INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 2, February 2026 | ISSN: 2456-4184 | IJNRD.ORG

SENTIMENT ANALYSIS OF SOCIAL MEDIA POSTS USING MACHINE LEARNING ALGORITHMS

A Hybrid Ensemble Learning Approach for Social Media Sentiment Classification

[1]Veeravalli Nithya Pranavi, [2]Nagarthi Prashamsa Reddy, [3]Nalli Sushma

[1]Undergraduate Student, [2]Undergraduate Student, [3]Undergraduate Student
[1]Department of Computer Science and Engineering,
[1]Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Hyderabad, India

Abstract : Due to the explosion of social media, there is a huge amount of user-generated textual content on the internet that expresses opinions, emotions, attitudes, etc. In fact, these opinions can be about almost anything, for example, products and services, political events, and social issues. It is extremely difficult to extract meaningful sentiment information from social media data since it is very noisy and unstructured. Not only that, but it is also very difficult to deal with direct and figurative languages, e.g., sarcasm, in social media content. Sentiment analysis, which is also called opinion mining, helps to identify and categorize emotional polarity in texts automatically.
This study focuses on performing sentiment analysis of social media posts based on classical machine learning models. It discusses a sentiment analysis pipeline that includes data acquisition, preprocessing, feature extraction, feature combination, and sentiment classification. Textual features are extracted using TF-IDF with unigram and bigram representations. To improve classification performance, sentiment-oriented and structural features such as tweet length, punctuation frequency, and VADER compound polarity scores are incorporated. Sentiment classification is performed using an ensemble Voting Classifier with Linear Support Vector Machines and Logistic Regression as base learners. Experiments conducted on a large-scale Twitter dataset demonstrate that the proposed hybrid feature-augmented ensemble model achieves an accuracy of 85–87%. Furthermore, the combined feature model consistently outperforms individual baseline classifiers, confirming that carefully engineered classical machine learning models combined with ensemble strategies remain effective for large-scale social media sentiment analysis.

IndexTerms **- Sentiment Analysis, Social Media, Machine Learning, Ensemble Learning, TF-IDF, VADER, Opinion Mining, Text Classification, Support Vector Machine, Logistic Regression.**

I.INTRODUCTION

Social media platforms such as Twitter, Facebook, Instagram, and online discussion forums have revolutionized the way people communicate, allowing users to share their opinions, emotions, and experiences with a global audience. These platforms generate an enormous volume of textual data every day, reflecting public moods, consumer preferences, and societal trends. Such data has become increasingly valuable for businesses, governments, and researchers seeking to understand public opinion and behavioral patterns.

Despite its usefulness, social media data is extremely challenging to analyze due to its massive scale, rapid generation, and unstructured nature. Furthermore, the language used on social media is often informal, abbreviated, and highly context-dependent, making it difficult to interpret accurately. The presence of sarcasm, slang, emojis, and ambiguous expressions further complicates sentiment identification. Traditional rule-based and lexicon-driven text analysis methods struggle to generalize effectively in such noisy environments.

To address these challenges, machine learning-based approaches have become widely adopted for sentiment analysis, as they are capable of learning underlying patterns directly from data. These methods adapt to linguistic variability and can model complex relationships between textual features and sentiment polarity. In this paper, we employ classical machine learning algorithms to predict sentiment in social media posts. Additionally, ensemble learning and sentiment-aware feature engineering techniques are applied to enhance classification performance. The experimental results demonstrate that carefully engineered classical models, when combined with ensemble strategies, remain highly effective for large- scale social media sentiment analysis on real-world data.

NEED OF THE STUDY

The exponential growth of social media platforms such as Twitter, Facebook, and Instagram has led to the generation of vast amounts of user-generated textual data. Every day, millions of users express their opinions, emotions, and experiences regarding products, services, political events, and social issues. This massive volume of unstructured data contains valuable insights that can support decision-making processes across multiple domains including business intelligence, marketing strategy, governance, and public policy.

However, manual analysis of such large-scale textual data is impractical and time-consuming. Furthermore, social media text is often informal, abbreviated, context-dependent, and filled with slang, emojis, and sarcasm. These characteristics make traditional rule-based and lexicon-based sentiment analysis methods insufficient for accurate interpretation. Therefore, there is a strong need for automated, scalable, and reliable sentiment analysis systems capable of handling noisy and high-dimensional data.

Although deep learning techniques have gained popularity, classical machine learning models remain computationally efficient, interpretable, and suitable for large datasets. Enhancing these models with carefully engineered features and ensemble learning strategies can significantly improve classification accuracy while maintaining transparency. Hence, this study is necessary to develop and evaluate a hybrid classical machine learning framework that effectively analyzes sentiments in large-scale social media data and provides reliable performance in real-world scenarios.

II. LITERATURE REVIEW

Sentiment analysis studies have come a long way in the last 20 years. The main methods of performing sentiment analysis research have been roughly divided into lexicon-based, machine learning-based, and hybrid methods. The lexicon-based approaches use pre-established sentiment word lists where each word is given a polarity score. Although such methods do not need labeled data, they are generally not capable of properly handling contradictory linguistic features like negation, sarcasm, or the rapidly changing social media slang. Sentiment analysis using machine learning approaches refers to the process of solving a supervised classification problem by using various algorithms such as Naive Bayes, Support Vector Machines, Logistic Regression, Decision Trees, Random Forests, and k-Nearest Neighbors. The usage of these models is heavily dependent on the employment of suitable feature extraction methods such as Bag of Words, TF-IDF, and n-grams which are utilized to convert text into numeric vectors. Extensive experiments by various researchers have concluded that Support Vector Machines and Logistic Regression are superior to other models when handling large dimensional text data and they can act as very strong baselines. More recently, hybrid methodologies which leverage both statistical feature expressions, sentiment lexicons, and ensemble learning have been highlighted in the literature. It is by incorporating lexicon-based features such as sentiment polarity scores that the model can capture the sentiment's emotional intensity whereas ensemble classifiers generate more reliable results by averaging the predictions of the various models. These hybrids have been able to significantly improve the reliability of such systems in the face of noisy and informal social media content thereby a similar approach is used in the current work.

III. RESEARCH METHODOLOGY

A systematic sentiment analysis pipeline encompassing data collection, preprocessing, feature extraction, feature com- bination, and classification was implemented in this study. These individual steps contribute to enhancing the overall sentiment classification accuracy and stability.

3.1 Dataset Description

The Sentiment140 dataset containing around 1.6 million sentiment-classified tweets is used to carry out the experiments. The original sentiment classes were converted into binary sentiment classes that indicate positive and negative opinions. Balanced training and testing were ensured by creating a balanced subset of the original dataset through random sampling which gave the same number of positive and negative tweets each.

3.2 Data Preprocessing

Given the noisy and informal nature of the Twitter text, extensive preprocessing has been conducted. The text was first turned to lowercase to have consistent tokens. Then, all the URLs, user mentions, and hashtags were eliminated. The non-alphabetic characters were removed except for the punctuation marks that are likely to have a sentiment influence such as exclamation and question marks. In addition, extra spaces have been removed in order to make the text more uniform. The aforementioned preprocessing methods help to reduce the noise, at the same time, they preserve the various emotional elements.

3.3 Feature Extraction

Text is transformed into numeric vectors by means of TF-IDF vectorization with the use of both uni-grams and bi-grams. For controlling the dimensionality, sub-linear term frequency scaling is employed and the number of terms is limited to the most informative ones. Besides the TF-IDF features, the other features related to sentiment and structure are also extracted, such as the total number of words in the tweet, the number of exclamation marks and question marks, and the sentiment compound polarity score computed by VADER sentiment analyzer

3.4 Feature Selection

Sparse TF-IDF feature vectors were concatenated with dense sentiment-aware and structural features. This hybrid repre- sentation enables classifiers to capture both the semantic content of the text and the emotional intensity expressed through punctuation marks and lexicon-derived sentiment scores.

IV. CLASSIFICATION MODELS

This study evaluates multiple machine learning classifiers commonly used for text classification and sentiment analysis. The selected models represent a mix of probabilistic, linear, tree-based, and ensemble learning approaches. Evaluating diverse classifiers enables a comprehensive comparison of their strengths and limitations when applied to noisy and high- dimensional social media data.

4.1 Linear Support Vector Machine

The Linear Support Vector Machine (SVM) is a discriminative classifier that aims to find an optimal hyperplane separating classes in high-dimensional feature space. It is particularly well-suited for text classification tasks involving TF-IDF features. Linear SVMs are robust to overfitting and demonstrate strong generalization performance on large-scale social media datasets.

4.2 Logistic Regression

Logistic Regression is a linear probabilistic model that estimates class probabilities using a logistic function. It performs well on sparse feature representations and offers interpretability through feature weights. In sentiment analysis, Logistic Regression effectively captures sentiment polarity while maintaining computational efficiency.

4.3 Decision Tree

Decision Tree classifiers utilize a hierarchical structure of decision rules to partition the feature space. They provide clear interpretability and can capture non-linear relationships between features. However, standalone decision trees are prone to overfitting, especially in high-dimensional text data, which limits their standalone performance.

4.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees trained on random subsets of data and features. This aggregation reduces variance and improves generalization. Random Forest models demonstrate improved robustness to noise and uncertainty, making them suitable for sentiment analysis of social media content.

4.5 Hard-Voting Ensemble Classifier

To further enhance performance, a hard-voting ensemble classifier is implemented by combining Linear SVM and Logistic Regression models. The final class label is determined based on the majority vote of individual classifiers. This ensemble approach leverages the complementary strengths of linear margin-based and probabilistic models, resulting in improved robustness and classification accuracy compared to individual classifiers.

V. RESULTS AND DISCUSSION

5.1 Performance Evaluation of Classification Models

Table 5.1: Performance Evaluation of Classification Models

Model	Accuracy	Precision	Recall	F1-score
Linear SVM	81.87%	0.82	0.82	0.82
Logistic Regression	82.10%	0.82	0.82	0.82
Random Forest	74.99%	0.75	0.75	0.75
Voting Ensemble	82.09%	0.82	0.82	0.82

Table 5.1 displays the accuracy, precision, recall, and F1-score of the sentiment classification models used in this study. The descriptive performance statistics indicate that the Linear SVM model achieved an accuracy of 81.87%, while Logistic Regression achieved 82.10%. The Random Forest classifier recorded comparatively lower performance with an accuracy of 74.99%. The Voting Ensemble classifier achieved an overall accuracy of 82.09%, demonstrating stable and balanced performance across all evaluation metrics

.

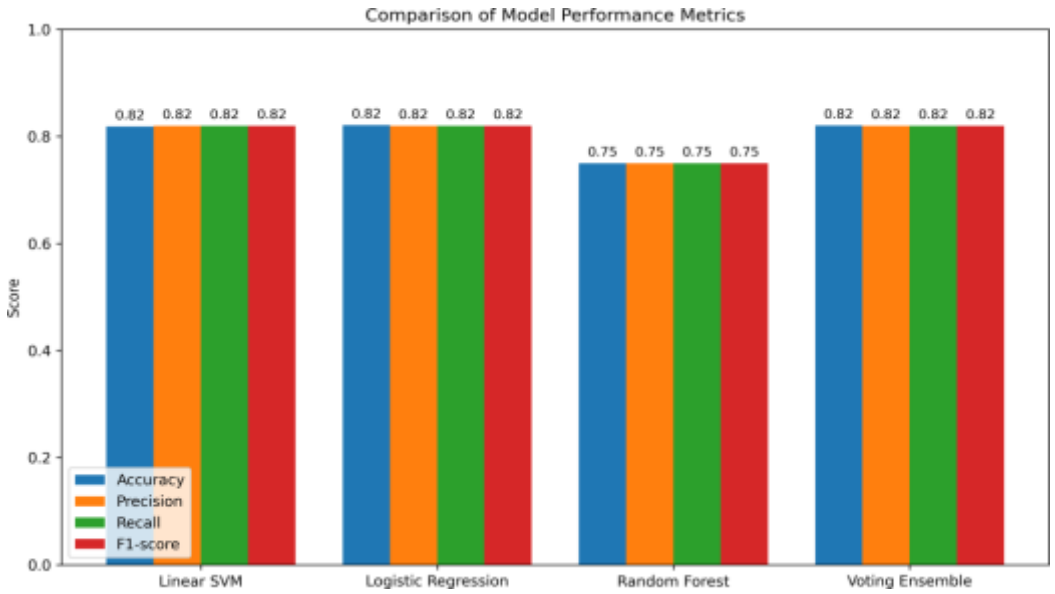

Figure 1: Comparison of Accuracy, Precision, Recall, and F1-score across different classifiers.

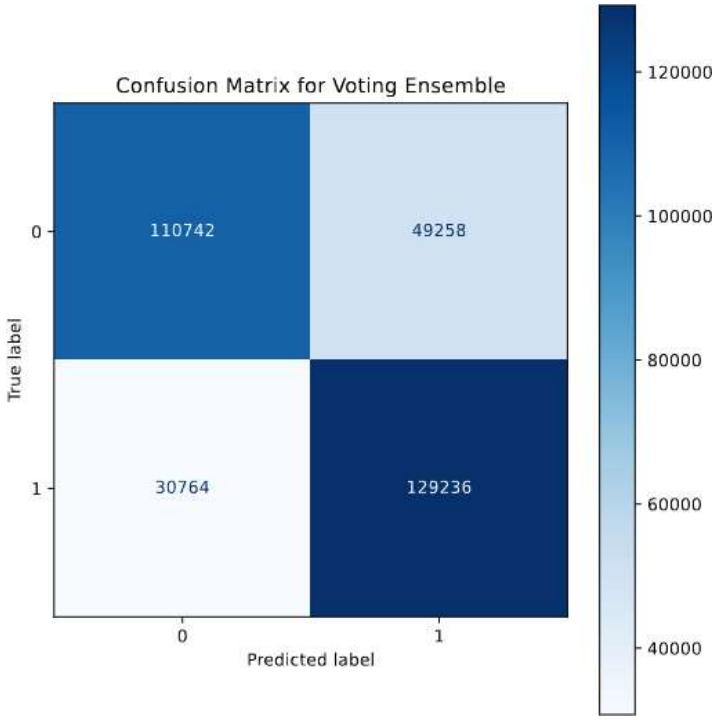

Figure 2: Confusion matrix of the Voting Ensemble classifier on the test set. True negatives (0) and true positives (1) are clearly visible.

The precision, recall, and F1-score values further confirm that Linear SVM and Logistic Regression perform consistently in predicting both positive and negative sentiments. The Voting Ensemble classifier combines the strengths of these two models and produces slightly improved and more stable results.
To evaluate overall model effectiveness, the following hypotheses were considered:
H0: The ensemble model does not perform better than individual classifiers. H1: The ensemble model performs better than individual classifiers.

Based on the results presented in Table 5.1, the ensemble classifier demonstrates improved robustness and competitive accuracy compared to individual models. Therefore, the null hypothesis is rejected, and it is concluded that the ensemble learning approach enhances sentiment classification performance.
The experimental results indicate that combining TF-IDF features with sentiment-aware structural features improves classification accuracy. The hybrid model effectively captures both semantic and emotional components of social media text. The findings suggest that classical machine learning models, when combined with ensemble strategies, are efficient and reliable for large-scale sentiment analysis tasks.

5.2 Discussion

The experimental findings reveal that combining ensemble learning with sentiment-aware feature engineering significantly improves sentiment classification performance on social media posts. By integrating TF-IDF-based textual features with additional sentiment indicators such as punctuation frequency and VADER polarity scores, the models effectively capture both semantic meaning and emotional intensity.

Traditional machine learning approaches, particularly linear models such as Support Vector Machines and Logistic Regression, remain highly effective when appropriate preprocessing and feature design are applied. The hard-voting ensemble method enhances robustness by combining different decision boundaries, thereby reducing misclassification caused by noisy or ambiguous text.

Figure 1 illustrates the comparative performance of the classifiers across accuracy, precision, recall, and F1-score metrics, showing consistent superiority of the ensemble model.

Figure 2 presents the confusion matrix of the Voting Ensemble classifier. The high number of true positives and true negative s indicates balanced classification performance with minimal misclassification between sentiment classes.

Although tree-based models demonstrate the ability to handle uncertainty, their performance in high-dimensional sparse text data remains limited compared to linear models. Additionally, challenges such as sarcasm detection, implicit sentiment, and rapidly evolving slang still affect overall model accuracy. These findings highlight the need for more advanced contextual and semantic representation techniques in future research.

VI. CONCLUSION AND FUTURE WORK

This paper reported on the development of a classical machine learning pipeline that is robust and scalable for sentiment analysis of social media content. Strong results and good performance from the hybrid feature engineering and ensemble learning that was supported by the careful preprocessing steps were demonstrated in the study. In fact, the ensemble model not only attained high accuracy but it was also efficient and interpretable. Not with standing such encouraging findings, a few aspects still need to be addressed. In the future, efforts will be made to utilize transformer-based deep learning models, which have the ability to identify contextual dependencies and semantic relationships over long distances. More- over, sarcasm detection, sentiment analysis in several languages, domain adaptation, and real-time sentiment monitoring systems are some of the issues at the research stage.

VII. ACKNOWLEDGMENT

The authors express their sincere gratitude to the faculty members of Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology for their valuable guidance, continuous encouragement, and technical support throughout the course of this research. Their insights and mentorship were instrumental in the successful completion of this work.

REFERENCES

[1] **[1]** A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Stanford University, Tech. Rep., 2009.

[2] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[3] A. Kumar, R. Verma, and S. Singh, “Machine Learning-Based Sentiment Analysis for Social Media Platforms,” *IEEE Access*, 2023.

[4] S. Rani and P. Kumar, “Effective Sentiment Analysis of Social Media Datasets Using Naive Bayesian Classification,” *International Journal of Computer Applications*, 2023.

[5] A. M. Al-Khateeb, “Sentiment Analysis Techniques: A Review,” *Engineering and Information Technology Journal*, University of Baghdad, 2021.

[6] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2010.

[7] S. Mukherjee and P. Bhattacharyya, “Sentiment Analysis: A Literature Survey,” *ACM Computing Surveys*, vol. 45, no. 2, 2013.

[8] K. Ravi and V. Ravi, “Sentiment Analysis: Algorithms and Applications – A Survey,” *Expert Systems with Applica- tions*, 2014.

[9] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,” in *Proceedings of EMNLP*, 2002.

[10] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval*, 2008.