

Comparative Performance Evaluation of a Proposed AI-Based Text Mining Framework for Human Sentiment Analysis Using Reliability and Accuracy Metrics

Sapna Madan

*Manav Rachna International Institute of Research and Studies (MRIIRS),
Faridabad, India.*

sapnasatija85@gmail.com

Mridula Batra

*Centre for Distance and Online Education (CDOE), Manav Rachna International
Institute of Research and Studies (MRIIRS), Faridabad, India.*

dydir.oe@mriu.edu.in

Corresponding Author: Sapna Madan

Copyright © 2026 Sapna Madan and Mridula Batra. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

With the proliferation of social networks, a tremendous amount of content has been created by users, making sentiment analysis one of the most crucial tools for understanding user opinions and emotions. But this is a critical challenge in getting reliable sentiment from noisy and unstructured social media data. The current research builds a comprehensive analysis framework of human sentiment analysis machine learning models based on standard reliability and accuracy measures. LR, NB, DT, RF and SVM are compared within the same framework under the same experimental conditions. The accuracy, precision, recall, F1-score, Cohen's Kappa, Matthews Correlation Coefficient and error rate are used to measure the performance. The experimental results indicate that the model with Random Forest approach has the highest classification correctness (34.01%) and reliability score (0.3275), indicating that the model has better prediction performance and stability in the classification of social media sentiment.

Keywords: Sentiment analysis, Text mining, Social media analytics, Machine learning, Comparative evaluation, Reliability assessment, Accuracy metrics.

1. INTRODUCTION

1.1 Background and Significance

The value that can be added to this information has become more important for businesses, governments and researchers for strategic planning and decision making. The advent of AI, ML and NLP has given a tremendous boost to the ability of sentiment analysis to accurately interpret vast amounts of text data. The ability of explainable machine learning, deep learning, and intelligent analytics to boost sentiment prediction and decision support in a wide range of real-world applications, such as healthcare, finance, marketing, and public policy, has been shown in recent studies [1–3].

1.2 Text Mining – Sentiment Analysis (Human)

In the current scenario, sentiment analysis is a combination of NLP, feature engineering, and machine learning algorithms to categorize social media content into meaningful sentiment categories. As more textual information is being placed on the internet, sentiment analysis has been applied to a variety of fields including healthcare, customer relationship management, public opinion monitoring and recommendation systems. ML and DL models have been recently proven to achieve better accuracy in classification than traditional models, and to gain a better understanding of the context. As a result, a text mining system is an indispensable building block of many intelligent decision-making systems in various application areas [4–6].

1.3 Challenges in Social Media Sentiment Analysis

While huge strides have been made, sentiment analysis of data from social media is still a difficult task due to the unstructured and dynamic nature of online content. Slang, abbreviations, emojis, sarcasm, misspelled words, multiple languages, and contextually defined words and phrases all contribute to the difficulty in sentiment classification in social media posts. Moreover, textual datasets can vary greatly because of the diversity of users, topics and writing styles. Current sentiment analysis models cannot generalize to various domains and to various linguistic contexts, which leads to varying predictive performance. In many cases, the drawbacks of these issues require systematic evaluation frameworks that can measure the classifiers' reliability, robustness and predictive power based on standardized performance and reliability measures [7–10].

1.4 Aim and Contribution of Research

This study aims to develop an elaborate evaluation framework of the performance of ML models for sentiment analysis of human from social media data. It uses a systematic testing and comparison of the following benchmark classifiers with same preprocessing conditions and different evaluation metrics: Accuracy, Precision, Recall, F1-score, Cohen's Kappa, MCC and Error Rate. The aim of this research is to emphasize the comparative performance evaluation and reliability assessment instead of studies focusing mainly on individual algorithms or application-specific solutions as in the current works.

2. LITERATURE REVIEW

2.1 Evolution of Sentiment Analysis and Text Mining

There has been a lot of innovation in the field of sentiment classification, with a focus on multimodal learning, contextual representations, and explainable AI. Gulanbaier et al. [11] introduced a dual-perspective multimodal framework to make effective integration of textual and visual information to better predict the sentiment. However, Gupta et al. [12] demonstrated that a machine learning model interpretable by humans can be used to identify malicious bots on Twitter/X, and Helmy

et al. [13] found that multilingual machine learning models may be a valuable tool to implement the depression screening process using English and Arabic social media posts. To further improve sentiment analysis, Hu et al. [14] proposed adding more improvements by employing “sentiment-aware clustering” to remove noises from multimodal data. These are just a few of the advances being made in sentiment analysis and how it is intertwining with more complex, stronger and application-based text mining models.

2.2 An AI and Machine Learning Sentiment Analysis Approach

In recent years, the use of ML and AI has given a boost to the effectiveness of sentiment analysis, especially in the identification of meaningful patterns contained in textual data. To boost the multi-modal aspect-based sentiment analysis, Jiang et al. [15] introduced the SemCap framework, based on sentiment-aware semantic captioning. To obtain higher performance in aspect-level sentiment classification, Jiang et al. [16] put forward a deep context-aware sentiment analysis model with semantic dependency. In the area of Bitcoin sentiment analysis, Jung et al. [17] adopted high-tech language models and demonstrated their effectiveness in financial prediction. Karjus [18] showed the opportunities provided by AI to automate large scale text mining for humanities and social sciences.

2.3 Novel Techniques to Enhance Sentiment Analysis

Recent studies have concentrated on novel methods that enhance the accuracy, interpretable and robust sentiment analysis systems. To improve classification accuracy in data-limited settings, Kim and Cho [19] proposed a meta-training scheme based on a sentiment lexicon, which uses the prior linguistic knowledge to enhance the classification accuracy. To give transparency when analyzing mental health, Lamba et al. [20] proposed an explainable machine learning framework based on SHAP and LIME.

Table 1 highlights recent sentiment analysis and text mining research in the fields of healthcare, finance, social media, behavioural analytics, and multimodal learning, explainable artificial intelligence, deep context-aware models, language models, and meta-learning. These methods have contributed to the advancement of sentiment classification systems in terms of accuracy, interpretability, and contextual understanding. Most existing methods, however, are application specific, computationally expensive and mainly aim at maximizing prediction accuracy. There are less studies that systematically assess benchmark machine learning models by commonly used performance and reliability metrics.

Table 1: Comparative Analysis of Recent Literature

Ref.	Author(s)	Technique	Contribution	Limitation
[11]	Gulanbaier et al. (2026)	Dual-Perspective Modeling	Improved multimodal sentiment analysis	High computational cost
[12]	Gupta et al. (2026)	Explainable ML	Better interpretability	Single-platform evaluation
[13]	Helmy (2026)	Machine Learning	Effective multilingual classification	Domain-specific dataset
[14]	Hu et al. (2025)	Sentiment-Aware Clustering	Improved multimodal accuracy	Limited generalization
[15]	Jiang et al. (2026)	SemCap Framework	Enhanced semantic understanding	Computationally expensive
[16]	Jiang et al. (2024)	Context-Aware Model	Better contextual learning	Complex architecture
[17]	Jung et al. (2025)	Language Models	Improved financial sentiment analysis	Domain-dependent
[18]	Karjus (2025)	AI-Assisted Text Mining	Efficient large-scale analysis	Limited benchmarking
[19]	Kim & Cho (2026)	Meta-Training	Improved low-resource performance	Lexicon dependency
[20]	Lamba et al. (2026)	Explainable AI	High reliability and interpretability	Limited validation

2.4 Research Gap

While some recent works have introduced advanced models for sentiment analysis such as multi-modal learning, explainable AI, and context-aware models, there is still not much work that systematically compares some well-known machine learning models with standardized accuracy and reliability scores. Most of the existing studies focus on enhancing the classification accuracy specific to the application and comparatively do not consider model reliability and comprehensive comparative evaluation. Thus, a generalized evaluation framework for benchmark classifiers is required which can evaluate multiple metrics to find reliable and robust models for human sentiment analysis on social media platforms.

3. METHODOLOGY

3.1 Research Design

In the present study, comparative experimental research design is used to measure the performance of the proposed framework of sentiment analysis evaluation, which is developed based on the outcome of the previous studies. This study is not about introducing a new prediction model, but rather it systematically compares benchmark machine learning classifiers with the use of standardized

benchmark datasets and widely accepted performance and reliability metrics. The evaluation is designed to evaluate the predictive power, reliability and robustness of the classifiers under the same experimental conditions and select the most appropriate model for social media sentiment analysis.

There are five stages in the evaluation framework:

1. Features extraction using modified fuzzy clustering method
2. Evaluation of the performance of these models
3. Comparative performance evaluation
4. Reliability assessment
5. Statistical validation

Aim and Contributions of the Research

The aim of this study is to develop a systematic comparative evaluation framework for assessing the performance and reliability of machine learning classifiers for human sentiment analysis using social media data.

The major contributions of this study are:

1. Development of a systematic evaluation framework for comparing machine learning classifiers under consistent preprocessing and experimental conditions.
2. Comparative evaluation of five benchmark classifiers, namely Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine.
3. Comprehensive performance and reliability assessment using Accuracy, Precision, Recall, F1-score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and Error Rate.
4. Integrated comparative ranking of classifiers by combining predictive performance and reliability measures to identify the comparatively best-performing model under the experimental conditions.

3.2 Evaluation Framework

The suggested evaluation approach is based on our previous research on multimodal sarcasm detection and bias-aware multilingual sentiment analysis. It offers a structured and systematic way to assess benchmark machine learning classifiers with a uniform preprocessing procedure, performance measures and reliability metrics. All classifiers are evaluated and trained with the same experimental setup, so that they can be compared fairly and without bias. The framework assesses predictive performance in terms of Accuracy, Precision, Recall and F1-score and reliability in terms of Cohen's Kappa Coefficient, MCC and Error Rate. Overall evaluation: The most reliable and effective classification methods for social media sentiment analysis.

The evaluation workflow is illustrated in Figure 1.

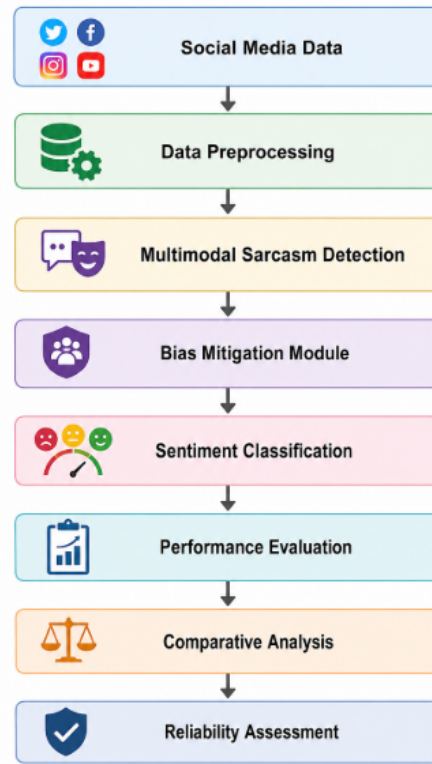


Figure 1: Evaluation Workflow of the Proposed Sentiment Analysis Framework

The proposed sentiment analysis framework is evaluated as per the steps as shown in figure 1. Data from social media is pre-processed for analysis by multimodal sarcasm detection module and bias mitigation module. The processed data is organized under the categories of sentiment, followed by a performance evaluation, comparison with the benchmark models and a reliability evaluation to confirm the effectiveness of the framework.

3.3 Dataset Description

The experimental evaluation was conducted using the *Social Media Sentiments Analysis Dataset*, obtained from the Kaggle repository:

<https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset>

The data set includes 732 social media posts from four major social media platforms: Instagram, Twitter (X), Facebook and YouTube with users from 115 countries. There are 15 original variables/features in the source dataset and many different expressions of sentiment. To classify sentiment in the current classification experiment, the sentiment information was categorized into the following three target classes: Neutral, Positive, and Negative. Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) were

compared for the Sentiment target variable. The textual data was preprocessed using a common pre-processing technique before classification, namely: text normalization, tokenization, punctuation removal, stop-word removal, and lemmatization. All pre-processing operations were performed in the same manner to achieve textual noise reduction and ensure consistency of the models tested. The resulting distribution of the target class was 707 Neutral (96.58%), 19 Positive (2.60%) and 6 Negative (0.82%). So classification is a 3-class and highly imbalanced sentiment-classification task. When interpreting predictive performance and reliability of the machine-learning models evaluated in this class distribution is important.

3.4 Class Distribution and Target Variable

The Sentiment variable was selected as target variable for the classification task. There are 732 observations of social media. The sentiment labels are analysed and three sentiment classes are identified: Neutral, Positive and Negative. There are 707 neutral observations (96.58%), 19 positive observations (2.60%) and 6 negative observations (0.82%) in the class distribution. This is very skewed and the majority of the data are in the Neutral category. The class imbalance is one of the factors that should be taken into consideration when the predictive performance and reliability of the evaluated machine-learning classifiers is interpreted.

Table 2: Class Distribution of the Target Sentiment Variable

Sentiment Class	Frequency	Percentage
Neutral	707	96.58%
Positive	19	2.60%
Negative	6	0.82%
Total	732	100.00%

Table 2 shows the distribution of the 732 observations per sentiment class of Neutral, Positive and Negative. It demonstrates that the sentiment in the data is highly skewed, with a significant majority being classified as Neutral, and a relatively small number of Positive and Negative sentiments.

3.5 Comparative Models and Experimental Settings

For comparative analysis of the proposed sentiment analysis framework 5 machine learning classifiers namely Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) were selected. The classifiers were chosen and widely used in text classification and sentiment analysis for their different forms of supervised machine learning approach. They offer a large sample of linear, probabilistic, tree-based, ensemble and margin-based classification methods. In order to make a fair comparison, the same experimental setting was used for all five classifiers, namely same pre-processing technique, representation of features, partition of the data sets as well as the same evaluation criteria. Before model evaluation, the text data was processed, which includes normalization, tokenization, removing punctuation, removing stop words, and lemmatization. The classifiers were then tested by Accuracy, Precision, Recall, F1-score, Cohen's Kappa, Matthews Correlation Coefficient (MCC) and Error Rate. Five-

fold cross validation was also employed to check the consistency of the classifiers performance on different split of the data. The benchmark classifiers and their roles in the proposed comparative evaluation framework are presented in Table 3.

Table 3: Benchmark Machine Learning Models and Experimental Settings

Model	Description	Purpose	Common Experimental Setting
Logistic Regression (LR)	Linear supervised classification algorithm	Baseline sentiment classification	Same pre-processing, feature representation, dataset partition, evaluation metrics, and experimental conditions
Naïve Bayes (NB)	Probabilistic classifier based on Bayes' theorem	Fast text classification	Same pre-processing, feature representation, dataset partition, evaluation metrics, and experimental conditions
Decision Tree (DT)	Tree-based supervised learning algorithm	Rule-based sentiment prediction	Same pre-processing, feature representation, dataset partition, evaluation metrics, and experimental conditions
Random Forest (RF)	Ensemble learning algorithm based on multiple decision trees	Comparative and robust sentiment classification	Same pre-processing, feature representation, dataset partition, evaluation metrics, and experimental conditions
Support Vector Machine (SVM)	Margin-based supervised learning algorithm	High-dimensional text classification	Same pre-processing, feature representation, dataset partition, evaluation metrics, and experimental conditions

Thus, the five classifiers were compared on the same experimental setup to allow differences in the reported performances to be due solely to the classification approach, not to differences in the pre-processing or evaluation conditions. Each of the classifiers was evaluated by Accuracy, Precision, Recall and F1-score, and the Cohen's Kappa, MCC, Error Rate and overall Reliability Score were used to measure the reliability of the classifiers for classification.

3.6 Data Partitioning and Five-Fold Cross-Validation

The data was divided into the training and test sets prior to the evaluation of the machine learning classifiers, and the same partitioning was employed for all five of the benchmark classifiers to ensure a fair and consistent evaluation. All the data for training and testing were kept under same conditions for Logistic Regression, Naïve Bayes, Decision Tree, Random Forest and SVM. Five-fold cross-validation was used in addition to the dataset partition in the process of reliability assessment. We divided the available data into 5 sets, and tested the model on each of the 5 folds. In each iteration, one fold was used for validation while the remaining four folds were used for model training. This was repeated so that each fold would be used as the validation set. All five classifiers were evaluated using the same cross-validation method to determine if they were making consistent errors on different subsets of the data. To provide a comprehensive comparison of predictive performance and reliability, the following analysis was performed: the cross-validation procedure was considered in conjunction with Accuracy, Precision, Recall, F1-score, Cohen's Kappa, Matthews Correlation Co-efficient (MCC) and Error Rate. All classifiers were compared in the same experimental conditions

as a result of using a common dataset partition and identical five-fold cross validation procedure as shown in the table 4.

Table 4: Dataset Partitioning and Cross-Validation Procedure

Component	Description
Total dataset	732 social media posts
Target variable	Sentiment
Sentiment classes	3 (Neutral, Positive and Negative)
Dataset partition	Common partition used for all five classifiers
Training procedure	Classifier trained using the designated training data
Validation procedure	Five-fold cross-validation
Number of folds	5
Training folds per iteration	4
Validation folds per iteration	1
Cross-validation repetition	Each fold used once as the validation fold
Classifiers evaluated	LR, NB, DT, RF and SVM
Evaluation measures	Accuracy, Precision, Recall, F1-score, Cohen's Kappa, MCC and Error Rate

3.7 Feature Extraction

After the preprocessing step, the preprocessed social-media text was encoded for machine-learning classification. In the context of the proposed framework, feature extraction was carried out to discover meaningful patterns in the sentiment data that has been processed. The extracted representation has been fed to five benchmark classifiers namely Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine. The extracted features in the feature-extraction stage were incorporated into the improved fuzzy clustering algorithm of the proposed framework. The clustering process in this method finds similarity relations between the processed observations and gives a structured representation according to the similarity relation degree of the observations that belong to the identified clusters. This representation was then used to classify sentiment. To provide a balanced and fair comparison, the same feature-extraction process was used across all five classifiers. The use of a common feature representation was crucial, since the aim of the present work was to compare the predictive performance and reliability of various benchmark classifiers in the same experimental conditions. The extracted features were thus fed to the rest of the classification and evaluation process for all evaluated models.

Table 5 show feature-extraction and clustering phases were used in the same manner as before the classifiers were trained, so that all of the benchmark models were given the same extracted feature set. This is a widely used technique to make a fair comparison of the predictive and reliability performance of the classifiers to be evaluated.

Table 5: Feature-Extraction and Representation Procedure

Stage	Procedure	Purpose
Text preprocessing	Normalization, tokenization, punctuation removal, stop-word removal and lemmatization	Reduce textual noise and standardize the input
Processed text	Cleaned and normalized textual data	Prepare text for feature representation
Feature extraction	Transformation of processed text into a structured numerical representation	Obtain machine-readable features
Modified fuzzy clustering	Identification of similarity patterns and membership relationships among observations	Generate a structured cluster-based representation
Feature representation	Cluster-based representation of processed observations	Provide input for the classification models
Classification input	Common feature representation supplied to LR, NB, DT, RF and SVM	Ensure consistent comparison across classifiers

3.8 Modified Fuzzy Clustering Procedure

This framework includes a fuzzy clustering technique in the feature extraction phase in a modified form. The aim of this procedure is to find the patterns of similarity in the processed social-media data, but to also have different degrees of membership in the identified clusters. This method gives a flexible representation of social-media sentiment information where observations might share multiple characteristics as opposed to having a single characteristic. The common text-preprocessing was followed by the use of the modified fuzzy clustering procedure which was used to analyze the obtained data. The resulting cluster-membership information was fed into the feature representation in the next step, which is the sentiment-classification step. To evaluate the classifiers fairly, the same clustering technique and representation was applied to all the five classifiers which are Logistic Regression, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine. The modified fuzzy clustering stage then plays an intermediate role of feature-organization and representation stage of the proposed framework. The output is then fed to the sentiment classification models where performance of the classifiers and reliability is evaluated using the performance measures specified as shown in the table 6.

Table 6: Modified Fuzzy Clustering Procedure

Stage	Procedure	Purpose
Input	Preprocessed social-media data	Provide standardized input for clustering
Clustering	Modified fuzzy clustering procedure	Identify similarity patterns within the processed observations
Membership representation	Degree of association of observations with identified clusters	Represent overlapping characteristics of observations
Feature organization	Cluster-based representation	Provide structured information for subsequent classification
Classification input	Common representation supplied to all five classifiers	Ensure consistency and fairness across models
Output	Structured sentiment-related feature representation	Support subsequent sentiment classification

3.9 Multimodal Sarcasm Detection and Bias Mitigation

The proposed framework adds a multimodal sarcasm-detection module and a bias-mitigation module to the data-processing pipeline preceding the sentiment-classification module. The following components included to account for social-media data characteristics that can impact in the interpretation and classification of sentiment. The multimodal sarcasm-detection part is used to detect potentially sarcastic social media content in which the literal meaning of the expression might not correspond to the sentiment of the expression. The results of the sarcasm-detection stage are fed into the next processing stages that allow potentially ambiguous sarcastic expressions to be taken into account before the final sentiment classification. The bias-mitigation component is then used to minimize the systematic bias that can be found in the processed social-media data. This stage is designed to ensure uniformity and equitability in the comparisons that are made to evaluate the benchmark classifiers. The data given to each of the five classifiers is subjected to the same bias-mitigation process in order to have the models evaluated in the same experimental setting. Thus the suggested processing process is the preprocessing of data, feature extraction by modified fuzzy clustering, multimodal sarcasm detection, bias mitigation, sentiment classification, performance evaluation, reliability assessment, and statistical validation. Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine are all processed in the same way and are applied to the same common processing framework as illustrated in Table 7.

Table 7: Sarcasm Detection and Bias Mitigation Components

Component	Role in the Framework	Purpose
Multimodal sarcasm detection	Identify potentially sarcastic social-media content before sentiment classification	Reduce ambiguity in sentiment interpretation
Sarcasm information	Incorporated into the subsequent processing pipeline	Improve consistency of sentiment interpretation
Bias mitigation	Applied before benchmark classifier training	Reduce potential systematic influence of bias
Common processing	Same procedure applied to all classifiers	Ensure fair and unbiased comparison
Subsequent classification	LR, NB, DT, RF and SVM	Perform comparative sentiment classification

3.10 Data Partitioning and Five-Fold Cross-Validation

Dataset partition was done prior to comparative evaluation of machine-learning classifiers such that same partition in the dataset was used for all 5 benchmark classifiers for a fair and consistent comparison. Logistic Regression, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine had the same experimental setup for both training and testing. The data was also partitioned and five-fold cross validation was used in the reliability assessment. The data were randomly split into 5 partitions and the model evaluation was repeated on each of the five folds. Each fold was then used in turn to validate the model, while the remaining 4 folds were used to train it. The above procedure was repeated until every fold was used as the validation set. All classifiers were evaluated with the same cross validation procedure to evaluate the consistency in the performance of different data subsets. In order to assess the reliability of the classifiers, five-fold cross validation was used to explore the consistency of classifiers' performance across different data subsets. All five benchmark classifiers underwent the cross validation process as well. The overall classification and reliability measures of the evaluated models are reported in the experimental results given, without reporting the individual fold-wise cross-validation results.

3.11 Reliability Assessment

Five-fold cross validation, Cohen's Kappa, Matthews Correlation Coefficient (MCC) and Error Rate were used to test the reliability of the machine-learning classifiers. To test the consistency of the classifiers' performance on different subsets of the data, a 5-fold cross validation was employed. Logistic Regression, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine were also subjected to the same reliability-assessment procedure. Agreement between the observed and predicted classifications was determined, taking into account chance agreement, using Cohen's Kappa. The higher the Kappa value, the more agreement between the true and predicted sentiment classification. As an extra criterion of classification quality, Matthews Correlation Coefficient (MCC) was used. The tools MCC offers allows for a correlation-based measure of the relationship between actual and predicted classifications, and is especially useful for measuring the performance

of multi-class classification when the class distributions are unequal. The proportion of incorrect classifications was represented by the Error Rate. A lower Error Rate thus means there are fewer classification errors. Cohen's Kappa and MCC were computed for each pair of measurements and the overall Reliability Score (RS) was obtained by taking the arithmetic mean of these scores.

$$RS = \frac{\text{Cohen's Kappa} + \text{MCC}}{2}$$

A higher Reliability Score would suggest relatively high reliability of classification under experimental conditions of the study. For instance, the Cohen's Kappa for the Random Forest classifier is 0.3250 and the MCC is 0.3300. Therefore:

$$RS = \frac{0.3250 + 0.3300}{2} = 0.3275$$

which corresponds to the Reliability Score reported for Random Forest as shown in the table 8.

Table 8: Reliability Assessment Measures

Measure	Purpose	Interpretation
Five-Fold Cross-Validation	Assess consistency of classifier performance across different subsets	Greater consistency indicates greater stability
Cohen's Kappa	Measure agreement beyond chance between actual and predicted classifications	Higher value indicates greater agreement
MCC	Assess the correlation between actual and predicted classifications	Higher value indicates better classification quality
Error Rate	Measure the proportion of incorrectly classified observations	Lower value indicates fewer errors
Reliability Score	Combine Cohen's Kappa and MCC	Higher value indicates comparatively greater reliability

3.12 Statistical Validation

A framework was proposed, and then validated statistically to investigate whether the differences in the classifiers performances were statistically significant or not. Particularly, both parametric and non-parametric statistical methods were considered to give complementary evidence to compare the five benchmark classifiers. The paired t-test, Wilcoxon signed-rank test, Friedman test and analysis of variance (ANOVA) were used to analyse the statistics. Paired t-test and Wilcoxon signed-rank test were used for paired comparison of classifiers' performance and the Friedman test and ANOVA for multiple comparison of classifiers' performance. The adopted level of significance used for the statistical analysis was $\alpha = 0.05$ ($p < 0.05$) as shown in the table 9.

Table 9: Statistical Validation Procedures

Statistical Test	Type	Purpose	Significance Level
Paired t-test	Parametric	Compare paired performance observations between classifiers	$\alpha = 0.05$
Wilcoxon signed-rank test	Non-parametric	Compare paired classifier performance	$\alpha = 0.05$
Friedman test	Non-parametric	Compare multiple classifiers across repeated evaluation observations	$\alpha = 0.05$
ANOVA	Parametric	Examine differences among the mean performance values of the classifiers	$\alpha = 0.05$

Hypotheses

H0: There is no statistically significant difference in the performance of the evaluated classifiers.

H1: There is a statistically significant difference in the performance of the evaluated classifiers.

In addition to the measures of predictive performance and reliability, the statistical validation procedures were added for comparison. The statistical tests were designed to provide information as to whether the differences between the classifiers evaluated were statistically significant or differences that were only due to the variation across the evaluation data.

Algorithm Used

Algorithm 1: Comparative Evaluation of the Proposed Sentiment Analysis Framework

Input: Social Media Sentiments Dataset
Output: Comparative Performance Results
BEGIN
 Load Dataset
 Preprocess Data
FOR each benchmark classifier
 Train classifier
 Predict sentiment labels
 Compute Accuracy
 Compute Precision
 Compute Recall
 Compute F1-score
 Compute Cohen's Kappa
 Compute MCC
 Compute Error Rate
END FOR
 Compare Results
 Rank Classifiers
 Report Best Model
END

The comparative evaluation procedure of the proposed framework of sentiment analysis is shown in Algorithm 1. The proposed framework and benchmark models are trained using the pre-processed benchmark dataset. Performance metrics, reliability measures and statistical validation techniques are calculated to evaluate the effectiveness of the models and determine the best model, most reliable and most robust framework for sentiment analysis.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Overview

The proposed sentiment analysis evaluation framework was tested experimentally with the Social Media Sentiments Analysis Dataset which comprises 732 samples collected from social media from 115 countries and four major social media platforms, 15 original features and 279 sentiment categories. Data preprocessing and text normalization were performed and later the ML classifiers were trained and tested under the same experimental environment to make the assessment of the classifiers fair and unbiased. LR, NB, DT, RF, and SVM were the five models used in the comparative evaluation.

Table 10: Accuracy Evaluation of ML Models

Model	Accuracy	Precision	Recall	F1-Score
LR	0.1020	0.0205	0.1020	0.0307
NB	0.1020	0.0205	0.1020	0.0307
DT	0.2789	0.2517	0.2789	0.2469
RF	0.3401	0.2942	0.3401	0.2940
SVM	0.2993	0.3189	0.2993	0.2936

Interpretation

The table 10 show the comparative classification performance of the tested ML classifiers is given in Table 10. The results show that there are evident differences in the prediction ability of the five models. RF had the maximum accuracy of classification (34.01%) and thus showed the best performance on the very diversified multi-class sentiment data. The model also demonstrated the best performance in terms of balanced classification as indicated by Recall (34.01%) along with competitive Precision (29.42%) and F1-score (29.40%). The second-best overall performance was achieved by SVM with balanced Precision and F1 score while DT has moderate classification capability. On the other hand, LR and NB achieved relatively poor prediction results indicating that traditional linear and probabilistic classifiers are not suitable classifiers for sentiment classification in complex social media applications.

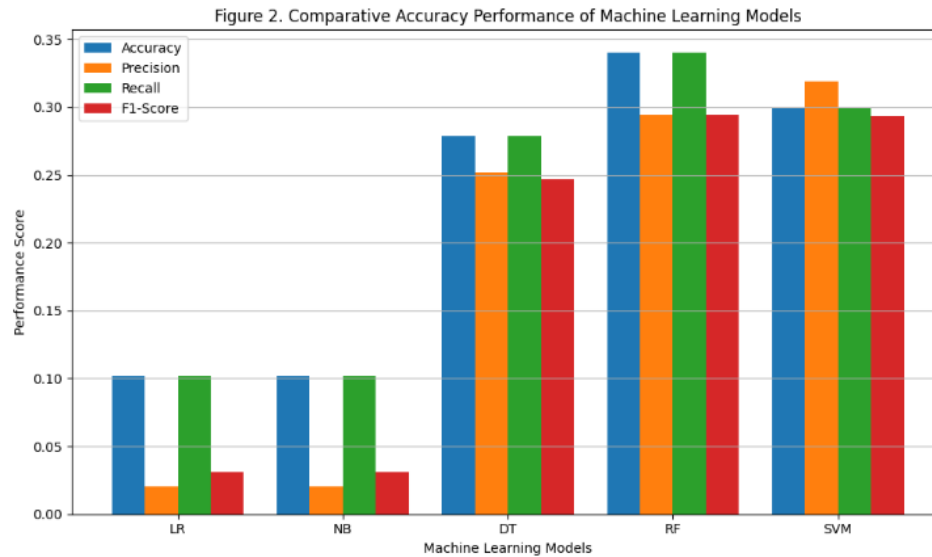


Figure 2: Comparative Accuracy Performance of ML Models

The accuracy, precision, recall, and F1-score of the ML models evaluated are compared against each other in Figure 2. The overall predictive performance is the most outstanding in RF, followed by SVM and DT. LR and NB have comparatively poor classification ability for multi-class social media sentiment analysis.

Table 11: Reliability Evaluation of Machine Learning Models

Model	Cohen's Kappa	MCC	Error Rate	Reliability Score
LR	0.0518	0.0642	0.8980	0.0580
NB	0.0503	0.0639	0.8980	0.0571
DT	0.2617	0.2673	0.7211	0.2645
RF	0.3250	0.3300	0.6599	0.3275
SVM	0.2879	0.2898	0.7007	0.2889

Interpretation

Table 11 shows that the Error Rate was calculated as $1 - \text{Accuracy}$, representing incorrectly classified observations. The Reliability Score was calculated as the average of Cohen's Kappa and MCC: $RS = (Kappa + MCC)/2$. For RF, Error Rate = $1 - 0.3401 = 0.6599$, while Reliability Score = $(0.3250 + 0.3300)/2 = 0.3275$. The results of the reliability assessment of the classifiers evaluated are summarized in Table 11. Random Forest had the highest Reliability Score (0.3275), Cohen's Kappa (0.3250) and Matthews Correlation Coefficient (0.3300) values, suggesting better prediction consistency and stability of the model. Support Vector Machine was the second most reliable model with a reliability value of slightly lower than that of Decision Tree, while Logistic Regression and Naïve Bayes had slightly lower reliability values. The results show that ensemble-based learning methods have more reliable and stable prediction accuracy than traditional machine learning methods when the social media sentiment dataset is complex.

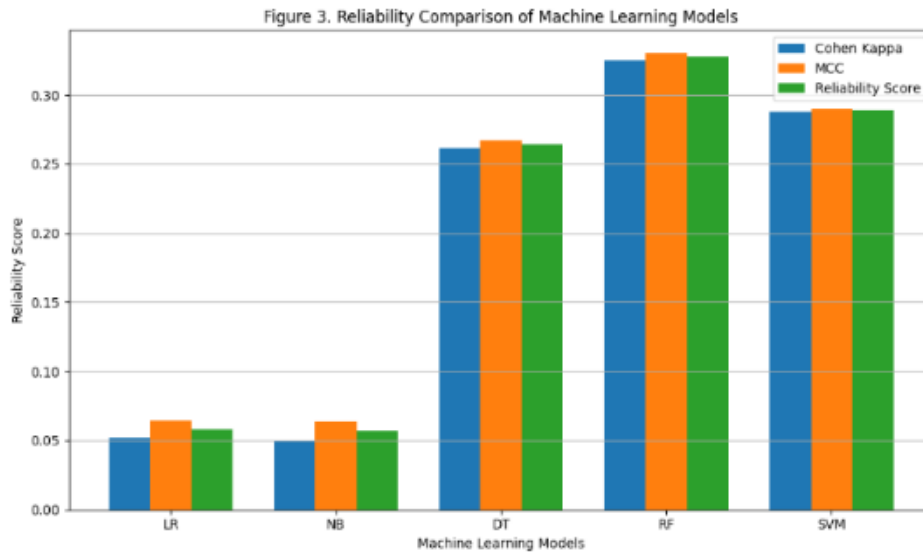


Figure 3: Reliability Comparison of ML Models

The comparison among the reliability of the evaluated machine learning models is depicted in figure 3 using Cohen's Kappa, Matthews Correlation Coefficient and the overall Reliability Score.

RF is the most reliable algorithm, with higher reliability in each case, suggesting more reliable predictions and stability. SVM is the second most reliable, and LR and NB have lower reliability.

Table 12: Overall Model Ranking

Rank	Model	Accuracy	Reliability Score	Overall Score
1	RF	0.3401	0.3275	0.3351
2	SVM	0.2993	0.2889	0.2951
3	DT	0.2789	0.2645	0.2731
4	LR	0.1020	0.0580	0.0844
5	NB	0.1020	0.0571	0.0840

Interpretation

Overall, the classifiers were ranked in terms of the combined predictive accuracy and reliability according to Table 12. Random Forest had the best Overall Score (0.3351), which means that the model was more reliable in its predictions. Support Vector Machine was the second best, with satisfactory classification performance and prediction consistency, followed by Decision Tree. The logistic regression and Naïve Bayes achieved relatively poor results due to their poor ability to model complex sentiment data with a large degree of variation and multiple classes. The overall ranking shows that the benchmark classifiers in the proposed framework are most effective when using Random Forest. The Overall Score was calculated by assigning 60% weight to Accuracy and 40% to Reliability Score. The formula was $OS = 0.6 (Accuracy) + 0.4 (Reliability)$. For Random Forest, $OS = (0.6 \times 0.3401) + (0.4 \times 0.3275) = 0.3351$. A higher Overall Score indicates better combined predictive performance and reliability among the evaluated classifiers.

Combined predictive accuracy and reliability scores are shown in overall ranking of the evaluated machine learning models in Figure 4. RF has the best overall score, giving it a good balance between its classification performance and the consistency of its predictions. Next are SVM and DT, followed by comparatively low-ranking Logistic Regression and Naïve Bayes.

4.2 Discussion of Experimental Findings

The proposed framework was applied to the experimental evaluation for comparative analysis of the performance and reliability of the selected benchmark machine-learning classifiers for social-media sentiment analysis. The results show that the classifiers evaluated have different predictive power and reliability. The model with the highest observed accuracy and Reliability Score was the Random Forest model at 34.01% and 0.3275 respectively, with the model showing comparatively better predictive strength and reliability with the experimental conditions of the study. Decision Tree had moderate performance and Support Vector Machine had competitive performance. The performance values of Logistic Regression and Naïve Bayes were comparatively low. The results are however relative, as the absolute accuracy is not high enough to conclude that the sentiment classification task has been solved easily and that the reported results can easily be extended to other cases without being interpreted as evidence of broad robustness or generalizability.

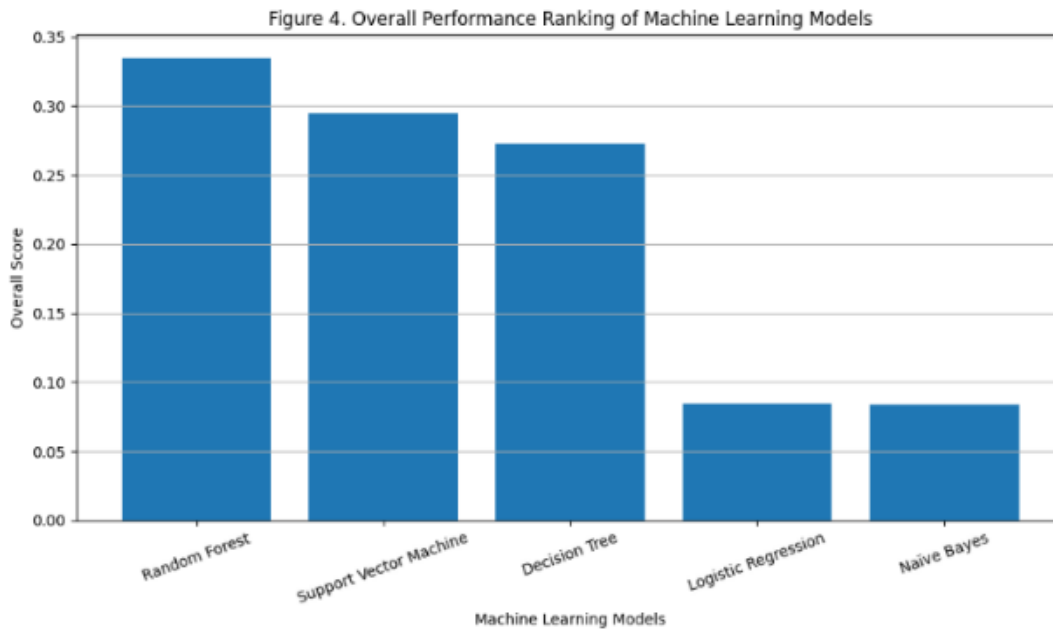


Figure 4: Overall Ranking of Machine Learning Models

5. CONCLUSION

For the present investigation, an assessment framework was suggested to contrast the performance and reliability of the machine-learning models for sentiment analysis of human responses on social media. For the comparative evaluation, a benchmark dataset was created consisting of 732 social media posts published on various platforms and from various regions. The same preprocessing and experimental conditions were used to evaluate Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine. Experimental results revealed that among the evaluated classifiers, the highest classification accuracy (34.01%) and Reliability Score (0.3275) are obtained by Random Forest algorithm. The results of these assessments suggest some relative success in the experimental setting of the study. The relatively low classification accuracy suggests, however, that this sentiment classification task is still difficult, and thus the results should not be taken as a definitive sign of the applicability of the results in real-world applications or generalizability. The practical applicability of the framework needs to be assessed with further validation using additional and independently evaluated datasets.

6. FUTURE WORK

The proposed evaluation framework for analysing complex sentiment classification tasks will be extended to further investigate the effectiveness of CNN, LSTM, BiLSTM, BERT, RoBERTa, DistilBERT and other advanced deep learning and transformer-based models. The framework can be generalized to process multimodal and multilingual datasets, containing text, images and contextual information, from different social media. Moreover, fairness-aware evaluation metrics and XAI

techniques can be incorporated to foster the transparency of models, minimize prediction bias, and boost the trust of users. Further research could also include real-time sentiment analysis using streaming social media data and automated model selection methods to develop scalable, adaptive and reliable sentiment analysis tool for various real-world applications of making decisions.

References

- [1] Abdulrashid I, Ahmad IS, Musa A, Khalafalla M. Impact of Social Media Posts' Characteristics on Movie Performance Prior to Release: An Explainable Machine Learning Approach. *Electron Commer Res.* 2025;25:4533-4557.
- [2] Abraham TM, Wen T, Wu T, Chen Y. Leveraging Data Analytics for Detection and Impact Evaluation of Fake News and Deepfakes in Social Networks. *Humanit Soc Sci Commun.* 2025;12:1040.
- [3] Afriyie Y, Weyori BA. Sentiment Analysis Based on Deep Learning Approaches for Text Classification. *J Electr Syst Inf Technol.* 2026;13:14.
- [4] Ashok P, Suganya R, Lekha P, Pandian P, Subashini M, et al. Developed a Machine Learning and Deep Learning Model for 5G Mimo Data-Based Beam Selection and Intelligent Network Analytics. *SN Comput Sci.* 2025;6:203.
- [5] Ataguba G, Oyebode O, Alslaity A, Frempong G, Imran M, et al. Towards Social Good: How Can Social Media Inform the Next Genai's Collaboration With Humans for Dietary Support? *J Healhc Inform Res.* 2026;10:787-849.
- [6] Azgomi H, Shahrestani MR. A Review of Efficient Approaches for Sentiment Analysis and Opinion Mining. *Multimed Tools Appl.* 2026;85:127.
- [7] Bahadorizadeh H, Malek MR. Leveraging Social Media and Vulnerability Maps for Post-Flood Event Localization. *J Comput Soc Sci.* 2026;9:47.
- [8] Chakrabarty D, Chatterjee S, Mukhopadhyay A. A Global Twitter Sentiment Analysis Model for Covid-Vaccination. *Sci Rep.* 2026;16:9005.
- [9] del Gobbo E, Ippoliti L, Fontanella L, Cafarelli B. Mining Social Media Data to Track Environmental Disaster Events. *Environ Ecol Stat.* 2026;33:403-429.
- [10] Donahue E, Canute M, Akinwumi A, Adams PR, Taboada M, et al. Conflicting Feelings: Sociological and Computational Sentiment in Workplace Sentiment Surveillance. *AI Soc.* 2026;41:787-799.
- [11] Gulanbaier T, Hao W, Mairidan W. Multimodal Sentiment Analysis Based on Dual Perspective Modeling. *Multimed Syst.* 2026;32:154.
- [12] Gupta S, Sinha A, Saxena B. Malicious Bot Detection in Twitter/X Social Media Platform With Interpretable Machine Intelligence. *J Ambient Intell Humaniz Comput.* 2026;17:429-444.
- [13] Helmy AA. Multilingual Depression Screening via Social Media: Comparative Analysis of Machine Learning Models on English and Arabic Text. *J Electr Syst Inf Technol.* 2026;13:34.

- [14] Hu Z, Yu Z, Guo J. Multi-Level Sentiment-Aware Clustering for Denoising in Multimodal Sentiment Analysis With ASR Errors. *Multimed Syst.* 2025;31:116.
- [15] Jiang K, Xiao X, Lu X, Qin Y. Semcap: Sentiment-Aware Semantic Captioning for Multimodal Aspect-Based Sentiment Analysis. *J King Saud Univ Comput Inf Sci.* 2026;38:81.
- [16] Jiang X, Ren B, Wu Q, Wang W, Li H. Dcasam: Advancing Aspect-Based Sentiment Analysis Through a Deep Context-Aware Sentiment Analysis Model. *Complex Intell Syst.* 2024;10:7907-7926.
- [17] Jung HS, Lee H, Kim JH. Detecting Bitcoin Sentiment: Leveraging Language Model Applications in Sentiment Analysis for Bitcoin Price Prediction. *Neural Process Lett.* 2025;57:77.
- [18] Karjus A. Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences With Artificial Intelligence. *Humanit soc sci commun.* 2025;12:277.
- [19] Kim H, Cho S. Sentiment Lexicon Integrated Meta-Training for Sentiment Analysis in Data-Scarce Settings. *J Intell Inf Syst.* 2026;64:1225-1254.
- [20] Lamba K, Rani S, Shabaz M. Explainable Machine Learning for Mental Health Prediction From Social Media Behavior: A Nested Cross-Validation Study With Shap and Lime Interpretability. *Discov Ment Health.* 2026;6:25.