Sentiment Analysis of IMDB Movie Reviews: Natural Language Processing for Opinion Mining

[1]Mrs. D. Kanakasatya, [2]M. Prema Kumari, [3]L. Hemanth Venkatesh, [4]S. Bhaskar Sai Manikanta, [5]S. Ramya

[1]*Assistant Professor, Srinivasa Institute of Engineering and Technology*

[1234]*UG Scholar, Srinivasa Institute of Engineering and Technology*

doi.org/10.64643/IJIRTV12I10-195128-459

Abstract: The rapid growth of online platforms has significantly increased the volume of user-generated textual data. Movie review websites such as IMDb allow users to share opinions about movies, actors, and overall cinematic experiences. These reviews contain valuable information that can be used to understand audience sentiment and preferences. However, analyzing such large volumes of textual data manually is extremely difficult and time-consuming. Therefore, automated sentiment analysis techniques have become essential for extracting meaningful insights from user reviews.

This research proposes a machine learning-based sentiment analysis system for classifying IMDb movie reviews into positive and negative categories. The system uses Natural Language Processing (NLP) techniques to process and analyse textual data effectively. Initially, the textual reviews undergo preprocessing steps such as text cleaning, tokenization, stop-word removal, and normalization. After preprocessing, the textual data is transformed into numerical feature representations using the Term Frequency–Inverse Document Frequency (TF-IDF) technique.

To perform sentiment classification, multiple machine learning algorithms are applied, including Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). In addition, an ensemble learning method is implemented to combine predictions from these individual models to improve overall performance. The experiments are conducted using the IMDb dataset containing 50,000 labelled movie reviews. Experimental results demonstrate that the proposed approach effectively identifies sentiment polarity and achieves high classification accuracy. The study highlights the effectiveness of combining TF-IDF feature extraction with machine learning algorithms for sentiment analysis tasks.

Keywords: Sentiment Analysis, Natural Language Processing, Opinion Mining, IMDb Dataset, TF-IDF, Logistic Regression, Naive Bayes, Support Vector Machine, Ensemble Learning, Text Mining.

I. INTRODUCTION

In recent years, the internet has become a major platform for users to express their opinions and share experiences about products, services, and entertainment content. Among these platforms, movie review websites such as IMDb play an important role in allowing audiences to publish their opinions about movies. These reviews reflect the viewers' perceptions regarding storylines, acting performances, cinematography, and overall movie quality. As a result, movie reviews provide valuable information for producers, filmmakers, and viewers.

However, the increasing number of reviews generated every day makes it difficult to manually analyze and interpret audience opinions. Traditional manual analysis methods are inefficient and cannot handle large-scale datasets effectively. This challenge has led to the development of automated sentiment analysis techniques that can process and classify textual data efficiently.

Sentiment analysis, also known as opinion mining, is a subfield of Natural Language Processing that focuses on identifying and extracting subjective information from text. It aims to determine whether a piece of text expresses a positive, negative, or neutral sentiment. In the context of movie reviews, sentiment analysis helps identify whether a review reflects a positive or negative opinion about a movie.

Despite the advancements in sentiment analysis, analysing textual data remains challenging due to several factors. Reviews often contain informal language, sarcasm, negation words, and context-dependent expressions that complicate sentiment detection. Therefore, effective preprocessing and feature extraction techniques are necessary before applying machine learning algorithms.

In this project, a sentiment analysis framework is developed for IMDb movie reviews using Natural Language Processing techniques. The framework includes text preprocessing, TF-IDF feature extraction, and multiple machine learning models such as Logistic Regression, Naive Bayes, and Support Vector Machine. Additionally, an ensemble learning approach is used to combine predictions from multiple classifiers to improve overall accuracy.

The main objective of this study is to evaluate the effectiveness of different machine learning algorithms for sentiment classification and develop an efficient system capable of accurately predicting sentiment polarity in movie reviews.

II. LITERATURE REVIEW

Sentiment analysis has been widely studied in the field of Natural Language Processing and text mining. Researchers have proposed various techniques for extracting opinions from textual data, including machine learning approaches, rule-based systems, and deep learning methods.

One of the earliest studies in sentiment analysis was conducted by Pang and Lee, who applied machine learning algorithms such as Naive Bayes, Maximum Entropy, and Support Vector Machines to classify movie reviews. Their research demonstrated that machine learning techniques could effectively outperform traditional rule-based approaches in sentiment classification tasks.

Later, Maas et al. introduced the Large Movie Review Dataset, commonly known as the IMDb dataset. This dataset consists of 50,000 labelled movie reviews and has become a widely used benchmark dataset for sentiment analysis research. The availability of this dataset enabled researchers to develop and evaluate various machine learning models for opinion mining.

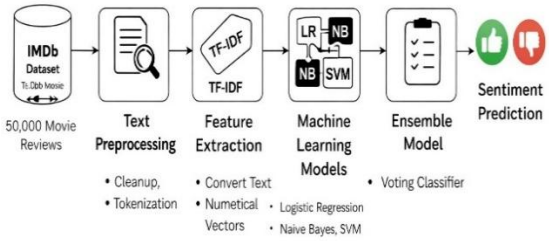
Feature extraction techniques also play a crucial role in sentiment analysis. One of the most commonly used methods is Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to words based on their importance within a document relative to the entire dataset. This approach helps highlight important words while reducing the influence of commonly occurring terms.

Support Vector Machines have also been widely applied in text classification problems due to their ability to handle high-dimensional data effectively. Studies have shown that SVM models often achieve high accuracy in sentiment analysis tasks, particularly when combined with appropriate feature extraction techniques.

More recently, researchers have explored deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for sentiment analysis. Although these models can achieve high performance, they require large computational resources and extensive training time.

Therefore, traditional machine learning models combined with effective feature extraction methods remain popular due to their computational efficiency, interpretability, and strong performance in text classification tasks.

III. SYSTEM ARCHITECTURE



The proposed sentiment analysis system follows a structured pipeline that converts raw textual reviews into sentiment predictions. The system begins by collecting movie reviews from the IMDb dataset.

These reviews are then processed through several preprocessing steps to clean and standardize the text.

After preprocessing, the cleaned text is transformed into numerical feature vectors using TF-IDF vectorization. These numerical representations capture the importance of words in the dataset and allow machine learning algorithms to analyse textual data effectively.
The feature vectors are then used to train multiple classification models, including Logistic Regression, Naive Bayes, and Support Vector Machine. Each model learns patterns in the training data and predicts the sentiment of unseen reviews.
To further improve prediction accuracy, an ensemble learning approach is implemented. The ensemble model combines the predictions of multiple classifiers and produces a final sentiment prediction based on the majority vote of the individual models.
Finally, the system evaluates the performance of the models using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

IV. METHODOLOGY

Data Preprocessing:
Text preprocessing is an essential step in sentiment analysis because raw textual data often contains noise and inconsistencies. In this project, several preprocessing techniques are applied to prepare the dataset for analysis.
Initially, all text is converted to lowercase to maintain consistency. Punctuation marks, special characters, and HTML tags are removed to eliminate unnecessary symbols from the dataset. Stop words such as “the,” “is,” and “and” are removed because they do not contribute significant meaning to sentiment classification.
Tokenization is then applied to split the text into individual words or tokens. Lemmatization is also performed to reduce words to their base form, which helps reduce dimensionality and improves model performance.

Feature Extraction using TF-IDF:
After preprocessing, the textual data is transformed into numerical features using the TF-IDF technique. TF-IDF measures the importance of a word in a document relative to the entire dataset.
Term Frequency measures how often a word appears in a document, while Inverse Document Frequency reduces the weight of commonly occurring words across the dataset. By combining these two measures, TF-IDF assigns higher weights to important words that are more likely to contribute to sentiment classification.
The TF-IDF representation creates a high-dimensional feature space where each word is represented as a numerical value. This numerical representation allows machine learning algorithms to process textual data effectively.

MACHINE LEARNING ALGORITHMS: Logistic Regression: Logistic Regression is a supervised learning algorithm commonly used for binary classification problems. In this project, it is used to classify movie reviews into positive or negative categories. Logistic Regression estimates the probability that a given review belongs to a particular class using a sigmoid function.

Naive Bayes: Naive Bayes is a probabilistic classification algorithm based on Bayes’ theorem. It assumes independence between features and calculates the probability that a review belongs to a particular sentiment class. Naive Bayes is widely used for text classification because it performs well with high-dimensional data.

Support Vector Machine: Support Vector Machine is a powerful machine learning algorithm used for classification tasks. It works by finding the optimal hyperplane that separates data points belonging to different classes. SVM is particularly effective in text classification because it can handle sparse and high-dimensional feature spaces.

Ensemble Method: The ensemble learning technique is used to improve the overall performance of the sentiment classification system. In this study, a Voting Ensemble classifier is implemented to combine the predictions from multiple machine learning models including Logistic Regression, Naive Bayes, and Support Vector Machine.

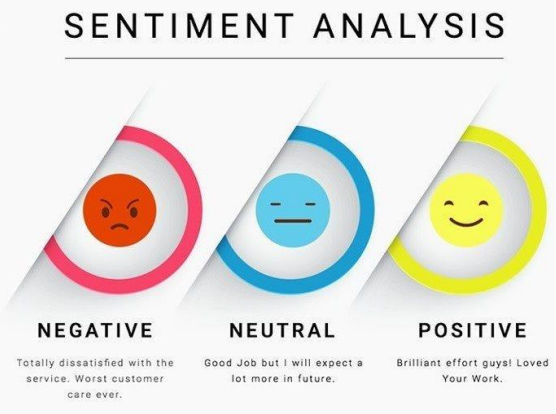
The Voting Ensemble works by collecting predictions from each individual classifier and determining the final sentiment based on the majority voting strategy.

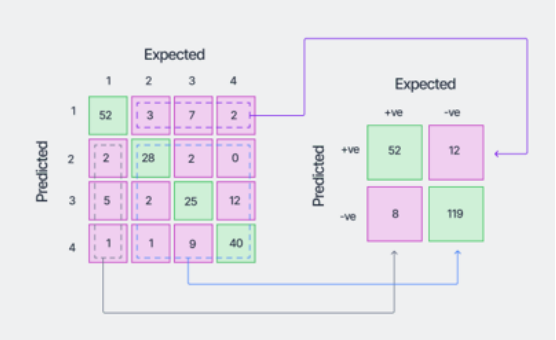
If two or more classifiers predict a positive sentiment, the final output is classified as positive; otherwise, it is classified as negative.

This hybrid approach leverages the strengths of multiple algorithms and reduces the weaknesses of individual models. By combining multiple classifiers, the ensemble model increases prediction stability, improves accuracy, and reduces the risk of misclassification.

V. RESULTS

The experimental evaluation of the proposed sentiment analysis system demonstrates the effectiveness of machine learning algorithms in classifying textual movie reviews into positive and negative sentiments. After preprocessing the dataset and converting the textual data into numerical vectors using the TF-IDF feature extraction method, the classification models were trained using the training dataset and evaluated using the testing dataset. The results indicate that the models were able to capture significant sentiment patterns present in the movie reviews and classify them with high accuracy.





Among the individual classifiers, Logistic Regression produced strong performance due to its ability to model relationships between textual features and sentiment labels effectively. Naive Bayes also performed efficiently because of its probabilistic nature and its ability to handle high-dimensional text data. Support Vector Machine achieved slightly higher accuracy compared to other individual models due to its capability to identify optimal decision boundaries within the feature space. This demonstrates that SVM is particularly suitable for text classification tasks where the feature space is sparse and high dimensional.

The ensemble model further improved the classification performance by combining predictions from Logistic Regression, Naive Bayes, and Support Vector Machine. By integrating multiple models, the ensemble approach reduced the effect of individual model errors and produced more reliable predictions. The improvement in accuracy indicates that combining multiple classifiers can significantly enhance the robustness and stability of sentiment classification systems. These results confirm that ensemble learning is an effective strategy for improving the performance of machine learning models in opinion mining tasks.

VI. CONCLUSION

The findings of this study demonstrate that machine learning techniques combined with Natural Language Processing methods can effectively analyse large volumes of textual data and extract meaningful insights from user opinions. By applying text preprocessing and TF-IDF feature extraction, the system was able to convert unstructured textual movie reviews into structured numerical representations suitable for machine learning models.

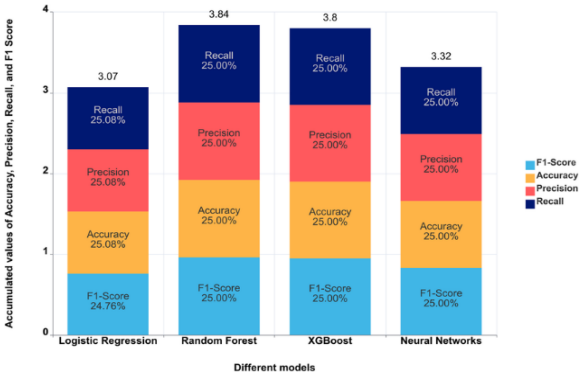
The experimental results show that Logistic Regression, Naive Bayes, and Support Vector Machine are capable of performing sentiment classification with high accuracy. Among these models, Support Vector Machine achieved strong performance due to its ability to identify optimal classification boundaries within high-dimensional feature spaces. Furthermore, the ensemble learning approach improved overall system performance by combining predictions from multiple classifiers and reducing the impact of individual model errors.

Overall, the proposed sentiment analysis framework provides a scalable and efficient solution for opinion mining in textual datasets. The system can be applied to various domains such as product reviews, social media analysis, and customer feedback evaluation. Future improvements may involve incorporating advanced deep learning techniques and contextual language models to enhance the ability of the system to understand complex linguistic patterns and improve sentiment prediction accuracy.

The trained sentiment classification model was also deployed as a real-time web application using the Flask framework, allowing users to enter movie reviews and instantly obtain sentiment predictions through an interactive interface.

VII. DISCUSSION

The results obtained from this research highlight the importance of proper feature extraction and model selection in sentiment analysis tasks. The TF-IDF technique played a significant role in identifying important words that contribute to sentiment classification.



By assigning higher weights to meaningful words and reducing the influence of common words, TF-IDF helped the machine learning algorithms focus on the most relevant textual features within the dataset. This improved the ability of the models to differentiate between positive and negative sentiments in movie reviews.

Another important observation from the experimental analysis is the variation in performance among different classification algorithms. Logistic Regression provided stable and interpretable results, while Naive Bayes demonstrated faster computation and efficient learning even with large datasets. Support Vector Machine showed superior performance due to its ability to handle complex decision boundaries and high-dimensional data spaces. These differences highlight the strengths and limitations of each algorithm when applied to text classification problems.

Despite achieving high accuracy, certain challenges remain in sentiment analysis. Some movie reviews contain mixed opinions where both positive and negative sentiments appear in the same text, making classification more difficult. Additionally, sarcasm, irony, and contextual expressions can sometimes lead to misclassification. These challenges indicate that while traditional machine learning models perform well, there is still room for improvement. Future research can explore deep learning models and contextual language representations to further enhance sentiment analysis performance.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proc. ACL-02 Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.

[2] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” Proc. 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 142–150.

[3] J. Ramos, “Using TF-IDF to determine word relevance in document queries,” Proc. First Instructional Conf. on Machine Learning, 2003.

[4] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[5] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, Sebastopol, CA, USA: O’Reilly Media, 2009.

[7] A. Zhang, Z. Lipton, M. Li, and A. Smola, Dive into Deep Learning, Cambridge University Press, 2021.

[8] Y. Liu, J. Ott, N. Goyal, et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, 2019, pp. 4171–4186.
[10] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," Advances in Neural Information Processing *Systems*, vol. 28, 2015.
[11] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
[12] M. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.
[13] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF-IDF, LSI and multi-words for text classification," Expert Systems with Applications, vol. 38, no. 3, pp. 2758–2765, 2011.
[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv*:1301.3781*, 2013.
[15] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Systems, vol. 89, pp. 14–46, 2015.
[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[17] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.
[18] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," Proc. EMNLP, 2013.
[19] S. Kiritchenko, X. Zhu, and S. Mohammad, "Sentiment analysis of short informal texts," Journal of Artificial Intelligence Research, vol. 50, pp. 723–762, 2014.