Machine Learning-Based Sentiment Analysis On Twitter Data

k.sangeetha
School of Engineering and Technology
Sharda unversity
Greater, Noida

* Palvi Gupta
School of Engineering and Technology
Sharda unversity
Greater, Noida

***Abstract*—Twitter and other social media platforms produce huge amounts of user content which displays current public emotions and feelings. The study faces difficulties because people use informal expressions and slang terms and short forms and emoji symbols and unstructured text. The research focuses on detecting emotions in Twitter data through classical machine learning methods which include Logistic Regression and Linear Support Vector Machine (SVM) and Decision Tree classifiers. The preprocessing procedures together with feature extraction methods create structured data from original tweets which can be used for classification purposes.The models use standard performance metrics which include accuracy and precision and recall and F1-score to check their ability to find emotional categories. The comparison results show that Logistic Regression and Linear SVM perform well with high-dimensional textual features while Decision Trees provide understandable classification results which make them vulnerable to data noise. The research demonstrates that traditional machine learning methods can effectively analyze social media sentiment and emotion, which helps businesses track public opinion and analyze customer feedback.**

***Index Terms*—Emotion Analysis, Sentiment Classification, Twitter, Social Media Analytics, Machine Learning, Logistic Regression, Support Vector Machine, Decision Tree**

I. Introduction

In today's digital era, social media platforms and online marketplaces generate a massive volume of user-generated content.Social media platforms including Twitter,Instagram,YouTube,Facebook,Amazon and Flipkart e-commerce sites permit people to share their views and personal stories while connecting with others around the world. Twitter functions as an exceptional research tool because it delivers brief messages that people around the world can access at any time. The platform enables users to generate billions of tweets each day which creates a vast collection of data that shows how people feel and what they think.Tweets present multiple difficulties for Twitter data analysis because they feature informal writing styles and common shortcuts and regional dialects and emoticons and mocking remarks and multiple languages. The traditional machine learning methods through which businesses implement sentiment and emotion classification use Logistic Regression,Linear Support Vector Machines and Decision Trees as their primary methods. The models use text preprocessing and feature extraction methods to change unstructured tweet data into structured test data which can be used for classification purposes.Multiple areas of study depend on accurately detecting both sentiment and emotion. Businesses use it to optimize their marketing efforts while increasing customer contentment, policymakers employ it to track public sentiment, and researchers use it to study how public emotional states and mental health evolve over time. This research project conducts a comprehensive assessment of Logistic Regression ,Linear SVM and Decision Tree classifiers to evaluate their capacity for Twitter emotional analysis through social media data. The research results establish a foundation for constructing dependable systems that analyze sentiment, which can be used in opinion mining, customer feedback evaluation, and extensive public emotion tracking.

II. Litreature Review

Recent studies have shown that traditional machine learning models continue to analyze social media platforms' sentiment and emotion data especially on Twitter. Singh et al. (2024) conducted a comparative study using Logistic Regression, SVM, and Decision Tree classifiers on Twitter sentiment datasets which showed that linear classifiers work well in social text environments with noisy conditions [1]. Roja and Durairaj (2023) tested Logistic Regression and SVM on post-COVID tweets and found that the models successfully classified multiple sentiment categories while achieving high accuracy rates [2]. The analysis of Twitter datasets by Bohra and Kumari (2024) showed that both Logistic Regression and SVM achieved strong performance results which demonstrated how linear models efficiently handle text classification tasks [3]. Kumar and Jayapratha (2025) created a system which used Logistic Regression, SVM, and Random Forest with TF-IDF vectorization to demonstrate how traditional classifiers process short-text social media content [4].Hidayat and Aminulhaq (2025) studied sentiment classification for app reviews through SVM and Decision Tree and Logistic Regression, which showed SVM as the most effective model while Decision Tree delivered quick and understandable outcomes [5]. Brandão et al. (2024) studied how feature selection and cross-validation methods optimize classical classifiers which include SVM and Decision Tree, showing that proper data preparation leads to significant performance gains [6]. Alabdulkarim et al. (2024)

studied Amazon product reviews and found that Logistic Regression and SVM achieved similar accuracy results when tested against some deep learning models, which demonstrates that traditional methods work effectively with actual datasets [7].Nurlanuly (2025) proposed a hybrid approach which combines machine learning with transformer models yet traditional classifiers such as Logistic Regression and SVM proved to deliver effective baseline results for social network sentiment monitoring [8]. Mantika et al. (2024) used Naïve Bayes and Logistic Regression on Twitter political data to demonstrate that traditional models remain valuable in fast-changing social settings [9]. Nithya et al. (2023) conducted sentiment analysis through various machine learning models that included Logistic Regression and SVM to show how standard NLP preprocessing methods improve model accuracy [10].Ahmed and Abdulmajed (2025) conducted a model comparison study which included SVM and Random Forest models to analyze political Twitter datasets, demonstrating that evaluation metrics such as precision and recall and F1-score are crucial for assessing traditional classifiers [11]. Singh and Jaiswal (2023) conducted a comparative study across several ML models, emphasizing the sustained relevance of Logistic Regression for social media sentiment tasks [12]. Paul et al. (2025) studied how Bangla comments express different emotions, proving that Linear SVM successfully processed multiple languages through social media content [13]. Another recent study in 2026 analyzed event-related English tweets using Logistic Regression, SVM, and Decision Tree, confirming that classical classifiers remain valuable for emotion classification when combined with proper feature extraction techniques [14].Finally Padhy et al. (2024) demonstrated that SVM and Logistic Regression can be applied to Twitter sentiment datasets through the use of TF-IDF features, which showed that traditional models continue to hold value for practical use alongside new deep learning technologies [15].

III. Methodology

The study employs a quantitative experimental design to investigate emotional responses through Twitter data analysis which uses traditional machine learning methods. The method transforms unstructured social media text into structured data elements that enable sentiment classification while testing three different models which include Logistic Regression and Linear Support Vector Machine (SVM) and Decision Tree.

A. Dataset Description

The research used a dataset which contains 73913 user-created text samples that were obtained from different social media sites. The research team collected data from multiple sources, which enabled them to capture different writing styles and speaking patterns and thematic writing styles, thus creating a more diverse collection of linguistic elements for their research. The dataset includes various subjects, which range from technology and healthcare to digital entertainment and online services and e-commerce, making it suitable for sentiment analysis across multiple domains.The processing of each text instance results in classifying it into one of four sentiment categories, which consist of positive, negative, neutral and irrelevant. The multi-class labeling method produces more precise sentiment evaluations because it enables better understanding of sentiment through its multiple available categories. The first analysis of raw data showed various problems that typically arise with social media data, which included duplicate records and different capitalization patterns and hyperlinks and emojis and hashtags and user mentions. The researchers processed these problems during the dataset cleaning process, which helped them achieve better dataset quality and consistent dataset performance.The existing sentiment class distribution showed moderate imbalance before the preprocessing work began. The pie chart demonstrates that positive samples represent 30.5% of the dataset while negative samples account for 26.2% and neutral samples make up 24.1% of the dataset. The dataset contains 19.2% of irrelevant texts, which make up the remaining content. The model will develop a tendency to predict majority classes because the distribution of classes exists in an unbalanced state.The dataset demonstrates how social media users interact with each other by using informal language and abbreviations and sarcastic remarks and specialized vocabulary. The study shows that the system needs the ability to process both noisy data and short unfinished thoughts. The real-world data set creates authentic challenges because it tests machine learning and deep learning models in realistic sentiment analysis situations.Social media texts often contain unwanted elements such as web links and user tags and emojis and hashtags and extra punctuation which do not directly help in identifying sentiment. The solution to this problem required the implementation of a systematic text preprocessing procedure. The process included contraction expansion as its main component which transformed shortened word forms into their complete versions with the help of a contraction handling library. The conversion process enables uniformity throughout the text while it enhances correct token production in later processing phases.

B. Text Preprocessing

A structured preprocessing pipeline was applied to the raw tweets because it helps to enhance the quality of textual data which enables accurate sentiment analysis.

Lowercasing:

The process converted all characters to lowercase because it created a standard format which treated tokens that differed only through capitalization as identical.

Removal of URLs, user mentions, and hashtags:

The process eliminated hyperlinks and user references and hashtag symbols through regular expressions because these elements lack direct value for sentiment interpretation.

Punctuation and non-alphabetic character removal:

The process removed all punctuation marks and digits and non-alphabetic symbols to preserve only essential textual content.

Tokenization:

The preprocessed text was segmented into individual tokens using the Natural Language Toolkit (NLTK) which enabled word-level analysis.

Stopword removal:

The system removed common English stopwords which included ”the” and ”is” and ”an” and ”to” because these words typically lack important sentiment value.

Lemmatization:

The process used the WordNet Lemmatizer to convert tokens into their dictionary-standard forms. The choice of lemmatization over stemming creates linguistically valid root words which maintain their original semantic meanings.

IV. Visualization

A. Confusion Matrix Analysis

To evaluate the performance of the models, confusion matrices were generated for the Logistic Regression, Linear SVM, and Decision Tree classifiers.

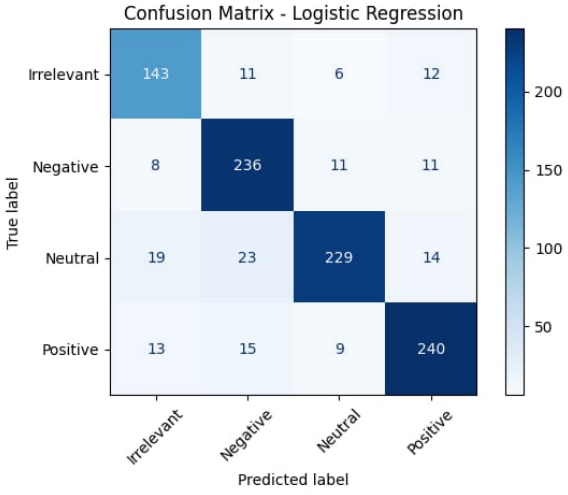

Fig. 1. Confusion Matrix – Logistic Regression model.

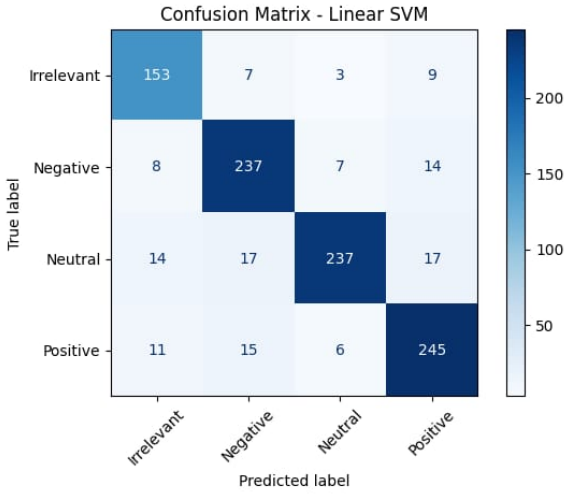

Fig. 2. Confusion Matrix – Linear SVM model.

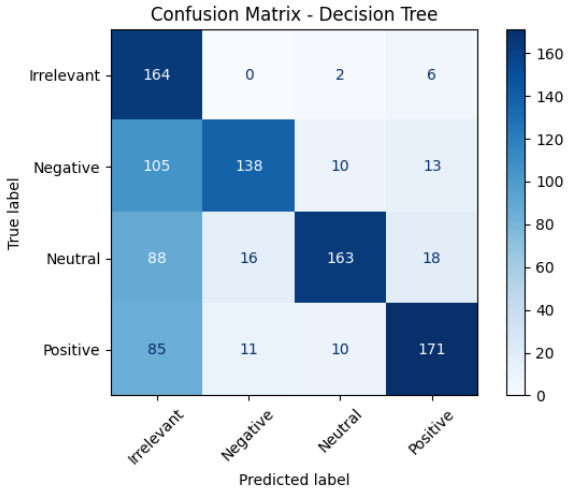

Fig. 3. Confusion Matrix – Decision Tree model.

Figure 1:The confusion matrix for the Logistic Regression model shows its performance results. The model correctly classified sentiments through its diagonal values which show accurate results and used its off-diagonal values to show cases of misclassification. The Logistic Regression model correctly categorizes ”positive” and ”negative” tweets, but it fails to distinguish between ”neutral” and ”irrelevant” tweets because of its inability to detect minor emotional shifts.

Figure 2:The Linear SVM confusion matrix appears in the current section. Linear SVM handles strongly polarized emotions with the same effectiveness as Logistic Regression but fails to differentiate between ”neutral” and ”irrelevant” sentiment categories. The SVM system provides better stability than Logistic Regression but it cannot learn contextual features necessary for accurate emotion identification.

Figure 3:The Decision Tree confusion matrix is presented through this display. Decision Trees produce results that can be understood by users yet their performance declines when they encounter noisy data which affects their ability to detect subtle sentiment between mixed and ambiguous sentiment classes.

B. Class-wise Performance Metrics

The three models were evaluated through confusion matrices which were supported by class-wise precision and recall and F1-score measurements. The visualization displays how the model performs across various sentiment categories while it shows which sentiment categories the model excels at and struggles to handle.

From Figure 4The F1-score analysis for three models displays their performance through the comparison of three different classes. The graph demonstrates that Logistic Regression and Linear SVM perform better than their competitors when it comes to both positive and negative class assessments while neutral and irrelevant tweets present greater difficulties for their systems. The Decision Tree model demonstrates acceptable performance through its F1 score results but its

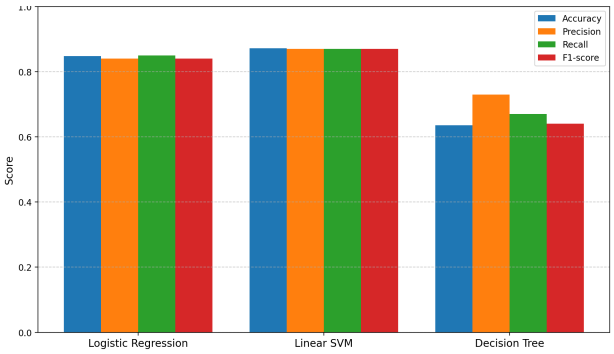

Fig. 4. Class-wise Precision, Recall, and F1-Score for Logistic Regression, Linear SVM, and Decision Tree models.

overall F1 score results show that it struggles with social media data which contains extraneous information. The visualization demonstrates that linear models effectively process high-dimensional textual features whereas tree-based models generate understandable yet less accurate outcomes.

V. Results and Discussion

The results of the study were presented through two different sections which retained their original structure. The study examined three classifiers which included Logistic Regression, Linear SVM and Decision Tree to test their effectiveness on a Twitter dataset that contained 73913 processed samples. The researchers used standard performance metrics to evaluate the models which included accuracy and precision and recall and F1-score. The classical machine learning techniques can perform textual sentiment analysis but their success rate differs according to the specific sentiment type which they need to analyze.The confusion matrices (Figures 1, 2, and 3) show that all three models succeed in detecting highly emotional states which include positive and negative feelings. The system struggles to separate neutral tweets from irrelevant tweets. The two methods Logistic Regression and Linear SVM display identical error patterns because they cannot differentiate between neutral and irrelevant categories without using contextual features. The Decision Tree classifier shows understandable results but its accuracy drops when it processes disorganized and messy social media content.The class-wise performance metrics shown in Figure 4 further support these observations. The two methods achieve their F1 score through positive and negative classifications because they handle high dimensional textual data effectively. The Decision Trees enable people to understand their process but they lead to decreased F1 output for neutral and irrelevant categories because they cannot handle data disturbances and they fail to detect minor emotional changes. Overall, the results suggest that linear classifiers such as Logistic Regression and Linear SVM are well-suited for detecting clear-cut sentiments in social .

VI. Conclusion and Future Work

The study examined how emotions are detected from Twitter data through the use of three machine learning methods which include Logistic Regression and Linear Support Vector Machine (SVM) and Decision Tree classifiers. The research applied systematic text preprocessing and feature extraction techniques to transform raw, noisy tweets into structured data suitable for classification. The models were evaluated using confusion matrices and class-wise performance metrics which included precision and recall and F1-score. The results show that Logistic Regression and Linear SVM achieve better performance than Decision Trees when they need to process high-dimensional textual features for detecting strongly polarized sentiment classes that include positive and negative tweets. The models failed to differentiate between neutral and irrelevant tweets because social media content contains emotional content that is difficult to identify. Decision Trees provide interpretable results but their ability to process noisy and informal language shows decreased strength. Overall, the research shows that traditional machine learning approaches can provide a reliable and efficient method for emotion analysis on social media. However, challenges remain in detecting subtle, mixed, or multilingual sentiments, highlighting opportunities for future research.

A. Future Work

Future research should study advanced feature representations which allow multilingual emotion detection to improve classification of ambiguous and nuanced emotions. The model performance will improve when the dataset expands to include multilingual content which enables better detection of sarcasm and slang and emoji usage. The research will discover effective methods to track public sentiment and evaluate customer opinions on various social media platforms by studying both ensemble techniques and real-time sentiment analysis systems.

References

[1] R. Singh, N. Kulshrestha, A. Sinha, M. Agarwal, and B. Sinha, "Twitter sentiment analysis using machine learning algorithms: A comparative analysis," in *Advancements in Communication and Systems*. SCRS, India, 2024, pp. 135–144.

[2] N. Roja and T. Durairaj, "Evaluation of logistic regression and svm on post-covid twitter sentiment," *International Journal of Research in IT and Computer Science*, vol. 12, no. 7, 2023.

[3] A. Bohra and S. Kumari, "Comparative study of logistic regression and svm on twitter sentiment analysis," *Journal of Artificial Intelligence Research*, 2024.

[4] K. Kumar and P. Jayapratha, "Machine learning based twitter sentiment analysis using tf-idf features," *IJSRST*, 2025.

[5] R. Hidayat and A. Aminulhaq, "Sentiment classification of app reviews using svm, decision tree, and logistic regression," 2025.

[6] R. Brandão, J. Silva, and P. Fernandes, "Optimization of classical machine learning models for text classification," *Elsevier ScienceDirect*, 2024.

[7] H. Alabdulkarim, F. Alshammari, and S. Ahmed, "Application of ml techniques for amazon product review sentiment analysis," *MDPI Algorithms*, 2024.

[8] N. Nurlanuly, "Hybrid machine learning and transformer approach for social media sentiment monitoring," *arXiv preprint arXiv:2502.17143*, 2025.

[9] L. Mantika, R. Purnomo, and S. Hartono, "Political twitter sentiment classification using naïve bayes and logistic regression," 2024.

[10] N. Nithya, B. M. Devaraju, and H. A. Girijamma, "Twitter sentiment analysis using machine learning algorithms," *IJERT*, 2023.

[11] M. Ahmed and A. Abdulmajed, "Evaluation of ml models on political tweets," 2025.

[12] R. Singh and S. Jaiswal, “Comparative study of machine learning models for social media sentiment,” 2023.
[13] P. Paul, S. Chatterjee, and M. Roy, “Emotion classification in bangla social media comments using linear svm,” 2025.
[14] “Computational sentiment analysis for event-related english tweets using ml models,” *ResearchGate*, 2026.
[15] M. Padhy, S. Patnaik, and R. Das, “Svm and logistic regression for twitter sentiment analysis using tf-idf features,” *MDPI Algorithms*, 2024.