# Sentiment Analysis of IMDb Movie Reviews Using Various Machine Learning Models

Shashwat Tiwari 2010111038, Samyak Jain 2010110763 and Ayaan Khan 2010110175

***Shiv Nadar Institute of Eminence***

## Abstract

Sentiment analysis (SA) plays a crucial role in understanding public opinions, emotions, and preferences in various fields. This study focuses on the performance of different machine learning models for sentiment analysis of movie reviews from the IMDb dataset. We propose a comprehensive pre-processing pipeline tailored for movie reviews and train various models, including Logistic Regression, Decision Tree, Random Forest, KNN, SVM with different C values, Multinomial Gaussian Naive Bayes, and Gaussian Naive Bayes. The objective is to compare the performance of these models and identify the most suitable one for sentiment analysis of movie reviews. Additionally, we employ multiple data visualization techniques to better interpret and present the results obtained from our experiments.

# Introduction

Sentiment analysis, also known as opinion mining, is a critical task in natural language processing (NLP) that aims to classify a piece of text into different sentiment categories such as positive, negative, or neutral. It is particularly valuable in domains where un- derstanding public opinions, emotions, and preferences can provide actionable insights and contribute to better decision-making processes. In recent years, sentiment analysis has gained prominence due to the exponential growth of user-generated content on social media platforms, review websites, and online forums.

In this project, we focus on analyzing movie reviews from the IMDb dataset to perform sentiment analysis. Movie reviews provide insights into the public’s perception of films, which can be valuable to the film industry, critics, and viewers. By understanding the sentiments expressed in these reviews, stakeholders can make informed decisions regard- ing film production, promotion, and distribution.

Our research aims to develop a comprehensive and effective approach to perform sen- timent analysis on movie reviews. To achieve this, we propose a robust pre-processing pipeline tailored for movie reviews, which involves cleaning the text, tokenization, lemma- tization, stemming, and generating visualizations such as word clouds and histograms to understand review length and character distributions for both positive and negative re- views. Following the pre-processing steps, we train a diverse selection of machine learning models, including Logistic Regression, Decision Tree, Random Forest, KNN, SVM with different C values, Multinomial Gaussian Naive Bayes, and Gaussian Naive Bayes. By training and evaluating multiple models, we aim to identify the most suitable model for sentiment analysis of IMDb movie reviews.

# Literature Review/Related Works.

## ”An Effective Tokenization Algorithm for Information Retrieval Sys- tems” [1]

This research paper proposes a tokenization algorithm for information retrieval systems to enhance their efficiency and effectiveness. As the volume of web data increases, user expectations for precise results have grown, which depend on docu- ment storage and indexing patterns. Tokenization, an essential pre-processing step, generates tokens that reduce search space and form the basis for probabilistic IR scoring.

The paper introduces a tokenization approach based on document vectors, ensuring accurate token identification. It highlights the importance of ranking algorithms in finding relevant documents for user queries and notes that indexing algorithms’ effectiveness relies heavily on the quality of tokens generated during tokenization. The paper concludes that there is still room for improvement in the algorithm’s accuracy and efficiency.

In summary, this paper presents a tokenization approach based on document vec- tors, significantly improving the efficiency and effectiveness of information retrieval systems while reducing search space and storage requirements.

## ”Analyzing Sentiment using IMDb Dataset”[2]

It demonstrates the value of traditional machine learning (ML) algorithms in per- forming Sentiment Analysis (SA) on large volumes of textual data collected from the internet. Sentiment analysis, which detects the polarity of a statement as positive, negative, or neutral, is essential for understanding customer feedback, improving product quality, and monitoring market performance.

The paper presents an approach to perform sentiment analysis on the IMDb movie reviews dataset. The process involves retrieving the dataset, pre-processing the data by cleaning and removing HTML tags and stopwords, feature selection, text repre- sentation using Bag of Words (BoW), and feeding it to classifiers like Na¨ıve Bayes, Logistic Regression, Random Forest, and Decision Tree. Classifier performance is evaluated using six metrics, including confusion matrix, accuracy, precision, recall, F1 measure, and Area Under Curve (AUC).

In summary, this paper emphasizes the potential of traditional machine learning al- gorithms in analyzing sentiment and extracting valuable insights from vast amounts of textual data. The proposed approach demonstrates efficient processing and sen- timent analysis using the IMDb dataset, with a focus on evaluation metrics to determine algorithm efficiency and model performance.

## ”Application of Machine Learning for Sentiment Analysis of Movies Us- ing IMDB Rating”[3]

This paper proposes a framework using data mining and machine learning algo- rithms to analyze customer sentiment towards movies based on IMDB ratings. Sentiment analysis is crucial for understanding consumer feedback, and this com- putational approach classifies opinions as positive, negative, or neutral.

The study emphasizes the importance of sentiment analysis in areas like product, brand, and social media sentiment analysis. With most activity occurring online, companies can utilize social media to verify product reviews and analyze customer behavior. However, large data volumes and complex information require an efficient framework that can classify customer behavior using advanced machine learning algorithms.

The proposed framework involves data preprocessing and feature extraction, text vectorization, and the use of machine learning algorithms like Na¨ıve Bayes, Sup- port Vector Machine, and Logistic Regression. The study applies the framework to IMDB movie ratings and evaluates the accuracy, precision, and recall of the classification. Results show that the best performance is achieved with the SVM algorithm.

In conclusion, this paper offers a practical framework for movie sentiment analysis using IMDB ratings, emphasizing the importance of sentiment analysis in consumer behavior analysis and decision-making. The framework can be applied to various ap- plications, and the results demonstrate its potential for improving decision-making processes.

## ”On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes” [4]

In the realm of related works, the research paper titled “On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes” by Ng and Jordan (2001) presents a comprehensive comparison between discriminative and generative learning approaches used in classification tasks. The authors specifically examine logistic regression and naive Bayes algorithms to understand the conditions under which one outperforms the other.

The paper challenges the prevailing belief that discriminative classifiers are always superior to generative classifiers, highlighting that the performance of each algo- rithm is influenced by the training set size. Ng and Jordan (2001) demonstrate that while discriminative learning has lower asymptotic error, a generative classi- fier may approach its higher asymptotic error much faster.

The authors provide an insightful comparison between the naive Bayes model and its discriminative counterpart, logistic regression, showing that the generative model can perform better in certain situations. They identify two distinct regimes of performance based on the number of training examples, indicating that the choice between generative and discriminative classifiers depends on the specific problem and dataset at hand.

In conclusion, this research paper emphasizes the importance of evaluating both generative and discriminative classifiers for any given classification task, as the trade-offs between the two approaches should be carefully considered to determine the most suitable algorithm. This finding is particularly relevant to our project, as we explore various machine learning models, including logistic regression and naive Bayes, to determine the best-performing model for sentiment analysis of IMDb movie reviews.

## ”Sentiment Analysis on Movie Reviews Using Information Gain and K- Nearest Neighbor” [5]

Another relevant study in the context of our project is the research conducted on the effectiveness of various machine learning methods for sentiment analysis, including Naive Bayes (NB), K-nearest neighbor (KNN), Support vector machine (SVM), and Random Forest. In this study, the authors utilized the Polarity v2.0 dataset from the Cornell movie review dataset to test KNN with Information Gain feature selection in order to achieve better performance. The primary objectives of the research were to identify the optimal value for K in KNN and to compare KNN’s performance with other machine learning methods.

The study found that the optimal value for K in KNN is equal to 3 for the Po- larity v2.0 dataset. By employing 10-fold Cross Validation, the authors compared the performance of KNN, NB, SVM, and Random Forest both with and without Information Gain feature selection. While KNN initially achieved only 0.6 perfor- mance with K=3, its performance significantly improved to 0.968 after utilizing Information Gain, surpassing the performance of NB, SVM, and Random Forest.

The authors concluded that reducing irrelevant features has a greater impact on KNN than on other methods, and that feature selection with Information Gain improves the performance of all machine learning methods. They emphasized that the appropriate amount and quality of features are essential for better performance, particularly for KNN. The results of this study underscore the importance of effec- tive feature selection for achieving optimal performance in sentiment analysis tasks, which is relevant to our project as we compare the performance of different machine learning models, including logistic regression and naive Bayes, in analyzing IMDb movie reviews.

# Dataset Description [6]

[The IMDb dataset](https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?resource=download) used in this study consists of 50,000 movie reviews collected for the purpose of natural language processing and text analytics. This dataset is designed for binary sentiment classification, making it a valuable resource for understanding and analyzing the polarity of reviews as positive or negative. With substantially more data than previous benchmark datasets, the IMDb dataset offers a more comprehensive and robust training and testing ground for sentiment analysis. The dataset is divided into two

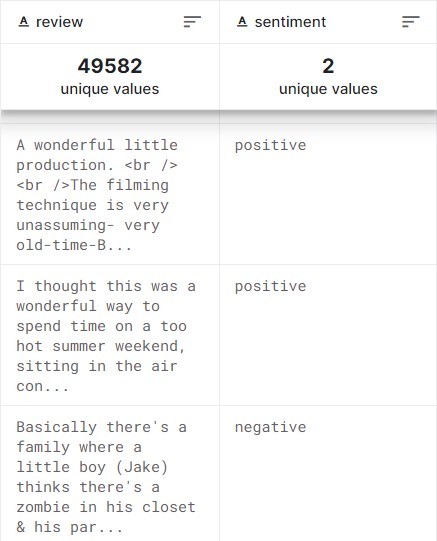
“ 

Figure 1: Dataset format

equal parts, with 25,000 highly polar movie reviews designated for training and another

25,000 for testing. This balanced split allows us to train various machine learning or deep learning algorithms on a large, representative sample and evaluate their performance on a separate, unseen set of reviews.

The reviews in the IMDb dataset are highly polar, meaning that they exhibit strong positive or negative sentiments. This characteristic of the data ensures that the models are able to learn from clear, distinct examples of both positive and negative sentiments. By predicting the number of positive and negative reviews using classification or deep learning algorithms, we can develop a better understanding of the underlying patterns and relationships in the movie review data and evaluate the effectiveness of different ap- proaches to sentiment analysis.

# Objective

The primary objective of this project is to develop a comprehensive and effective ap- proach to perform sentiment analysis on the IMDb movie reviews dataset. This dataset consists of 50,000 reviews, labeled as either positive or negative. By gaining a deeper understanding of the sentiment expressed in these reviews, our goal is to provide valu- able insights into customer opinions and preferences, ultimately contributing to better decision-making and performance in the entertainment industry.

To achieve this objective, we will undertake the following steps:

* Preprocess the IMDb movie reviews dataset by applying techniques such as tok- enization, lemmatization, stemming, and differentiating reviews based on length and polarity. This will prepare the data for model training and ensure a more accurate analysis.
* Visualize the most relevant words for both positive and negative reviews using word clouds. This will help identify common themes and sentiment-related terms in the dataset.
* Train various machine learning models, including Logistic Regression, Decision Tree, Random Forest, KNN, SVM with different C values, Multinomial Gaussian Naive Bayes, and Gaussian Naive Bayes, for sentiment analysis on the preprocessed dataset. This will allow us to explore the performance of different algorithms and identify the best-suited model for this task.
* Evaluate and compare the performance of the trained models using different metrics, such as accuracy, precision, recall, and F1 score. This will help determine the effectiveness of each model in predicting the sentiment of movie reviews.
* Identify the most suitable model for sentiment analysis of IMDb movie reviews based on the evaluation results. This will provide a reliable and accurate tool for understanding customer sentiment in the context of movie reviews.

By following these steps, we aim to create an effective approach to perform sentiment analysis on IMDb movie reviews and provide valuable insights into customer sentiment within the entertainment industry.

# Proposed Model

We will train and evaluate the following machine learning models for sentiment analysis on the IMDb movie reviews dataset. We will perform this process using both the original set of features as well as with feature selection for comparison.

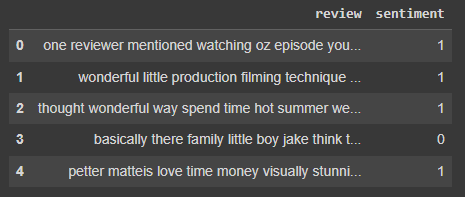


Figure 2: snippet of dataset values

* Logistic Regression: A linear classification model that uses the logistic function to estimate the probability of a review belonging to a specific sentiment class. We will train this model with and without feature selection to observe the impact on its performance.
* Decision Tree: A tree-based model that recursively splits the feature space into branches based on feature values. We will build decision trees with and without feature selection to assess the effect of reducing the number of features on the model’s accuracy.
* Random Forest: An ensemble method that combines multiple decision trees to produce a more accurate and robust classifier. We will train random forests with and without feature selection to analyze the performance changes due to feature selection.
* KNN: The k-Nearest Neighbors algorithm is a distance-based classifier that assigns a class to a review based on the majority class of its k closest neighbors. We will compare the performance of KNN models with and without feature selection.

SVM with C=1: Support Vector Machines are a set of algorithms that find the optimal hyperplane separating different classes. We will train SVM models with different regularization parameters (C=1) and compare their performance with and without feature selection.

* SVM with C=3: Similarly, we will train SVM models with another regularization parameter (C=3) to explore the effect of different regularization strengths in com- bination with feature selection.
* Multinomial Gaussian Naive Bayes: A probabilistic model that assumes indepen- dence between features and calculates class probabilities using the Bayes theorem. We will train this model with and without feature selection to examine its perfor- mance under different feature sets.

“

Figure 3: Word cloud positive reviews



Figure 4: Word cloud negative reviews

* Gaussian Naive Bayes: Similar to Multinomial Gaussian Naive Bayes, this model uses Gaussian distribution assumptions for the features. We will compare the per- formance of Gaussian Naive Bayes models with and without feature selection.

For each of the proposed models, we will follow the process: a . Split the dataset into training and testing sets.

b . Train the model on the training set with and without feature selection.

c . Evaluate the model on the testing set using performance metrics such as accuracy, precision, recall, and F1 score.

d . Compare the performance of the models with and without feature selection to understand the impact of feature selection on each algorithm.

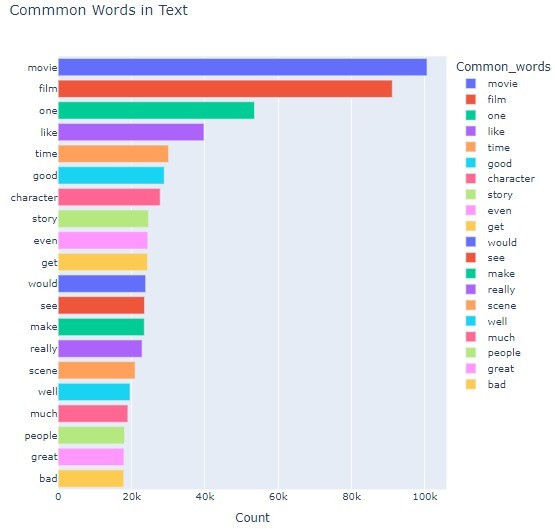


Figure 5: Common words chart

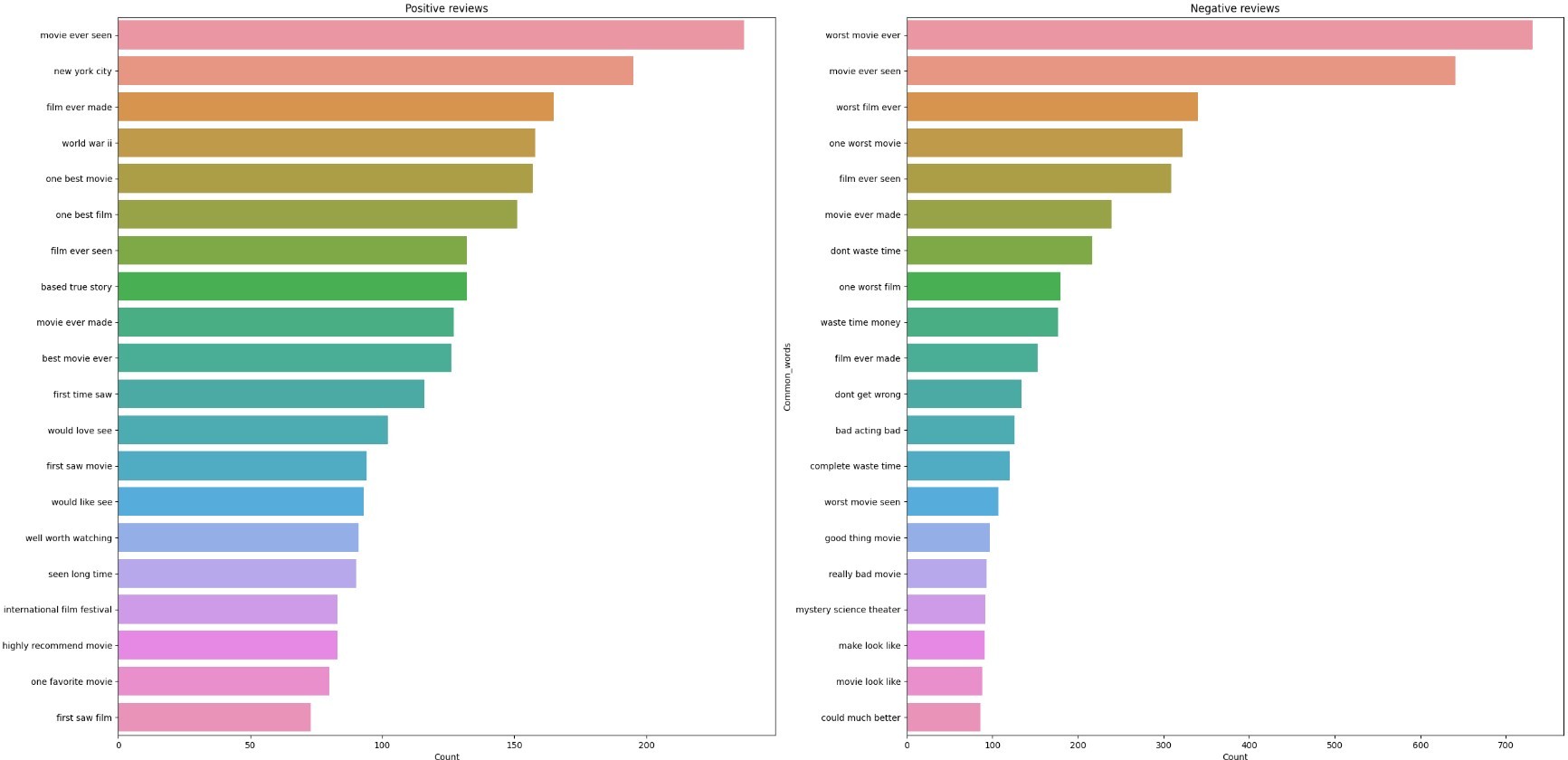


Figure 6: Common tri-grams chart for positive and negative reviews.

By evaluating the proposed models with and without feature selection, we will gain in- sights into the effectiveness of the feature selection process and the performance of each model in the context of IMDb movie reviews sentiment analysis.

# Methodology

Our methodology for sentiment analysis on the IMDb movie reviews dataset involves four main steps: data preprocessing, feature engineering, model training, and model evaluation.

* **Data Preprocessing**: Clean and normalize the dataset by removing irrelevant information, HTML tags, and stopwords. Tokenize, lemmatize, and stem the text, and generate visualizations such as word clouds and histograms to understand re- view length and character distributions for both positive and negative reviews.

Refine preprocessing techniques based on visualization analysis to improve input data quality.

* **Feature Engineering**: Convert preprocessed text data into numerical features using techniques like CountVectorizer or TF-IDF, allowing models to effectively analyze textual data.
* **Model Training**: Train a diverse selection of machine learning models, including Logistic Regression, Decision Tree, Random Forest, KNN, SVM with different C values, Multinomial Gaussian Naive Bayes, and Gaussian Naive Bayes on the trans- formed dataset. This variety enables exploration of different algorithms to identify the best-suited model for the task.
* **Model Evaluation**: Assess the performance of trained models using metrics such as accuracy, precision, recall, and F1 score. Compare model performance to deter- mine the most effective one for predicting the sentiment of movie reviews.

By following this methodology, we aim to create an efficient approach to perform sentiment analysis on IMDb movie reviews, providing valuable insights into cus- tomer sentiment within the entertainment industry..

# Experimentation and Results

In our experimentation process, we trained several models on the IMDb movie reviews dataset, including Gaussian Naive Bayes. Although Gaussian Naive Bayes is often effec- tive for text classification, it relies on the assumption that the features follow a Gaussian distribution. Unfortunately, our dataset’s features did not adhere to this distribution, which resulted in suboptimal model performance.

To address this issue, we attempted to perform feature selection on the dataset. How- ever, this step did not improve the accuracy as anticipated, as it inadvertently removed

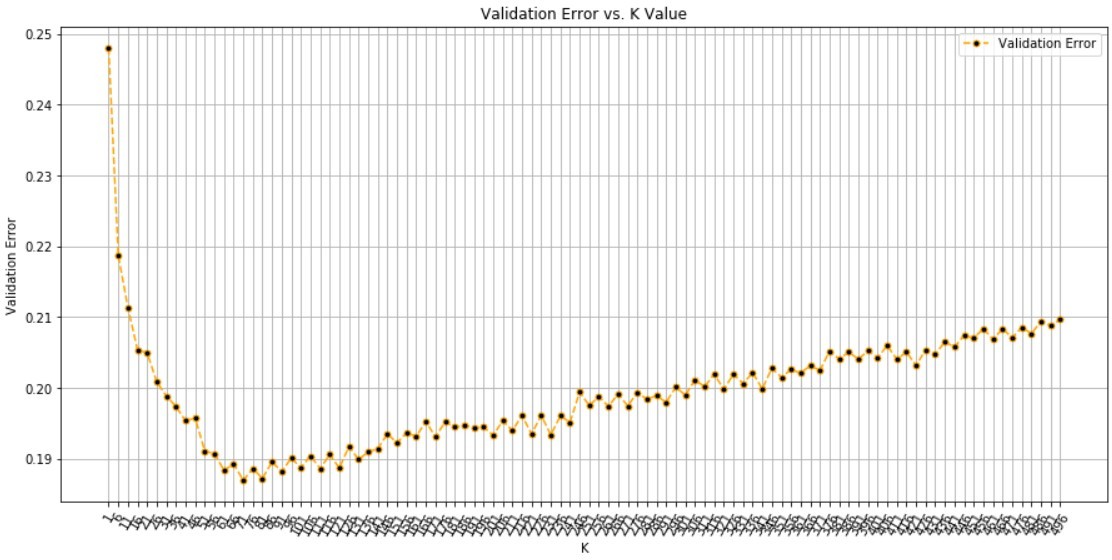


Figure 7: Validation error vs K value

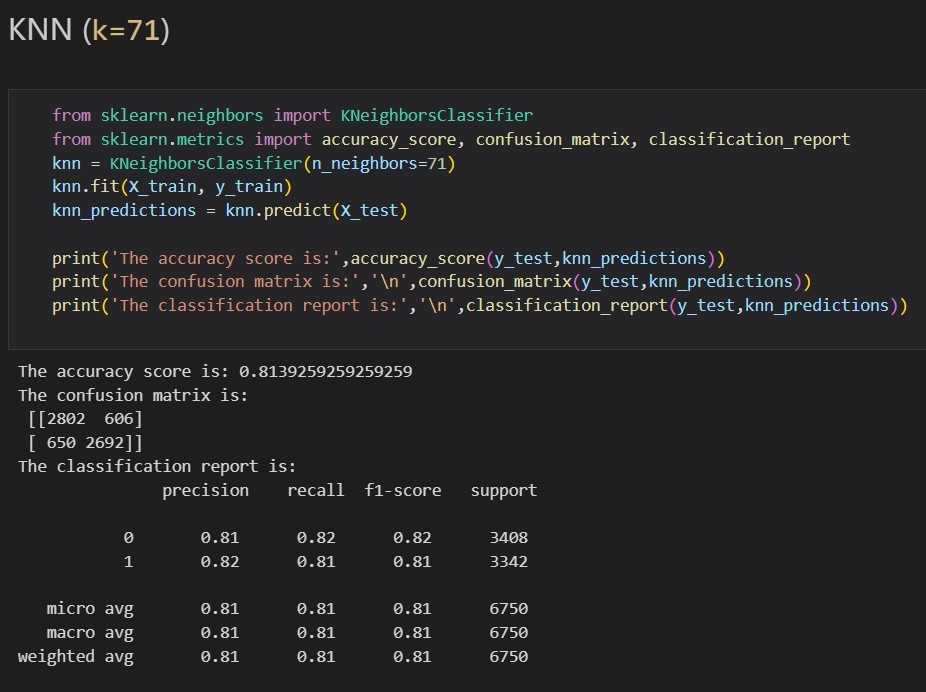


Figure 8: K Nearest Neighbours findings

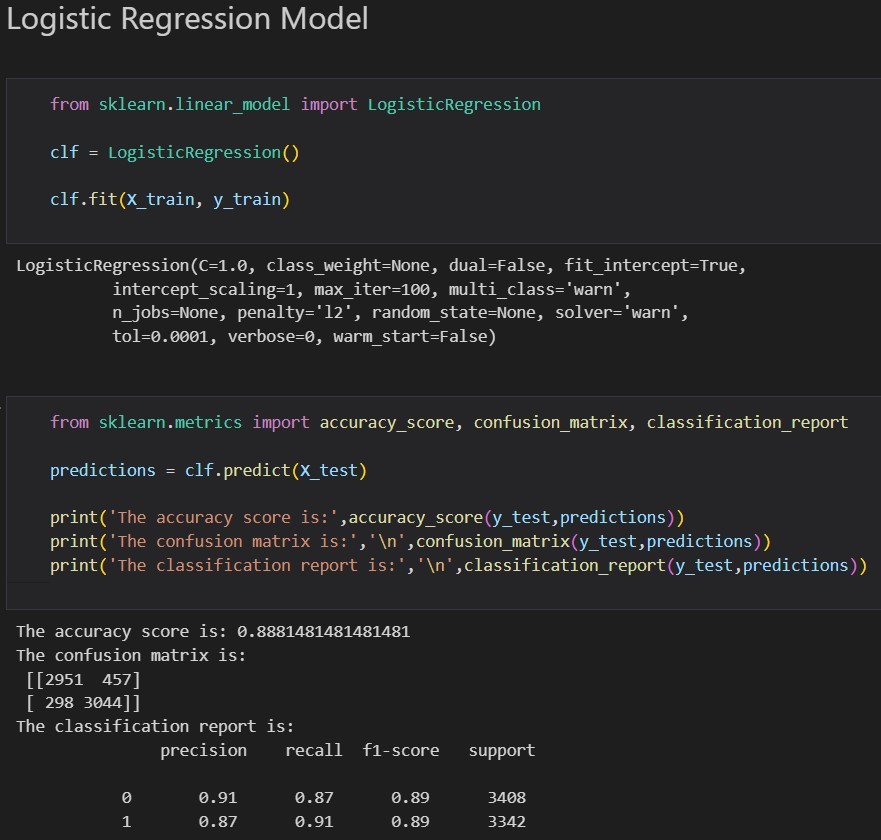


Figure 9: Logistic Regression Model findings

important words from the dataset. In order to refine our approach, we manually investi- gated the words that negatively affected the accuracy and added them to the list of stop words. This process allowed us to identify and remove words that contributed to lower

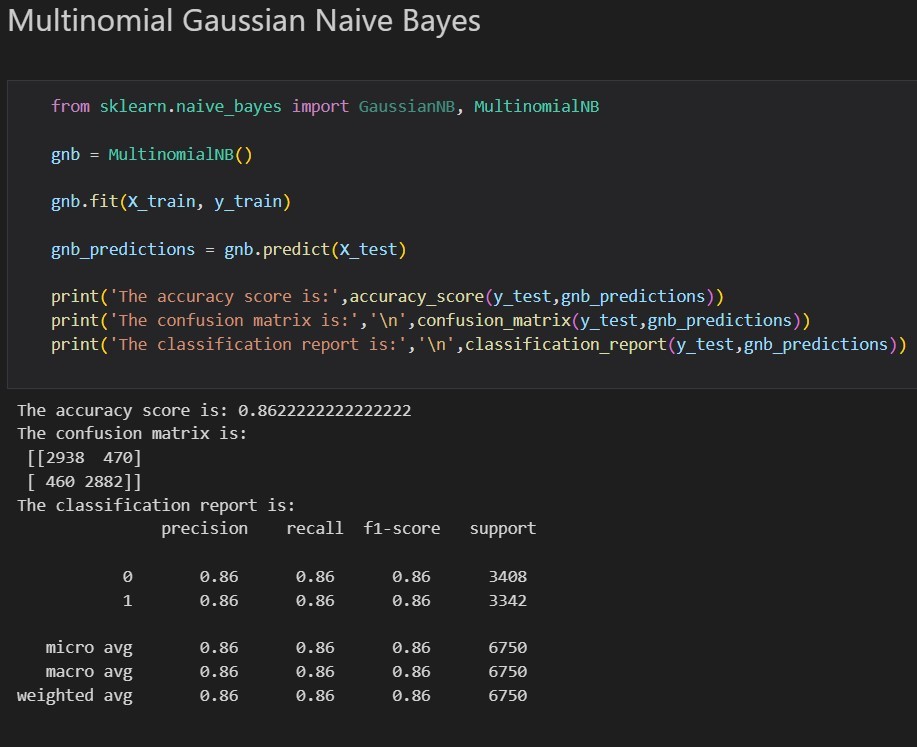


Figure 10: Multinomial Gaussian Naive Bayes findings accuracy rates, thus optimizing the input data for our models.

Upon improving the input data, we observed a noticeable enhancement in the perfor-

mance of the Logistic Regression model. This model outperformed the other algorithms, demonstrating its suitability for sentiment analysis in the context of IMDb movie re- views.The same can be inferred from the values of the F1 score that were observed and are presented in the figures attached. The combination of careful data preprocessing and the use of the Logistic Regression model proved to be a successful approach for our sen- timent analysis task, yielding the best results among the models tested with an accuracy of 0.888 with feature selection.

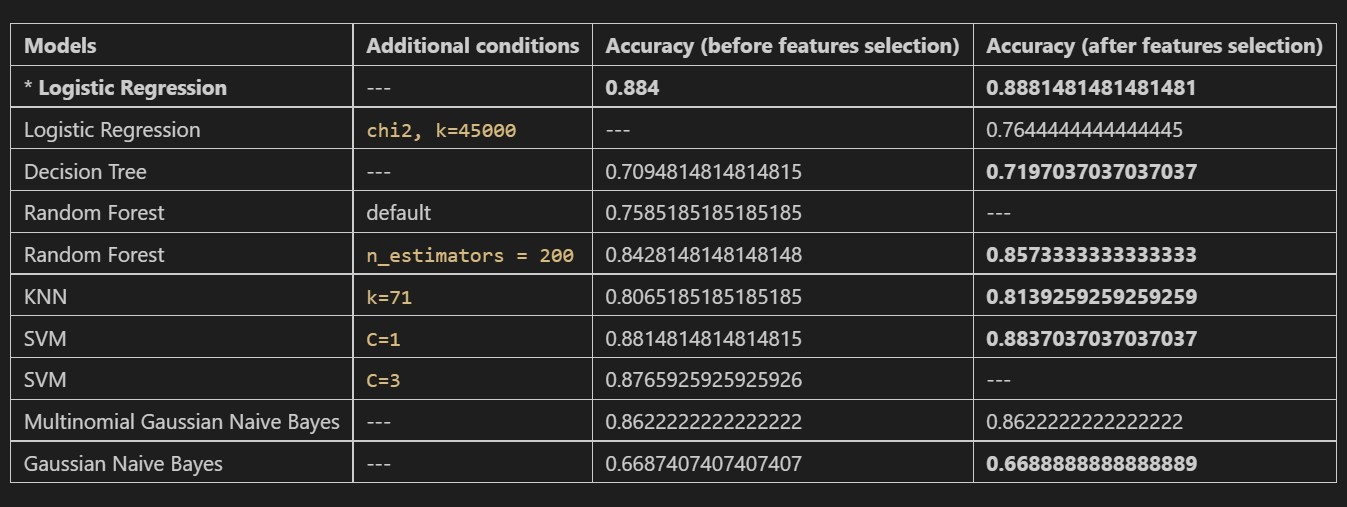


Figure 11: Findings Summary

# Conclusions and Limitations

In this study, we explored the application of various machine learning models, including Logistic Regression, Decision Tree, Random Forest, KNN, SVM with different C values, Multinomial Gaussian Naive Bayes, and Gaussian Naive Bayes for sentiment analysis of movie reviews from the IMDb dataset. We developed a comprehensive pre-processing

pipeline and applied feature classification techniques to enhance the extracted data and facilitate keyword extraction from user reviews.

Based on our experimental results, **Logistic Regression** emerged as the best-performing model compared to the other methods. We further improved the accuracy of our base- line models by refining the pre-processing pipeline, including adding more stop words to better capture relevant features.

While our research has achieved promising results in the domain of sentiment analysis for movie reviews, there are limitations and potential areas for future work:

* The scope of this study is limited to the IMDb dataset, which may not be represen- tative of all movie review sources. Expanding the research to include other sources of movie reviews, such as Rotten Tomatoes or Metacritic, would provide a more comprehensive understanding of the sentiment analysis performance across different platforms.
* Our methodology relies on a bag-of-words approach for feature representation, which can sometimes fail to capture the semantic meaning and context of the text. In future work, we could explore other feature representation techniques, such as word embeddings or sequence models like LSTM and Transformers, to better capture the semantic relationships between words.
* The current study focuses on binary sentiment classification (positive and negative). Future research could expand the scope to multi-class sentiment analysis, which would classify reviews into more fine-grained categories such as extremely positive, positive, neutral, negative, and extremely negative.
* The performance of the models can be further improved by incorporating additional features, such as review metadata (e.g., the reviewer’s rating, the number of likes on a review, or the reviewer’s historical rating patterns). This additional information could help models make better-informed decisions when classifying sentiments.

In conclusion, our study provides valuable insights into the performance of various ma- chine learning models for sentiment analysis of movie reviews. By understanding public opinions, emotions, and preferences, the film industry, critics, and viewers can make more informed decisions, ultimately benefiting the entertainment industry as a whole.

# References

1. Singh, Vikram Saini, Balwinder. (2014). An Effective Tokenization Algorithm for Information Retrieval Systems. Computer Science Information Technology. 4. 10.5121/csit.2014.4910.
2. S. Tripathi, R. Mehrotra, V. Bansal and S. Upadhyay, ”Analyzing Sentiment us- ing IMDb Dataset,” 2020 12th International Conference on Computational Intelli- gence and Communication Networks (CICN), Bhimtal, India, 2020, pp. 30-33, doi: 10.1109/CICN49253.2020.9242570.
3. S. Rathor and Y. Prakash, ”Application of Machine Learning for Sentiment Anal- ysis of Movies Using IMDB Rating,” 2022 IEEE 11th International Conference on

Communication Systems and Network Technologies (CSNT), Indore, India, 2022, pp. 196-199, doi: 10.1109/CSNT54456.2022.9787663.

1. Ng, Andrew, and Michael Jordan. 2001. “On Discriminative Vs. Generative Classi- fiers: A Comparison of Logistic Regression and Naive Bayes.” In Advances in Neural Information Processing Systems, edited by T. Dietterich, S. Becker, and Z. Ghahra- mani. Vol. 14. MIT Press.
2. Daeli, N. O. F., Adiwijaya, A. (2020, May 11). Sentiment Analysis on Movie Reviews using Information Gain and K-Nearest Neighbor. Journal of Data Science and Its Applications, 3(1), 1-7.
3. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In Proceed- ings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
4. [Kaggle Dataset link](https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?resource=download)