IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Sentiment Analysis of IMDB Movie Reviews Using Machine Learning and NLP Techniques

Shweta Daware
Centre for Distance and Online Education (CDOE)
University of Mumbai, Mumbai, Maharashtra, India

Abstract: *With the rapid growth of digital content, sentiment analysis has emerged as an effective NLP technique for identifying opinions and emotions from textual information. This research develops a machine learning-based sentiment classification system using the IMDb 50K Movie Reviews dataset, which contains an equal number of positive and negative reviews. The collected reviews were cleaned through several preprocessing techniques, including lowercase conversion, removal of HTML tags, punctuation, duplicate records, stop words, and lemmatization. Exploratory Data Analysis (EDA) was performed to investigate the characteristics of the dataset through sentiment distribution, review length analysis, word frequency, word clouds, and bigram visualization. The cleaned text was transformed into numerical representations using the TF-IDF vectorization technique before training multiple machine learning algorithms. Among the evaluated models, Logistic Regression produced the best classification performance with an accuracy of 89%. The results demonstrate that appropriate preprocessing and feature engineering significantly improve sentiment prediction while maintaining an efficient and interpretable classification model. The developed model can be adapted for various real-world applications, including recommendation systems, customer feedback evaluation, opinion mining, and social media analytics.*

Keywords: Natural Language Processing, Sentiment Analysis, IMDb Dataset, Logistic Regression, TF-IDF.

I. INTRODUCTION

The widespread use of online platforms has resulted in a massive increase in user-generated textual content, including product reviews, blogs, discussion forums, and social media posts. These textual opinions provide valuable information for understanding public perception and user satisfaction. Since manually examining such a large amount of data is impractical, automated sentiment analysis has become an effective solution for identifying whether opinions expressed in text are positive or negative. Movie reviews are widely used in sentiment analysis research because they contain diverse writing styles, emotional expressions, and subjective opinions. Correctly identifying the sentiment of movie reviews helps recommendation systems, audience feedback analysis, and decision-making within the entertainment industry. Recent advances in Natural Language Processing (NLP) and machine learning have made it possible to transform unstructured text into numerical representations that can be efficiently processed by classification algorithms. In this work, the IMDb 50K Movie Reviews dataset is used to develop a binary sentiment classification model. The dataset undergoes several preprocessing operations, including duplicate removal, lowercase conversion, elimination of HTML tags and punctuation, stop-word removal, tokenization, and lemmatization. Exploratory Data Analysis (EDA) is then performed to examine important characteristics of the dataset before feature extraction using TF-IDF. Finally, multiple machine learning algorithms are trained and evaluated to determine the most effective model for sentiment classification. The primary objective of this research is to build an accurate, efficient, and interpretable sentiment analysis system while demonstrating the contribution of text preprocessing and feature extraction techniques to overall classification performance. This study also compares the performance of different machine learning classifiers using standard evaluation metrics to identify the most suitable approach for sentiment classification. The proposed

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-37273

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

methodology emphasizes simplicity, computational efficiency, and interpretability, making it suitable for practical real-world applications.

II. LITERATURE REVIEW

Sentiment analysis has become one of the most widely studied applications of Natural Language Processing (NLP), as it enables the automatic identification of opinions, emotions, and attitudes expressed in textual data. The rapid growth of user-generated content on online platforms such as e-commerce websites, social media, and review forums has significantly increased the demand for accurate sentiment classification techniques. By analyzing customer opinions, sentiment analysis helps organizations understand user preferences, improve products and services, and support data-driven decision-making. Early research in sentiment classification primarily focused on traditional machine learning approaches. The pioneering work of Pang et al. demonstrated that algorithms such as Naïve Bayes, Maximum Entropy, and Support Vector Machines could effectively classify movie reviews based on sentiment. Their study highlighted the importance of feature representation and established a strong foundation for future research in opinion mining. As the field evolved, researchers explored various text preprocessing techniques and feature extraction methods to improve classification performance. Approaches including tokenization, stop-word removal, stemming, and lemmatization became standard preprocessing steps, while techniques such as Bag-of-Words and Term Frequency–Inverse Document Frequency (TF-IDF) were widely adopted to convert textual information into numerical representations suitable for machine learning models. Among these, TF-IDF has remained one of the most effective feature extraction techniques due to its ability to emphasize informative words while reducing the influence of commonly occurring terms. Recent studies by Faridi et al., Ipadeola et al., and Guo have shown that combining TF-IDF with Logistic Regression continues to provide competitive performance for sentiment classification tasks. Although deep learning models such as recurrent neural networks and transformer-based architectures have achieved state-of-the-art results in many NLP applications, traditional machine learning methods remain attractive because they require fewer computational resources, offer faster training times, and provide greater interpretability. These advantages make them particularly suitable for binary sentiment classification problems and educational or resource-constrained environments. Motivated by these findings, the present study employs comprehensive text preprocessing followed by TF-IDF feature extraction and Logistic Regression to classify sentiments in the IMDb 50K movie reviews dataset. In addition to model development, exploratory data analysis is performed to examine the characteristics of the dataset and provide insights into review length, word distribution, and sentiment balance before classification.

TABLE I: RELATED WORK IN SENTIMENT ANALYSIS

Author	Year	Dataset	Method	Key Finding
Pang et al.	2002	Movie Reviews	Naïve Bayes, MaxEnt, SVM	Established ML for sentiment analysis
Faridi et al.	2025	IMDb	Logistic Regression	Achieved over 90% accuracy using Grid Search and Active Learning
Ipadeola et al	2025	IMDb	TF-IDF, Word2Vec, SVM	TF-IDF with SVM delivered strong performance.
Guo	2026	IMDb 50K	TF-IDF + Logistic Regression, BERT	BERT slightly outperformed Logistic Regression, but the traditional approach remained efficient and interpretable

III. METHODOLOGY

3.1 Dataset Description

The experiments were conducted using the IMDb 50K Movie Reviews dataset, a publicly available benchmark dataset frequently used for binary sentiment classification tasks. It contains 50,000 movie reviews that are evenly divided into positive and negative sentiment categories. The balanced nature of the dataset minimizes class imbalance and provides a reliable benchmark for evaluating machine learning models designed for text classification.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

TABLE II: DATASET DESCRIPTION

Attribute	Description
Dataset Name	IMDb 50K Movie Reviews
Total Reviews	50,000
Positive Reviews	25,000
Negative Reviews	25,000
Classification Type	Binary(Positive/Negative)

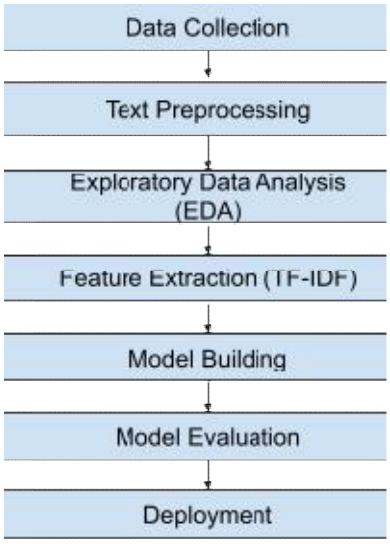

Fig. 1. Workflow of the Sentiment Analysis System.

The figure illustrates the complete NLP pipeline used in this study, starting from data collection to model deployment. It shows the sequential steps involved in building the sentiment classification system.

3.2 Text Preprocessing

The raw movie reviews contained noise such as HTML tags, punctuation, stop words, and inconsistent formatting, which could affect classification performance. During preprocessing, 418 duplicate reviews were removed, followed by lowercase conversion, HTML tag and punctuation removal, stop word removal, tokenization, and lemmatization. The cleaned reviews were then used for Exploratory Data Analysis (EDA) and TF-IDF feature extraction. These preprocessing steps improved text quality and enhanced the effectiveness of the sentiment classification model.

TABLE III: DATA PREPROCESSING

Preprocessing Step	Purpose
Duplicate Removal	Removed 418 duplicate reviews.
Lowercase Conversion	Standardizes the text by converting all characters into a uniform lowercase format.
HTML Tag Removal	Removes HTML elements from reviews.
Punctuation Removal	Eliminates unnecessary punctuation marks.
Stop Word Removal	Removes common words with little semantic value.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-37273

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Tokenization	Splits text into individual words (tokens).
Lemmatization	Converts words into their base or dictionary form.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand the characteristics and distribution of the IMDb 50K Movie Reviews dataset before model training. EDA helps identify patterns, trends, and relationships within the data, enabling better preprocessing and feature engineering. Various visualizations were created to examine the sentiment distribution, review length, word count, word clouds, frequently occurring words, and bigram patterns. These analyses provide valuable insights into the textual data and support the development of an effective sentiment classification model.

3.3.1 Sentiment Distribution

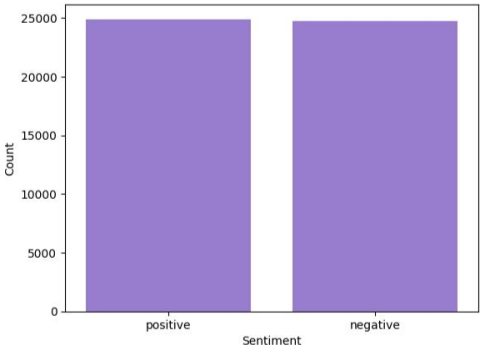

Fig.2 Positive and negative review distribution.

Figure 2 illustrates that the dataset remained well balanced after duplicate removal, with nearly equal numbers of positive and negative reviews, reducing class imbalance. This balanced distribution contributes to fair model training and improves the reliability of performance evaluation across both sentiment categories.

3.3.2 Word Count Analysis

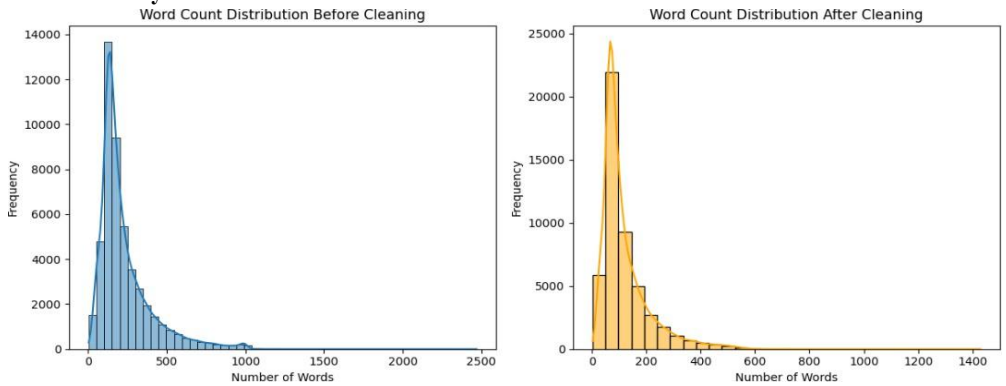

Fig.3 Comparison of Original and Cleaned Word Count Distribution

The average word count decreased from 231 to 120 words after preprocessing, indicating that unnecessary and repetitive words were effectively removed while preserving the meaningful content of each review

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

3.3.3 Word Cloud Analysis

Fig.4 Overall word Cloud of Dataset

The word cloud highlights the most frequent terms in the dataset, reflecting the dominant themes discussed in movie reviews

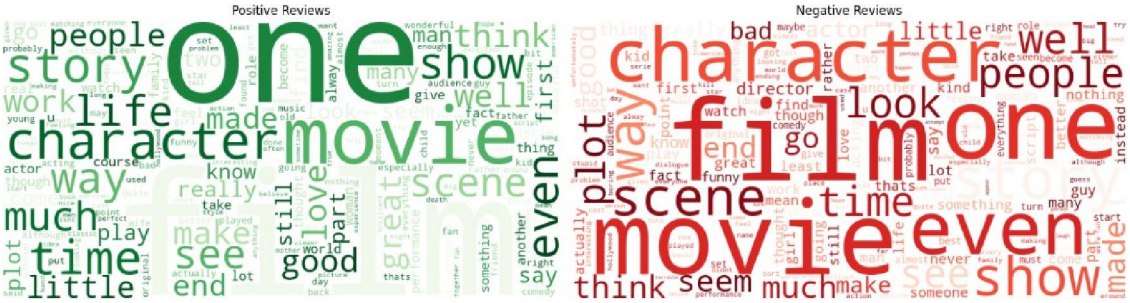

Fig.5 Compares Word Cloud in Positive and Negative Reviews.

Positive reviews emphasize words such as good, story, and character, while negative reviews contain terms like bad, highlighting differences in sentiment.

3.3.4 Most Frequent Words

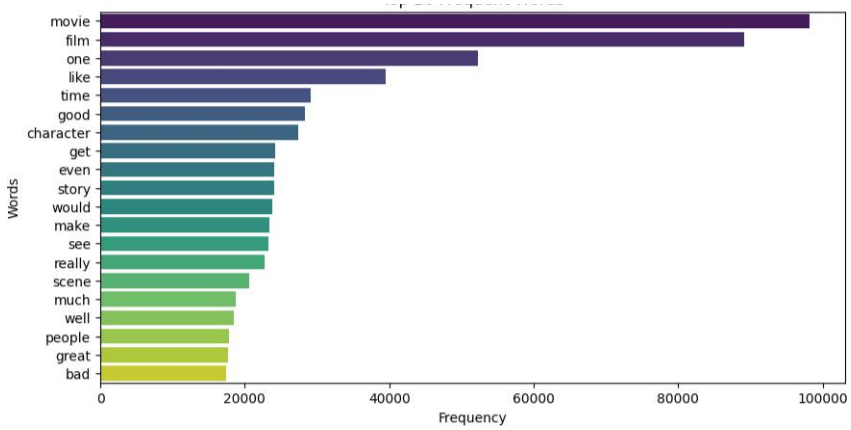

Fig.6 Top 20 Frequent Words

Shows The most frequently occurring words in the dataset are movie, film, one, like, and time. These words reflect the movie-review nature of the dataset.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

3.3.5 Bigram Analysis

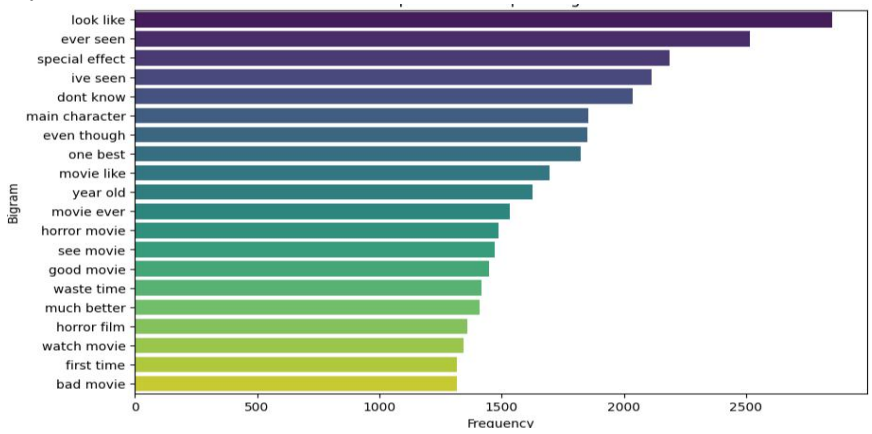

Fig. 7 Top 20 Frequent Bigrams in the Dataset.

Frequent bigrams such as good movie, bad movie, and waste time capture common opinion expressions and contextual information useful for sentiment classification.

3.4 Feature Extraction

Machine learning algorithms require numerical input rather than raw textual data. Therefore, after preprocessing, the cleaned movie reviews were transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF assigns greater importance to words that appear frequently within a review while reducing the influence of words that occur commonly across the entire dataset. To capture both individual words and meaningful word combinations, unigram and unigram–bigram TF-IDF representations were generated. The TF-IDF vectorizer was configured with a maximum of 5,000 features, while the unigram–bigram representation used an ngram range of (1,2). These numerical feature vectors served as the input for the classification algorithms, enabling them to learn linguistic patterns associated with positive and negative sentiments. The resulting sparse feature matrix provided an effective representation of the textual information while maintaining computational efficiency.

3.5 Model Building

3.5.1 Train- Test split

To evaluate the performance of the machine learning models, the dataset was divided into two subsets: training data and testing data. An 80:20 split ratio was used, where 80% of the data was used for training the models and the remaining 20% was used for testing their performance on unseen data..

3.5.2 Feature Input

The machine learning models were trained using numerical representations of the text data generated through the TF-IDF (Term Frequency–Inverse Document Frequency) technique. TF-IDF was used to convert the cleaned textual reviews into a structured format that can be processed by machine learning algorithms.

To capture both individual word importance and contextual relationships between words, both unigram and bigram representations were used. The unigram representation considers single words independently, while the bigram representation captures meaningful word pairs such as “special effects” or “story line,” which helps in improving sentiment understanding.

The resulting TF-IDF feature matrix is high-dimensional and sparse, where each feature corresponds to a unique term or word pair present in the dataset. This feature matrix serves as the input for all classification models used in this study.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

3.5.3 Models Used

In this study, multiple machine learning algorithms were applied to perform sentiment classification on movie reviews. The purpose of using different models was to compare their performance and identify the most effective approach for text classification.

The following models were used:

1. Logistic Regression:

A linear classification model that estimates the probability of a review belonging to a particular sentiment class. It is widely used for text classification tasks due to its efficiency and strong baseline performance.

2. Naïve Bayes:

A probabilistic classifier based on Bayes' theorem with the assumption of feature independence. It is particularly effective for text data and performs well in sentiment analysis tasks.

3. Support Vector Machine (SVM):

A supervised learning algorithm that finds the optimal hyperplane to separate different sentiment classes. It is highly effective in high-dimensional feature spaces such as TF-IDF vectors.

4. Random Forest:

An ensemble learning method that constructs multiple decision trees and combines their outputs for final prediction. It helps improve accuracy and reduces overfitting.

Each model was trained using the TF-IDF feature matrix and evaluated on the test dataset to compare their classification performance.

IV. MODEL EVALUATION

4.1 Evaluation Metrics

To evaluate the performance of the sentiment classification models, standard classification metrics were used, including Accuracy, Precision, Recall, and F1-score.

These metrics provide a comprehensive understanding of how well each model performs in distinguishing between positive and negative sentiments.

Accuracy measures the overall correctness of the model by calculating the proportion of correctly predicted reviews out of the total predictions.

Precision indicates the proportion of correctly predicted positive (or negative) reviews out of all reviews predicted as positive (or negative), reflecting the model's exactness.

Recall measures the ability of the model to correctly identify all actual positive (or negative) reviews, indicating completeness.

F1-score is the harmonic mean of precision and recall and provides a balanced measure of model performance, especially in cases of uneven class distribution.

These evaluation metrics are widely used in text classification tasks to assess and compare the performance of different machine learning models effectively.

4.2 Confusion matrix

The confusion matrix is used to evaluate the performance of classification models by comparing the actual labels with the predicted labels. It summarizes the results in terms of True Positives, True Negatives, False Positives, and False Negatives, providing a clear understanding of correct and incorrect predictions. This helps in analyzing the types of errors made by the model and gives deeper insight into its classification performance beyond standard accuracy metrics.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

4.3 Model Comparison

The performance of all machine learning models was evaluated using unigram and bigram TF-IDF features. Among the evaluated models, Logistic Regression achieved the best overall performance, followed by SVM. Naïve Bayes showed slight improvement with bigram features, while Random Forest produced comparatively lower performance. Overall, the results indicate that linear models are more effective for TF-IDF-based sentiment classification.

TABLE IV: MODEL COMPARISON

Model	Feature Type	Accuracy	Precision	Recall	F1-score
Logistic Regression	Unigram	0.89	0.89	0.89	0.89
	Bigram	0.89	0.89	0.89	0.89
Naive Bayes	Unigram	0.85	0.85	0.85	0.85
	Bigram	0.86	0.86	0.86	0.86
SVM	Unigram	0.88	0.88	0.88	0.88
	Bigram	0.88	0.88	0.88	0.88
Random Forest	Unigram	0.84	0.84	0.84	0.84
	Bigram	0.84	0.84	0.84	0.84

Overall, Logistic Regression emerged as the best-performing model for this dataset due to its balance of simplicity, efficiency, and strong generalization capability.

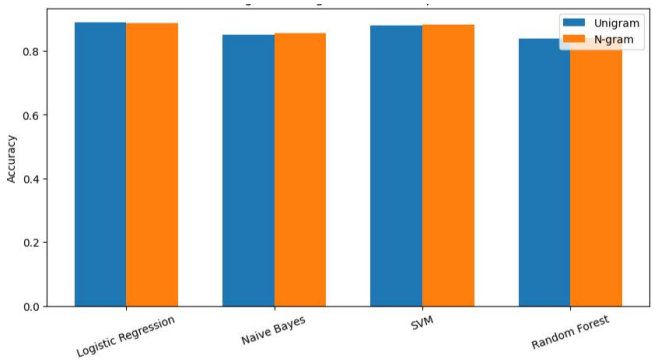

FIG.8 UNIGRAM VS N-GRAM MODEL COMPARISON

V. RESULTS DISCUSSION

5.1 Comparison Insight

The comparison of Logistic Regression, Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF) showed that linear models performed better than probabilistic and tree-based models for TF-IDF-based text classification. While SVM achieved competitive results, Naïve Bayes and Random Forest showed comparatively lower performance. Bigram features provided slight improvements by capturing contextual word relationships, although the overall difference from unigram features was minimal.

5.2 Best Model Selection

Based on the evaluation results, Logistic Regression achieved the highest and most consistent performance across both unigram and bigram TF-IDF feature sets, with an accuracy of 89%. It also demonstrated balanced precision, recall, and F1-score, indicating reliable classification of both positive and negative reviews. Its ability to handle high-dimensional sparse text data and maintain stable performance on unseen data makes it the most suitable model for this study.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

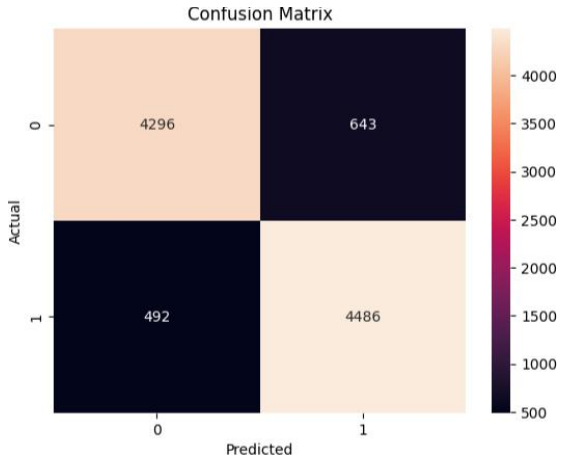

Fig. 9 Confusion Matrix of Best Model

Figure 9 presents the confusion matrix of the Logistic Regression model. The model correctly classified 4,296 negative reviews and 4,486 positive reviews, while only a small number of reviews were misclassified. These results indicate that the model effectively distinguishes between positive and negative sentiments, demonstrating strong classification performance.

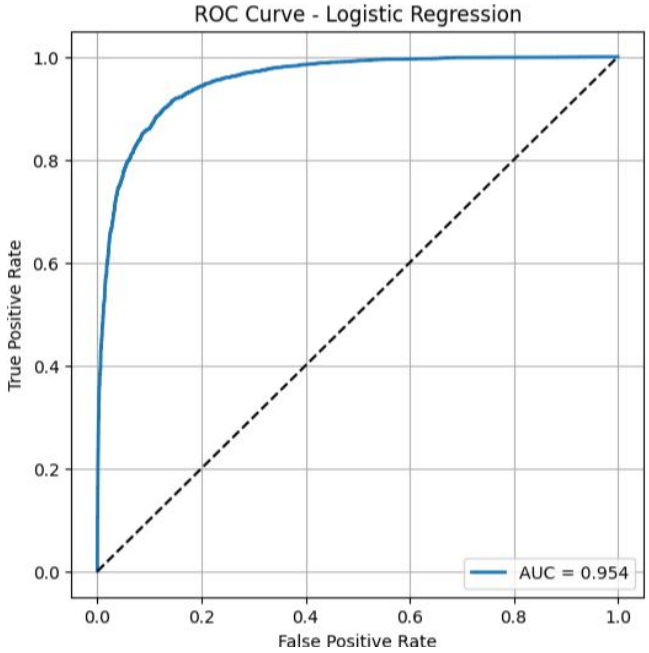

Fig. 10 ROC Curve of Best Model

Figure 10 illustrates the Receiver Operating Characteristic (ROC) curve of the Logistic Regression model. The curve remains close to the upper-left corner, indicating strong classification performance across different decision thresholds. The obtained Area Under the Curve (AUC) value of 0.954 demonstrates the model's excellent ability to distinguish between positive and negative movie reviews, confirming its effectiveness for sentiment classification.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

5.3 Overall Meaning of Results

The results demonstrate that traditional machine learning models combined with TF-IDF feature extraction can effectively classify movie review sentiments. Among the evaluated models, Logistic Regression provided the most balanced, stable, and accurate performance, confirming its suitability for sentiment analysis on the IMDb 50K Movie Reviews dataset.

VI. SYSTEM DEPLOYMENT

The trained sentiment analysis model was deployed using Streamlit to develop an interactive web application. The system enables users to input movie reviews and obtain real-time sentiment predictions. The entered text undergoes preprocessing and is then passed to the trained machine learning model, which classifies it into positive or negative sentiment. The deployment provides a simple and efficient user interface, demonstrating the practical applicability of the proposed model in real-world scenarios for automated text analysis.

The complete implementation of the proposed system, including the source code, trained models, and Streamlit application, is publicly available in the GitHub repository [14].

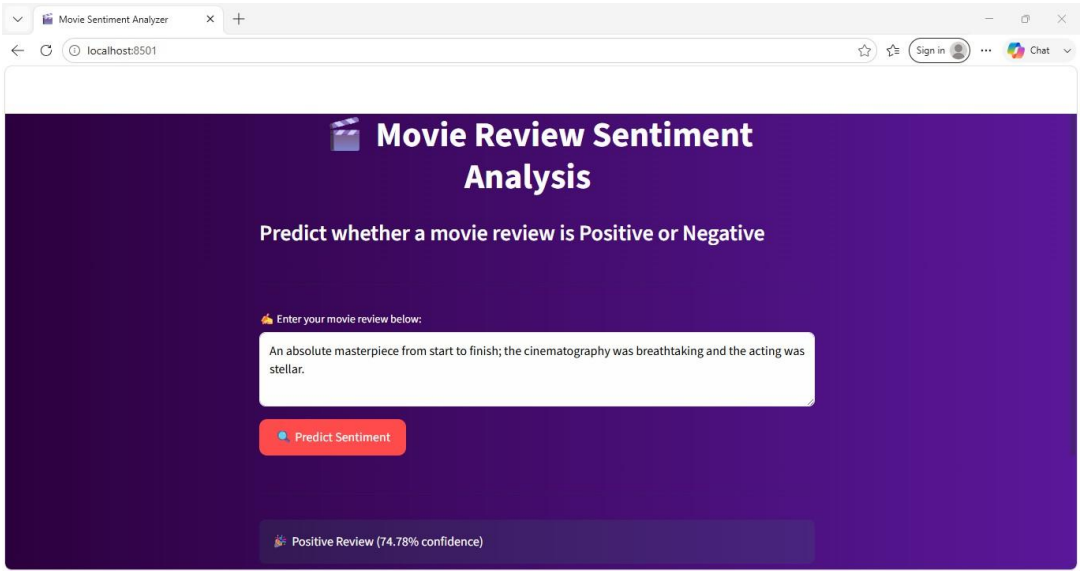

Fig. 11 Streamlit-based Sentiment Analysis Web Application.

VII. CONCLUSION

This study presented a machine learning-based approach for sentiment classification using the IMDb 50K Movie Reviews dataset. The workflow incorporated text preprocessing, exploratory data analysis, TF-IDF feature extraction, and multiple classification algorithms to develop an effective sentiment prediction system. Experimental results indicated that Logistic Regression performed best among all evaluated models, achieving an accuracy of 89% along with balanced precision, recall, and F1-score values. These findings suggest that proper text preprocessing and suitable feature representation significantly enhance model performance in sentiment analysis tasks. Furthermore, the developed Streamlit web application demonstrates the practical usability of the trained model by enabling real-time sentiment prediction. Overall, the proposed system provides an efficient and interpretable solution for binary sentiment classification and can be extended in future work to handle more complex natural language processing tasks and larger datasets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

VII. FUTURE SCOPE

Future work can improve this study by using advanced deep learning models such as LSTM and Transformer-based models like BERT, which can capture deeper contextual meaning in text. Larger and more diverse datasets can also be used to improve model generalization. Additionally, better text representations such as Word2Vec, GloVe, or contextual embeddings can be explored instead of TF-IDF. Future studies may also extend this work to multi-class sentiment classification and real-time sentiment analysis applications.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my guide/mentor for their valuable guidance, continuous support, and encouragement throughout the development of this project. Their suggestions and insights were instrumental in completing this work successfully. I am also thankful to my institution and faculty members for providing the necessary resources and learning environment to carry out this study. Finally, I would like to thank my friends and family for their constant motivation and support throughout this project.

REFERENCES

[1]. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[2]. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[3]. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing.
[4]. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
[5]. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval.
[6]. Faridi, M. A., et al. (2024). Sentiment Analysis of IMDb Movie Reviews Using Machine Learning Techniques. International Journal of Intelligent Systems and Applications.
[7]. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis.
Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 142–150.
[8]. Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the First Instructional Conference on Machine Learning.
[9]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.
[10]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.
[11]. Scikit-learn Documentation. https://scikit-learn.org/ [12]. NLTK Documentation https://www.nltk.org/
[13]. Streamlit Inc. (2025). Streamlit Documentation. https://docs.streamlit.io/
[14]. Shweta Daware, "IMDb Sentiment Analysis Using Machine Learning and NLP", GitHub Repository, 2026. Available: https://github.com/shwetajdaware1433-sys/IMDb-Sentiment-Analysis-Using-Machine-Learning