Volume 9, Issue 5, May – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

Sentiment Analysis of IMDb Movie Reviews

S.M. Yousuf Iqbal Tomal
Department of Computer Science and Engineering
BRAC University Dhaka, Bangladesh.

Abstract:- This paper presents a sentiment analysis project focusing on IMDb movie reviews, aimed at classifying reviews as either positive or negative based on their textual content. Utilizing a dataset of 50,000 IMDb movie reviews, sourced from Kaggle, the study addresses the binary classification challenge by employing pre-processing techniques such as TF-IDF vectorization. The dataset is split into training and testing sets, with models trained on the former and evaluated on the latter. Three machine learning algorithms—Logistic Regression, Random Forest, and Decision Tree—are implemented and compared using performance metrics including precision, recall, and F1-score. Results indicate that Logistic Regression outperforms other models in sentiment analysis classification. The report concludes by highlighting the project's contributions and suggesting avenues for future research, emphasizing the potential benefits of expanding sentiment types and dataset size.

I. INTRODUCTION

Sentiment analysis, an integral task in natural language processing, plays a pivotal role in discerning the polarity of textual content, thereby facilitating informed decision-making in various domains (Maas et al. 2011; Kiritchenko et al. 2014; Poria et al. 2016; Rodr´ıguez-Fernandez and Ortega 2017). In this context, this´ paper delves into the realm of sentiment analysis through the lens of IMDb movie reviews. The objective is to classify reviews as either positive or negative, leveraging a dataset comprising 50,000 labeled movie reviews sourced from Kaggle. By employing machine learning techniques, particularly logistic regression, random forest, and decision tree algorithms, this study endeavors to unravel the sentiment embedded within the textual corpus.

The significance of sentiment analysis in the realm of movie reviews lies in its potential to offer valuable insights to filmmakers, production houses, and audiences alike. Understanding audience sentiment towards a particular movie not only aids filmmakers in gauging audience reception but also empowers viewers to make informed choices regarding their cinematic preferences. Therefore, this paper seeks to contribute to the burgeoning field of sentiment analysis by addressing the specific nuances inherent in movie reviews.

Structured into several sections, this paper begins with an overview of the dataset utilized, delineating its characteristics and the binary classification problem it poses. Subsequently, the pre-processing steps involved in preparing the textual data for analysis are elucidated, including TF-IDF vectorization and data segmentation. Following this, the dataset is partitioned into training and testing sets to facilitate model training and evaluation.

II. RELATED WORK

The task of sentiment analysis, also known as opinion mining, has been extensively studied in various domains to extract and analyze sentiment or subjective information from text data. Maas et al. (2011) conducted a seminal work on sentiment analysis by proposing a method for learning word vectors to perform sentiment classification (Maas et al. 2011). Kiritchenko et al. (2014) focused on sentiment analysis of short informal texts, addressing the challenges of sentiment classification in informal language (Kiritchenko et al. 2014). Poria et al. (2016) explored aspect extraction for opinion mining using deep convolutional neural networks, contributing to the advancement of sentiment analysis techniques (Poria et al. 2016). Rodr´ıguez-Fernandez´ and Ortega (2017) analyzed the factors influencing sentiment analysis accuracy, shedding light on the challenges and opportunities in this field (Rodr´ıguez-Fernandez and Ortega 2017).´

In addition to sentiment analysis, research has investigated the impact of social question-and-answer (Q&A) sites on knowledge sharing in various communities. Vasilescu et al. (2015) examined how social Q&A sites are changing knowledge sharing dynamics in open source software communities, highlighting the role of social platforms in facilitating knowledge exchange (Vasilescu et al. 2015).

III. DATASET DESCRIPTION

The dataset utilized in this study comprises 50,000 IMDb movie reviews, sourced from Kaggle, a popular platform for hosting datasets and machine learning competitions. Each movie review within the dataset is labeled with its corresponding sentiment, categorized as either positive or negative. This binary classification scheme simplifies the sentiment analysis task, enabling the classification of reviews based on their polarity.

The IMDb movie review dataset offers a diverse and comprehensive collection of textual data, encompassing a wide range of movie genres, release dates, and viewer preferences. With an equal distribution of 25,000 positive and 25,000 negative reviews, the dataset provides a balanced representation of sentiment polarity, thereby minimizing bias during model training and evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

Each movie review in the dataset is represented as a text document, typically containing multiple sentences expressing the reviewer's opinion, critique, or evaluation of a particular film. The reviews vary in length, tone, and language style, reflecting the diversity of opinions prevalent among IMDb users.

Sentiment Count Table:

Sentiment	Count
positive	25000
negative	25000

Fig 1 Sentiment Count

Prior to model training, the dataset undergoes preprocessing steps to ensure compatibility with machine learning algorithms. Text data is transformed into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, which converts the textual corpus into a matrix of TF-IDF values representing the importance of each term in the context of the entire dataset.

The IMDb movie review dataset serves as a valuable resource for training and evaluating sentiment analysis models, offering researchers and practitioners a standardized benchmark for assessing the efficacy of different algorithms and techniques in classifying textual data based on sentiment polarity. The inclusion of this dataset in the study enables reproducibility and comparison with other sentiment analysis approaches, fostering collaboration and knowledge sharing within the research community.

First few rows of the dataset:

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Fig 2 First Few Rows of the Dataset

IV. METHODOLOGY

➢ *Data Loading and Preparation*

The IMDb movie review dataset is loaded from a CSV file obtained from Kaggle. The dataset contains two columns: 'review' and 'sentiment', representing the textual content of the reviews and their corresponding sentiment labels (positive or negative), respectively.

➢ *Exploratory Data Analysis (EDA)*

A preliminary analysis of the dataset is conducted to understand the distribution of classes (positive and negative sentiments) using visualizations, specifically a count plot. This step provides insights into the balance of classes within the dataset.

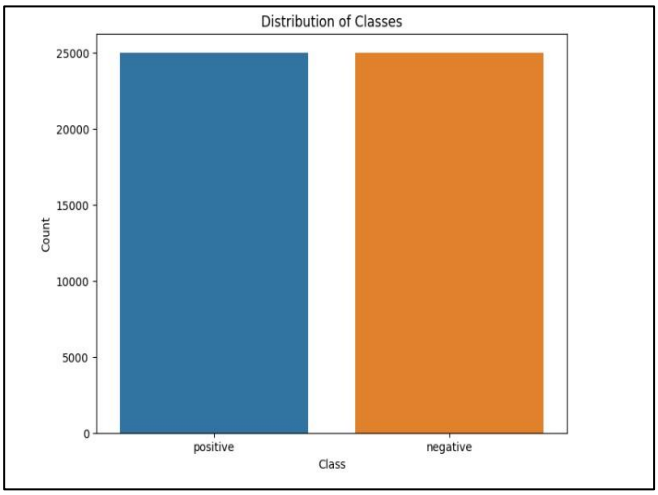

Fig 3 Distribution of Classes

V. EXPERIMENTAL SETUP

In this section, we provide a detailed overview of the experimental setup, including the process of splitting the dataset into training and testing sets.

A. Data Splitting

The IMDb movie review dataset was randomly split into training and testing sets using the traintestsplit function from the sklearn.modelselection module. The dataset, consisting of 50,000 reviews, was divided such that 80% of the data was allocated to the training set, while the remaining 20% was assigned to the testing set. This split ensured a sufficient amount of data for model training while preserving a separate set of unseen data for model evaluation.

B. Dataset Preprocessing

Before training the models, the dataset undergoes several preprocessing steps. This includes removing any irrelevant characters, such as punctuation marks and special symbols, and converting the text data to lowercase to ensure uniformity.

C. TF-IDF Vectorization

The TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique is used to convert the preprocessed text data into numerical features. TF-IDF assigns each term in the document a weight that reflects its importance in the document relative to the entire corpus. TF-IDF vectorization involves the following steps:

➢ *Term Frequency (TF) Calculation:*

The frequency of each term in the document is computed. TF is calculated using the formula:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

$$\text{TF}(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

➢ *Inverse Document Frequency (IDF) Calculation:*

The importance of each term in the entire corpus is computed. IDF is calculated using the formula:

$$\text{IDF}(t,D) = \log \frac{\text{Total number of documents in the corpus } N}{\text{Number of documents containing term } t}$$

➢ *TF-IDF Calculation:*

The TF-IDF score for each term in the document is computed as the product of its TF and IDF scores:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

TF-IDF vectorization transforms the text data into a matrix representation, where each row represents a document and each column represents a term, with the cell values being the TF-IDF scores.

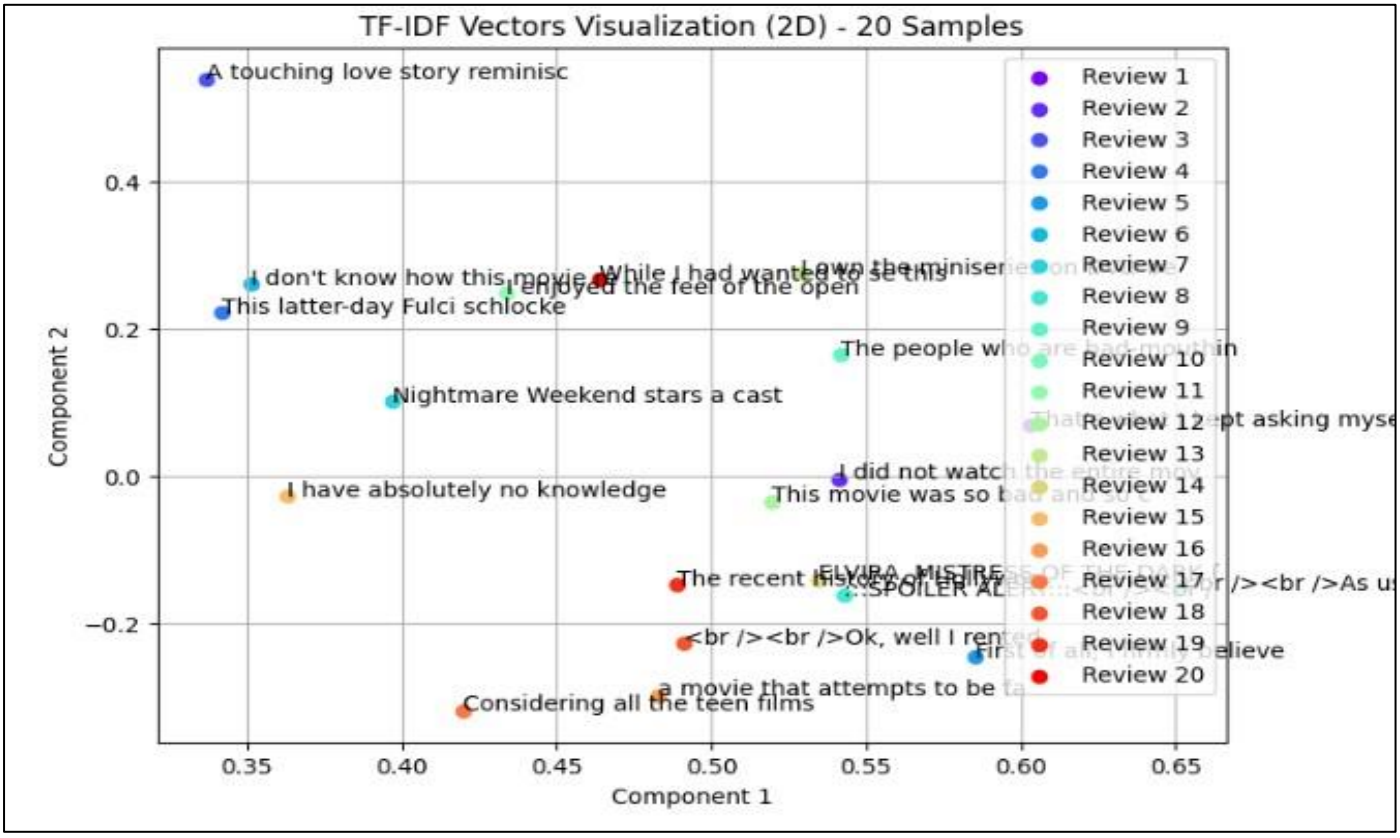

Fig 4 TF-IDF Vectors Visualization (2D)

➢ *In this Visualization:*

- Each point represents a 'Review' in the dataset.
- The position of a point in the 2D space is determined by the TF-IDF scores of the review's terms,reduced to two dimensions using TruncatedSVD.
- The colors distinguish different reviews in the dataset.
- Annotations provide a glimpse of the text content of each review (truncated for clarity).

TF-IDF vectorization is widely used in text mining and natural language processing tasks due to its ability to capture the semantic meaning of text data and identify important features for analysis.

D. Model Training

In this subsection, we describe the training process of the machine learning models used for sentiment analysis of IMDb movie reviews. We employ logistic regression, random forest, and decision tree algorithms for classification tasks.

➢ *Logistic Regression*

Logistic regression is a widely used statistical method for binary classification. Given a set of input features $X = \{x_1,x_2,...,x_n\}$, logistic regression models the probability $P(y = 1|X)$ of the positive class y as a logistic function of the linear combination of the input features:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

Where $\beta_0,\beta_1,\beta_2,...,\beta_n$ are the model parameters learned during training.

➢ *Random Forest*

Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. Each decision tree is trained on a bootstrap sample of the training data, and a random subset of features is considered at each split to promote diversity among the trees. The final prediction of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

random forest model is determined by aggregating the predictions of all individual trees.

➢ *Decision Tree*

A decision tree is a flowchart-like structure where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents a class label. Decision trees are simple yet powerful models for classification tasks and are particularly interpretable. The splitting criterion, such as Gini impurity or information gain, is used to determine the feature that best separates the data at each node.

After training the logistic regression, random forest, and decision tree models on the vectorized training data, predictions were made using the respective models on the vectorized test data. The performance of each model will be evaluated and compared in the subsequent evaluation section to determine their effectiveness in sentiment analysis.

VI. RESULTS

In this section, we present the results of our experiments, including the performance of each model on the testing set and any visualizations used for comparison.

➢ *Performance Metrics*

The performance of each model on the testing set was evaluated using standard metrics, including precision, recall, and F1-score. Table 1 provides a summary of these metrics for each model.

Table 1 Model Performance on Testing Set

Model	Precision	Recall	F1-score
Logistic Regression	0.887	0.909	0.898
Random Forest	0.857	0.839	0.848
Decision Tree	0.716	0.711	0.714

➢ *Visualizations*

We present visualizations to facilitate the comparison of model performances. Figure 5 displays bar charts illustrating the precision, recall, and F1-score of each model.

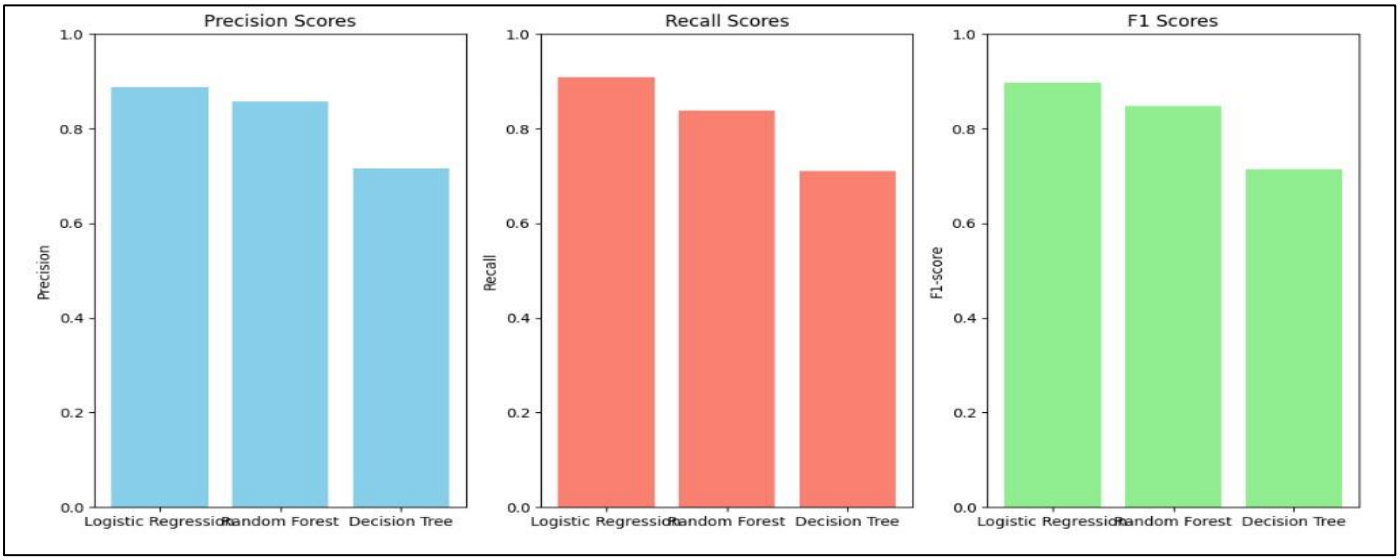

Fig 5 Bar Charts of Model Performance Metrics

Additionally, confusion matrices were generated to visualize the distribution of true positive, true negative, false positive, and false negative predictions for each model. Examples of confusion matrices for Logistic Regression, Random Forest, and Decision Tree models are shown in Figures 6, 7, and 8, respectively.

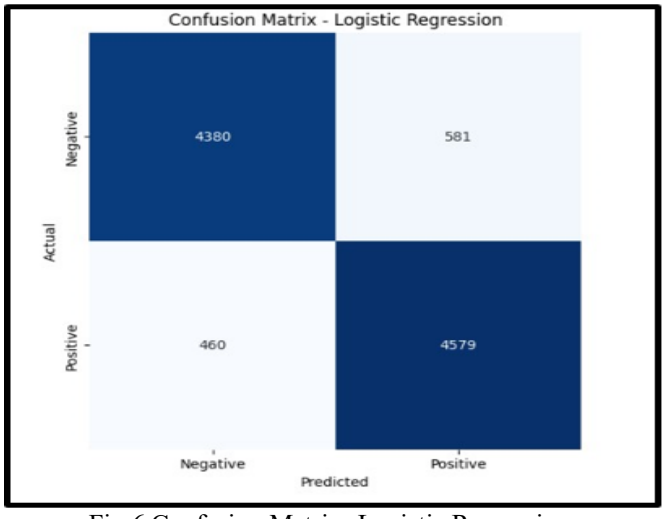

Fig 6 Confusion Matrix -Logistic Regression

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

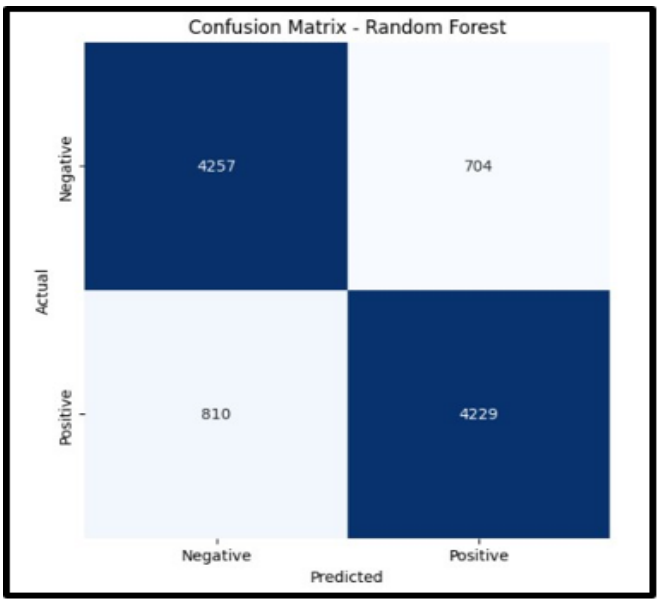

Fig 7 Confusion Matrix - Random Forest

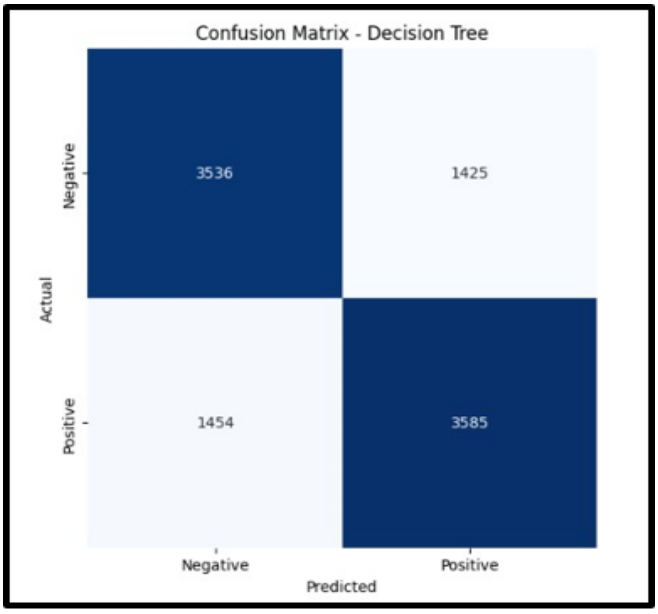

Fig 8 Confusion Matrix - Decision Tree

VII. DISCUSSION

In this section, we interpret the results of our experiments, discuss the implications of our findings, address any limitations or challenges encountered during the study, and suggest areas for future research.

➢ *Interpretation of Results*

Our experiments revealed that Logistic Regression performed better than Random Forest and Decision Tree models across multiple evaluation metrics such as precision, recall, and F1-score. Specifically, Logistic Regression achieved the highest precision, recall, and F1-score values of 0.887, 0.909, and 0.898, respectively. This suggests that Logistic Regression, with its linear decision boundary, is particularly well-suited for analyzing sentiment in IMDb movie reviews.

On the other hand, Random Forest, despite its ability to capture complex relationships within the data due to its ensemble nature, exhibited slightly lower performance compared to Logistic Regression, with an F1-score of 0.848. This indicates that while Random Forest is powerful, it may not be the optimal choice for sentiment analysis in this context.

Similarly, the Decision Tree model, known for its simplicity and interpretability, showed the lowest performance among the three models, with an F1-score of 0.714. This suggests that Decision Trees struggle to handle the nuances of natural language data effectively.

Overall, our findings emphasize the importance of selecting the appropriate model for sentiment analysis tasks. While more complex models like Random Forest may seem promising, simpler models like Logistic Regression can often yield superior results in specific scenarios, such as analyzing IMDb movie reviews.

➢ *Implications and Future Research*

Our findings have several implications for sentiment analysis tasks in the domain of movie reviews. Firstly, the effectiveness of Logistic Regression highlights the importance of considering both linear and nonlinear relationships in text data analysis. Future research could explore hybrid approaches that combine the strengths of linear and nonlinear models to further improve sentiment classification accuracy.

Additionally, while our study focused on binary sentiment classification (positive or negative), future research could extend the analysis to include more nuanced sentiment categories (e.g., neutral, mixed) to provide deeper insights into audience reactions towards movies.

Furthermore, our study encountered limitations related to the dataset size and feature representation. Expanding the dataset size and exploring advanced text representation techniques, such as word embeddings or contextual embeddings, could enhance model performance and generalizability.

Lastly, investigating the impact of domain-specific features, such as movie genres, actors, or release years, on sentiment analysis could lead to more tailored and context-aware sentiment classification models for movie reviews.

Overall, our study contributes to the understanding of sentiment analysis in movie reviews and paves the way for future research endeavors in this domain.

VIII. CONCLUSION

In this study, we conducted sentiment analysis on IMDb movie reviews using machine learning models. Our key findings can be summarized as follows:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1625

- Logistic Regression outperformed Random Forest and Decision Tree models in terms of precision,recall, and F1-score.
- The linear decision boundary of Logistic Regression proved effective for sentiment classification onIMDb movie reviews.
- Our results highlight the importance of considering both linear and nonlinear relationships in text dataanalysis.
- Future research could explore hybrid approaches combining the strengths of linear and nonlinear mod-els to improve sentiment classification accuracy.

In the context of sentiment analysis and IMDb movie reviews, our study underscores the significance of employing appropriate machine learning algorithms and feature representations to accurately classify sentiments expressed in textual data. By providing insights into audience reactions towards movies, sentiment analysis contributes to informed decision-making processes in the film industry, including marketing strategies, content creation, and audience engagement.

Overall, our study contributes to advancing sentiment analysis techniques in the domain of movie reviews and emphasizes the importance of leveraging machine learning methods for extracting valuable insights from textual data.

REFERENCES

[1]. Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[2]. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedingsofthe49thAnnualMeetingoftheAssociation for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.

[3]. Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

[4]. Sonia Rodr´ıguez-Fernandez and Francisco Ortega. Analysis of the factors influencing sentiment analysis´ accuracy. *Expert Systems with Applications*, 77:185–200, 2017.

[5]. Bogdan I. Vasilescu, Alexander Serebrenik, and Premkumar Devanbu. How social q&a sites are changing knowledge sharing in open source software communities. *IEEE Transactions on Software Engineering*, 41 (9):900–912, 2015.