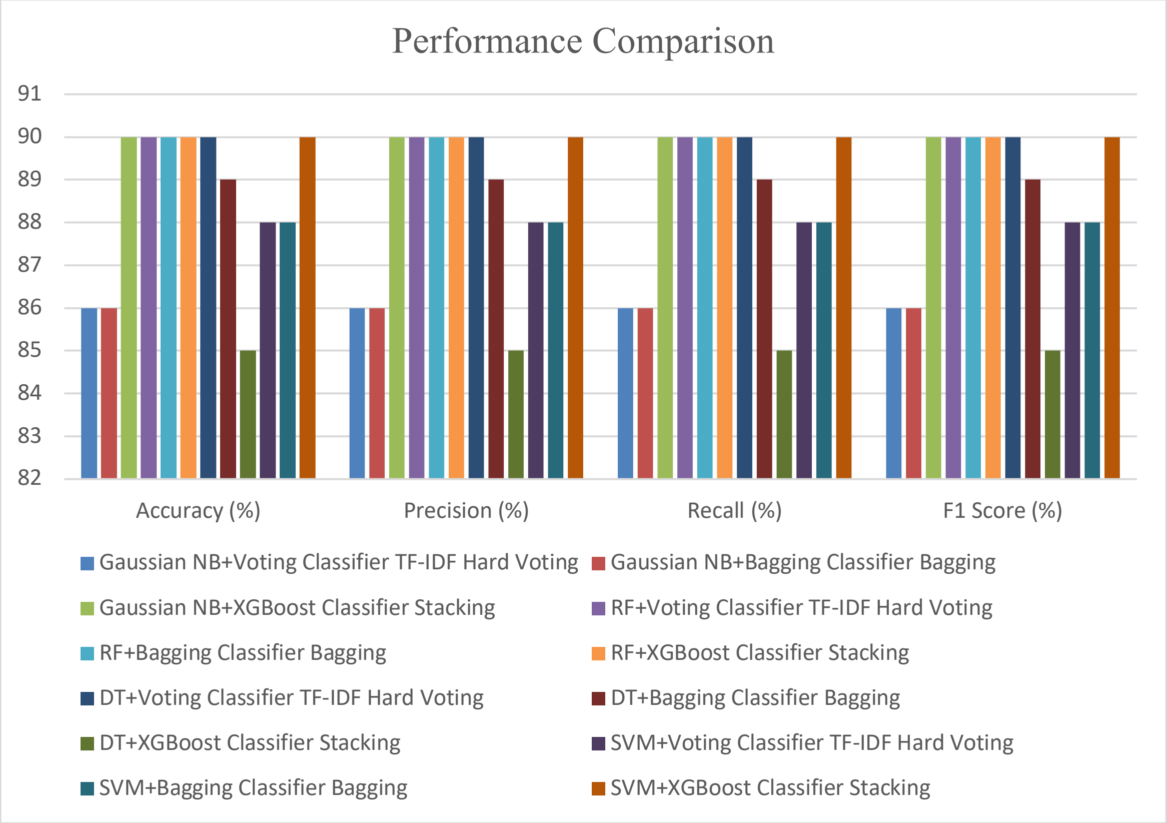

Fig. 7. Performance comparison of machine learning baseline with TD-IDF and ensemble classifier

The baseline DT classifier demonstrated a standalone accuracy of 82%. However, upon integration into an ensemble model utilizing a majority voting scheme, the accuracy experienced a remarkable improvement. Specifically, the ensemble DT with majority voting exhibited a notable accuracy increment of 8%, reaching an overall accuracy of 90%. Moreover, precision, recall, and F1-Score values also witnessed significant enhancements, further validating the compatibility of the ensemble voting technique with DT. This outcome underscores the effectiveness of ensemble methods in leveraging the complementary strengths of diverse classifiers, consequently amplifying the performance of sentiment analysis within the domain of tweets.

Upon a comprehensive examination of Table 2 and Figure 7, it becomes evident that the utilization of ensemble classifiers leads to a discernible enhancement in performance metrics, precision, recall, and F1-Score, surpassing those achieved by individual Machine Learning baselines. The ensemble approach serves to consolidate and strengthen the base learners, thereby enhancing predictive efficacy. Specifically, the combination of Gaussian NB, RF, and SVM integrated with stacking techniques employing XGBoost, yields a notable performance elevation to 90%.

Through the empirical investigation conducted, it has been empirically demonstrated that the integration of Random Forest with other ensemble learning methodologies confers a notable augmentation to its predictive efficacy. This is evidenced by the research findings, wherein all three composite approaches exhibited a predictive accuracy of 90%, thereby exceeding that achieved by the standalone Random Forest model.

Ensemble methods combining voting and bagging did not show any improvement in results for NB and SVM. NB and SVM generally exhibit lower variance compared to decision trees or other models, hence bagging does not significantly impact their performance. The majority voting mechanism may not effectively enhance the performance of NB and SVM because these models often provide consistent predictions that already align well with the overall distribution of the data.

The performance of Decision Trees (DT) in an XGBoost model with stacking ensemble techniques may only slightly increase overall accuracy from 82% to 85%, but it can lead to more significant improvements in precision, recall, and F1 score. This is because stacking can help the model capture more detailed patterns and improve its ability to correctly identify positive instances, thereby enhancing these specific performance metrics even if the overall accuracy gains are modest.

4.1 Confusion Matrix

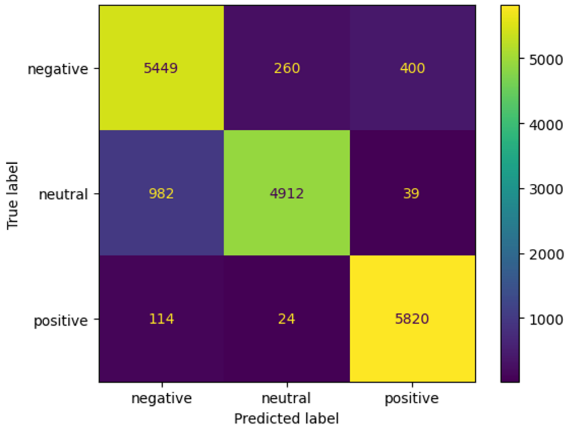

Fig. 8. Confusion matric single classifier using TF-IDF and random forest

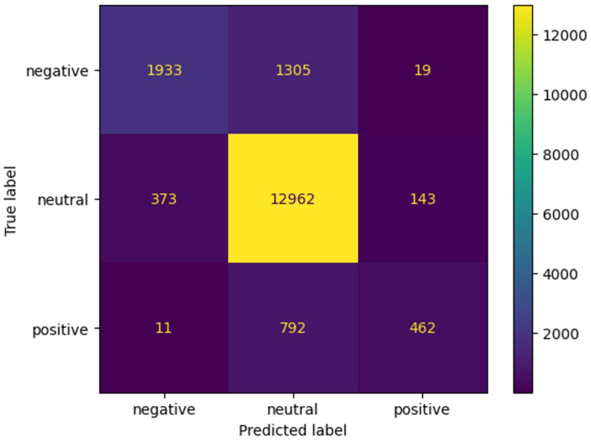

Fig. 9. Confusion matric single classifier using TF-IDF and random forest + stacking XGB classifier

A confusion matrix is a table that provides insight into the performance of a classification model by comparing its predicted labels to the true labels, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of the model's predictions. For the explanation of the confusion matrix in this experiment, we selected two sample matrices from a single machine learning classifier and an ensemble classifier, specifically from the Random Forest experimental results, to illustrate the differences between the base learner's results and the meta-learner's predictions. The confusion matrix comprises nine quadrants due to the inclusion of three features: positive, negative, and neutral. In this experiment, sentiment polarity data was labeled numerically, with 0 assigned to negative tweets, 1 to neutral tweets, and 2 to positive tweets. Twenty percent of the 90,000 dataset was used for testing in this classifier.

Utilizing TF-IDF and Random Forest in the baseline single classifier, as shown in Figure 8 means that:

i. In the first row, the model correctly classified 1,933 out of 3,257 negative reviews, misclassifying 1,305 as neutral and 19 as positive.
ii. In the second row, the model correctly classified 12,962 out of 13,478 neutral reviews, misclassifying 143 as positive and 373 as negative.
iii. In the third row, the model correctly predicted 462 out of 1,265 positive reviews, misclassifying 11 as negative and 792 as neutral.

Figure 9 show Confusion Matric Single Classifier using TF-IDF and Random Forest + Stacking XGB Classifier, explain that:

i. In the first row, the model correctly classified 5,449 out of 6,109 negative reviews, misclassifying 260 as neutral and 400 as positive.
ii. In the second row, the model correctly classified 4,912 out of 5,933 neutral reviews, misclassifying 39 as positive and 982 as negative.
iii. In the third row, the model correctly predicted 5820 out of 5,958 positive reviews, misclassifying 114 as negative and 24 as neutral.

Since accuracy measures how often the model is correct, when comparing the Random Forest (RF) classifier's accuracy of 85% to the RF with ensemble classifier's accuracy of 90%, it suggests that the ensemble classifier performs better in terms of overall correctness. Specifically, the ensemble classifier correctly predicts the class labels more frequently than the RF classifier when evaluated against the same dataset. This improvement in accuracy indicates that the ensemble classifier's combination of multiple base learners with a meta-learner has led to better predictive performance and a more reliable model.

5. Conclusion and Future Recommendation

In conclusion, the utilization of adaptive ensemble models in tweet sentiment analysis, employing machine learning and ensemble methods, has shown remarkable enhancements in predictive performance. By integrating diverse classifiers and ensemble techniques such as majority voting, bagging, and stacking with XGBoost, the ensemble approach effectively harnesses the complementary strengths of individual models, resulting in improved accuracy, precision, recall, and F1-Score metrics. The adaptable framework of the ensemble methodology allows for the creation of robust models capable of capturing subtle sentiment patterns in tweets, thereby advancing the effectiveness of sentiment analysis in this dataset.

In future research pursuits, it is imperative to continue exploring and refining ensemble techniques tailored specifically for tweet sentiment analysis. Moreover, extending the application of ensemble approaches to diverse datasets or domains beyond tweet sentiment analysis offers a fertile ground for research. By adapting and optimizing ensemble methods for various data modalities and problem domains, researchers can uncover new insights and methodologies that contribute to advancements in predictive modelling and analysis. Furthermore, the integration of deep learning techniques into ensemble frameworks presents an intriguing avenue for further exploration. By harnessing the representational power of deep learning models alongside the diversity and robustness of ensemble methods, researchers can potentially achieve even greater levels of predictive performance and generalization across a wide range of tasks and datasets. Through these efforts, future studies can advance the state-of-the-art in tweet sentiment analysis and contribute to a deeper understanding of public opinion dynamics in online discourse.

References

[1] Alrehili, Ahlam, and Kholood Albalawi. "Sentiment analysis of customer reviews using ensemble method." In *2019 International conference on computer and information sciences (ICCIS)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/ICCISci.2019.8716454

[2] Saleena, Nabizath. "An ensemble classification system for twitter sentiment analysis." *Procedia computer science* 132 (2018): 937-946. https://doi.org/10.1016/j.procs.2018.05.109

[3] Boyd, Danah M., and Nicole B. Ellison. "Social network sites: Definition, history, and scholarship." *Journal of computer-mediated Communication* 13, no. 1 (2007): 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

[4] Chen, Yanyu, Wenzhe Zheng, Wenbo Li, and Yimiao Huang. "Large group activity security risk assessment and risk early warning based on random forest algorithm." *Pattern Recognition Letters* 144 (2021): 1-5. https://doi.org/10.1016/j.patrec.2021.01.008

[5] Dangi, Dharmendra, Dheeraj K. Dixit, and Amit Bhagat. "Sentiment analysis of COVID-19 social media data through machine learning." *Multimedia tools and applications* 81, no. 29 (2022): 42261-42283. https://doi.org/10.1007/s11042-022-13492-w

[6] Dey, Sanjay, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey. "A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews." In *2020 international conference on contemporary computing and applications (IC3A)*, pp. 217-220. IEEE, 2020. https://doi.org/10.1109/IC3A48958.2020.233300

[7] Haque, Rezaul, Naimul Islam, Mayisha Tasneem, and Amit Kumar Das. "Multi-class sentiment classification on Bengali social media comments using machine learning." *International journal of cognitive computing in engineering* 4 (2023): 21-35. https://doi.org/10.1016/j.ijcce.2023.01.001

[8] Xuan, Shiyang, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. "Random forest for credit card fraud detection." In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, pp. 1-6. IEEE, 2018. https://doi.org/10.1109/ICNSC.2018.8361343

[9] Ade Iriani, Hendry, Daniel Herman Fredy Manongga, and Rung-Ching Chen. "Mining public opinion on radicalism in social media via sentiment analysis." (2020).

[10] Jackins, V., S. Vimal, Madasamy Kaliappan, and Mi Young Lee. "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes." *The Journal of Supercomputing* 77, no. 5 (2021): 5198-5219. https://doi.org/10.1007/s11227-020-03481-x

[11] Jain, Vipin, and Kanchan Lata Kashyap. "Multilayer hybrid ensemble machine learning model for analysis of covid-19 vaccine sentiments." *Journal of Intelligent & Fuzzy Systems* 43, no. 5 (2022): 6307-6319. https://doi.org/10.3233/JIFS-220279

[12] Khalili-Damghani, Kaveh, Farshid Abdi, and Shaghayegh Abolmakarem. "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries." *Applied Soft Computing* 73 (2018): 816-828. https://doi.org/10.1016/j.asoc.2018.09.001

[13] Kora, Rania, and Ammar Mohammed. "An enhanced approach for sentiment analysis based on meta-ensemble deep learning." *Social Network Analysis and Mining* 13, no. 1 (2023): 38. https://doi.org/10.1007/s13278-023-01043-6

[14] Kušen, Ema, and Mark Strembeck. "Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections." *Online Social Networks and Media* 5 (2018): 37-50. https://doi.org/10.1016/j.osnem.2017.12.002

[15] Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?." In *Proceedings of the 19th international conference on World wide web*, pp. 591-600. 2010. https://doi.org/10.1145/1772690.1772751

[16] Lim, Sunghoon, Conrad S. Tucker, and Soundar Kumara. "An unsupervised machine learning model for discovering latent infectious diseases using social media data." *Journal of biomedical informatics* 66 (2017): 82-94. https://doi.org/10.1016/j.jbi.2016.12.007

[17] Malla, SreeJagadeesh, and P. J. A. Alphonse. "COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets." *Applied Soft Computing* 107 (2021): 107495. https://doi.org/10.1016/j.asoc.2021.107495

[18] Minaee, Shervin, Elham Azimi, and AmirAli Abdolrashidi. "Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models." *arXiv preprint arXiv:1904.04206* (2019).

[19] Naseem, Usman, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim. "COVIDSenti: A large-scale

benchmark Twitter data set for COVID-19 sentiment analysis." *IEEE transactions on computational social systems* 8, no. 4 (2021): 1003-1015. https://doi.org/10.1109/TCSS.2021.3051189

[20] Pradha, Saurav, Malka N. Halgamuge, and Nguyen Tran Quoc Vinh. "Effective text data preprocessing technique for sentiment analysis in social media data." In *2019 11th international conference on knowledge and systems engineering (KSE)*, pp. 1-8. IEEE, 2019. https://doi.org/10.1109/KSE.2019.8919368

[21] Priya, Anu, and Abhinav Kumar. "Deep ensemble approach for COVID-19 fake news detection from social media." In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 396-401. IEEE, 2021. https://doi.org/10.1109/SPIN52536.2021.9565958

[22] Rana, Toqir A., and Yu-N. Cheah. "Sequential patterns rule-based approach for opinion target extraction from customer reviews." *Journal of Information Science* 45, no. 5 (2019): 643-655. https://doi.org/10.1177/0165551518808195

[23] Saha, Koustuv, John Torous, Sindhu Kiranmai Ernala, Conor Rizuto, Amanda Stafford, and Munmun De Choudhury. "A computational study of mental health awareness campaigns on social media." *Translational behavioral medicine* 9, no. 6 (2019): 1197-1207. https://doi.org/10.1093/tbm/ibz028

[24] Satrya, Wahyu Fadli, Ria Aprilliyani, and Emny Harna Yossy. "Sentiment analysis of Indonesian police chief using multi-level ensemble model." *Procedia Computer Science* 216 (2023): 620-629. https://doi.org/10.1016/j.procs.2022.12.177

[25] Sunitha, D., Raj Kumar Patra, N. V. Babu, A. Suresh, and Suresh Chand Gupta. "Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries." *Pattern recognition letters* 158 (2022): 164-170. https://doi.org/10.1016/j.patrec.2022.04.027

[26] Torizuka, Kenjiro, H. Oi, F. Saitoh, and Syohei Ishizu. "Benefit segmentation of online customer reviews using random forest." In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 487-491. IEEE, 2018. https://doi.org/10.1109/IEEM.2018.8607697

[27] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." *IEEE access* 6 (2018): 13825-13835. https://doi.org/10.1109/ACCESS.2018.2806394

[28] Zainuddin, Nurulhuda, Ali Selamat, and Roliana Ibrahim. "Hybrid sentiment classification on twitter aspect-based sentiment analysis." *Applied Intelligence* 48, no. 5 (2018): 1218-1232. https://doi.org/10.1007/s10489-017-1098-6

[29] Zhou, Xiaoyi, Pan Lu, Zijian Zheng, Denver Tolliver, and Amin Keramati. "Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree." *Reliability Engineering & System Safety* 200 (2020): 106931. https://doi.org/10.1016/j.ress.2020.106931