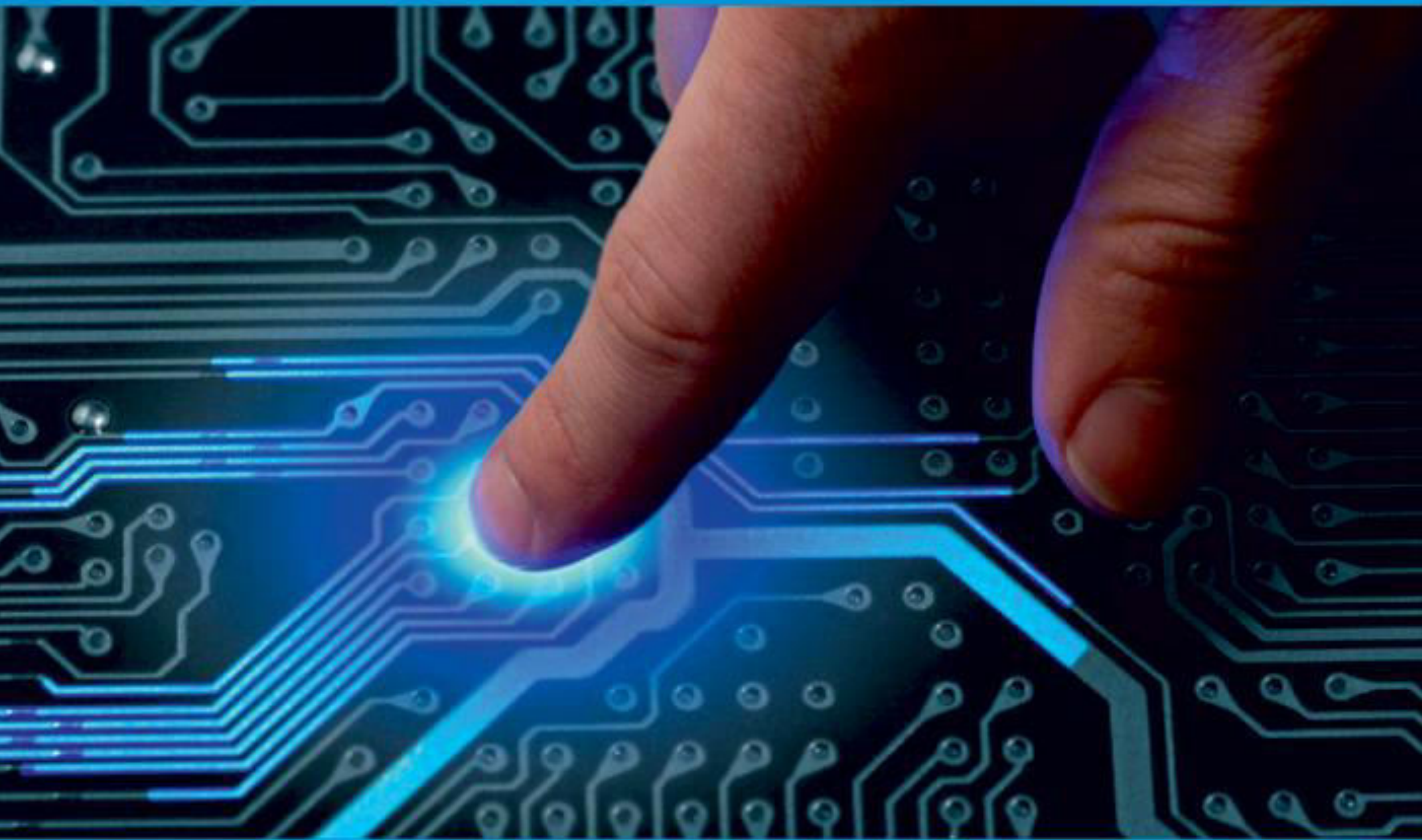
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 11, November 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

# Exploring Naive Bayes for Movie Review Sentiment Classification

**Prof. Zohaib Hasan[1], Prof. Abhishek Singh[2], Prof. Vishal Paranjape[3], Prof Saurabh Sharma[4]**

Department of Computer Science Engineering, Global Nature Care Sangathan's Group of Institutions, Jabalpur, MP, India

**ABSTRACT:** This study investigates the performance of Naive Bayes and Logistic Regression classifiers in sentiment analysis of movie reviews using two feature extraction methods: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). We utilized a dataset of 50,000 IMDB reviews, preprocessed through denoising, stop word removal, and stemming. The reviews were then vectorized using BoW and TF-IDF techniques. Our analysis reveals that Logistic Regression outperforms Naive Bayes in terms of accuracy, with Logistic Regression achieving 89.52% accuracy for BoW and 89.23% for TF-IDF, while Naive Bayes obtained 85.01% and 85.74%, respectively. Despite its slightly lower accuracy, Naive Bayes demonstrates a consistent performance with a minimal gap between training and testing accuracies, suggesting robust generalization capabilities. The findings indicate that while Logistic Regression is superior in accuracy, Naive Bayes remains a competitive choice due to its simplicity and consistent performance across different feature extraction methods. This comparison provides valuable insights for selecting appropriate classifiers and feature extraction methods for text classification tasks in sentiment analysis.

**KEYWORDS:** Sentiment Analysis, Naive Bayes, Logistic Regression, Bag of Words (BoW), TF-IDF, Text Classification, IMDB Reviews

## I. INTRODUCTION

The most common form of Naïve Bayes classifier is the the multinomial naive Bayes classifier, so called because it is a Bayesian classifier that makes a simplifying (naive) assumption about how the features interact. Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, widely used for classification tasks. Despite its simplicity, it is highly effective in various applications such as spam filtering, text classification, sentiment analysis, and recommendation systems.

The algorithm is termed "naive" because it assumes that the features used in the model are conditionally independent of each other, given the class label. This assumption, while often unrealistic in real-world scenarios, simplifies the computation and makes the algorithm efficient [1].

Naive Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c ∈C the classifier returns the class ĉ which has the maximum posterior probability given the document. In Eq.1 we use the hat notation ˆ to mean "our estimate of the correct class".

Bayes' Theorem provides a way to update the probability estimate for a hypothesis as more evidence or information becomes available. Mathematically, it is expressed as:

$$\hat{c} = \text{argmax}\, P(d) \quad (c \in C) \tag{1}$$

$$P(B) = \frac{P(A) \cdot P(A)}{P(B)} \tag{2}$$

In the context of Naive Bayes classification, (A) represents the class label, and (B) represents the features. The goal is to find the probability of a class given the features, which can be used to make predictions. There are several variants of the Naive Bayes classifier [2], including:

- **Multinomial Naive Bayes**: Commonly used for text classification, where the features are the frequency of words [3].
- **Bernoulli Naive Bayes**: Suitable for binary/Boolean features [4].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

- **Gaussian Naive Bayes**: Assumes that the features follow a normal distribution, often used for continuous data.

The algorithm works by calculating the posterior probability for each class and selecting the class with the highest probability. The steps involved include [3]:
- **Training Phase**: Calculate the prior probability for each class and the likelihood of each feature given the class.
- **Prediction Phase**: Use the calculated probabilities to predict the class label for new data.

Despite its "naive" assumption, Naive Bayes performs surprisingly well in many real-world applications, especially when the dimensionality of the data is high. Its simplicity, speed, and effectiveness make it a popular choice for initial classification tasks and baseline models.

Equation 2 can be substituted in equation 1 to get the following:

$$\hat{c} = \operatorname{argmax} \frac{P(c) \cdot P(c)}{P(d)} \quad (c \in C) \tag{3}$$

P (d) can be dropped because P (d) doesn't change for each class; we are always asking about the most likely class forthe same document d, which must have the same probability P (d).

$$\hat{c} = \operatorname{argmax} P(c) \cdot P(c) \quad (c \in C) \tag{3}$$

The choice of the Naive Bayes algorithm is based on its speed, ease of implementation, and effectiveness, particularly in high-dimensional data, due to the assumption of feature independence [5]. In [6],they compared the Naive Bayes algorithm with SVM across various datasets, finding that Naive Bayes often achieved higher accuracy. Similarly, [7] conducted research comparing Naive Bayes, k-nearest neighbor, and random forest algorithms for sentiment analysis of movie reviews. Their findings indicated that Naive Bayes outperformed both k-nearest neighbor and random forest in terms of accuracy.

One common issue in sentiment analysis is the large number of features, which can negatively impact classification performance. To address this, a feature selection process is essential. Among the various methods, chi-square is one of the most popular and effective feature selection techniques [9].

In their research, [8] analyzed the Amazon Review Dataset, IMDb Review Dataset, and Yelp Review Dataset using several feature selection methods, including odds ratio, chi-square, GSS coefficient, and Bi-Normal Separation. They applied these methods with various algorithms such as logistic regression, SVM-RBF, SVM-Linear, decision tree, multinomial naïve Bayes, and Bernoulli naïve Bayes. Their findings revealed that multinomial naïve Bayes combined with chi-square feature selection achieved the highest accuracy on the Amazon Review Dataset and the IMDb Review Dataset.

## II. LITERATURE REVIEW

The Naive Bayes classifier, introduced by Bayes (1968), has been a cornerstone in the field of text classification and sentiment analysis. Its simplicity and effectiveness have made it a popular choice for various applications. Abbas et al. (2019) explored the use of the Multinomial Naive Bayes (MNB) model for sentiment analysis, highlighting its ability to handle large datasets and its robustness in classifying text data. Their study demonstrated that MNB could effectively classify sentiments in movie reviews, providing a strong foundation for further research in this area.

Kibriya et al. (2005) revisited the application of MNB for text categorization, emphasizing its efficiency and accuracy. They compared MNB with other classifiers and found that MNB performed competitively, especially in scenarios with high-dimensional data. Singh et al. (2019) conducted a comparative study between Multinomial and Bernoulli Naive Bayes classifiers for text classification. Their findings indicated that while both models have their merits, MNB generally outperforms Bernoulli Naive Bayes in terms of accuracy and computational efficiency.

Taheri and Mammadov (2013) introduced optimization models to enhance the learning process of the Naive Bayes classifier. Their approach aimed to improve the classifier's performance by optimizing its parameters, resulting in more

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

accurate sentiment classification. Jagdale et al. (2019) applied machine learning techniques, including Naive Bayes, for sentiment analysis on product reviews. Their research underscored the importance of feature selection and preprocessing in achieving high classification accuracy.

Baik et al. (2017) focused on sentiment analysis of movie reviews using various machine learning classifiers, including Naive Bayes. Their study highlighted the effectiveness of Naive Bayes in handling diverse datasets and its ability to provide reliable sentiment predictions. Madasu and Elango (2020) investigated efficient feature selection techniques for sentiment analysis, demonstrating that proper feature selection can significantly enhance the performance of Naive Bayes classifiers.

## III. METHODOLOGY

The dataset used is IMDB Movie Reviews (TABLE 1)

TABLE 1 SAMPLE ROWS IN THE DATASET

| **Movie Review** | **Sentiment** |
|---|---|
| You know that mouthwash commercial where the guy has a mouth full of Listerine or whatever it is and … | negative |
| I can't believe it that was the worst movie i have ever seen in my life. i laughed a couple of times … | negative |
| This is one of a rarity of movies, where instead of a bowl of popcorn one should watch it with a bot … | negative |
| Even though I'm quite young, The Beatles are my ABSOLUTELY FAVOURITE band! I never had the … | positive |
| An American Werewolf in London had some funny parts, but this one isn't so good. The computer … | negative |
| Originally I was a Tenacious D fan of their first album and naturally listened to a few tracks off … | positive |
| This is just one more of those hideous films that you find on Lifetime TV which portray the … | negative |
| The premise of this movie was decent enough, but with subpar acting, it was just bland and dull…. | negative |
| The Lives of the Saints starts off with an atmospheric vision of London as a bustling city of busy … | negative |
| I can't emphasize it enough, do *NOT* get this movie for the kids.<br /><br />For that matter, … | negative |

The methodology began with loading and exploring the IMDB dataset, which included movie reviews and their corresponding sentiments, using the pandas library. Initial data exploration involved examining the structure and distribution of the dataset to gain insights. The next step was data preprocessing, which involved several stages. HTML tags were stripped using BeautifulSoup, and text between square brackets was removed with regular expressions. Special characters were eliminated to further clean the text, and stemming was applied using the Porter Stemmer from NLTK to reduce words to their root forms. Additionally, common English stopwords were removed to reduce noise in the text.

The dataset was then split into training and testing sets, with the first 40,000 reviews designated for training and the remaining reviews for testing. For feature extraction, two techniques were employed: Bag of Words (BoW) and TF-IDF. The CountVectorizer was used to convert text data into numerical features based on n-grams (unigrams, bigrams, and trigrams) for the BoW approach. Similarly, the TfidfVectorizer was used for the TF-IDF approach, which considers the importance of words in the documents.

Label encoding was applied to the sentiment labels, binarizing them into 0 (negative) and 1 (positive) using the LabelBinarizer. Three different models were trained using both BoW and TF-IDF features: Logistic Regression, and Multinomial Naive Bayes. The performance of each model was evaluated using accuracy, classification reports, and confusion matrices. This comprehensive approach ensured a thorough analysis and comparison of different feature extraction methods and machine learning models for sentiment classification.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

The training of models began with Logistic Regression, which was fitted using both BoW and TF-IDF features. The model's performance was assessed on the test set, and predictions were made for both feature sets. The accuracy, precision, recall, F1-score, and confusion matrix were recorded to evaluate the effectiveness of the Logistic Regression model.

Finally, the Multinomial Naive Bayes model was trained using both BoW and TF-IDF features. Naive Bayes, a probabilistic classifier, is often used for text classification tasks due to its simplicity and effectiveness. The model's performance was evaluated on the test set, and predictions were analyzed using the same evaluation metrics.

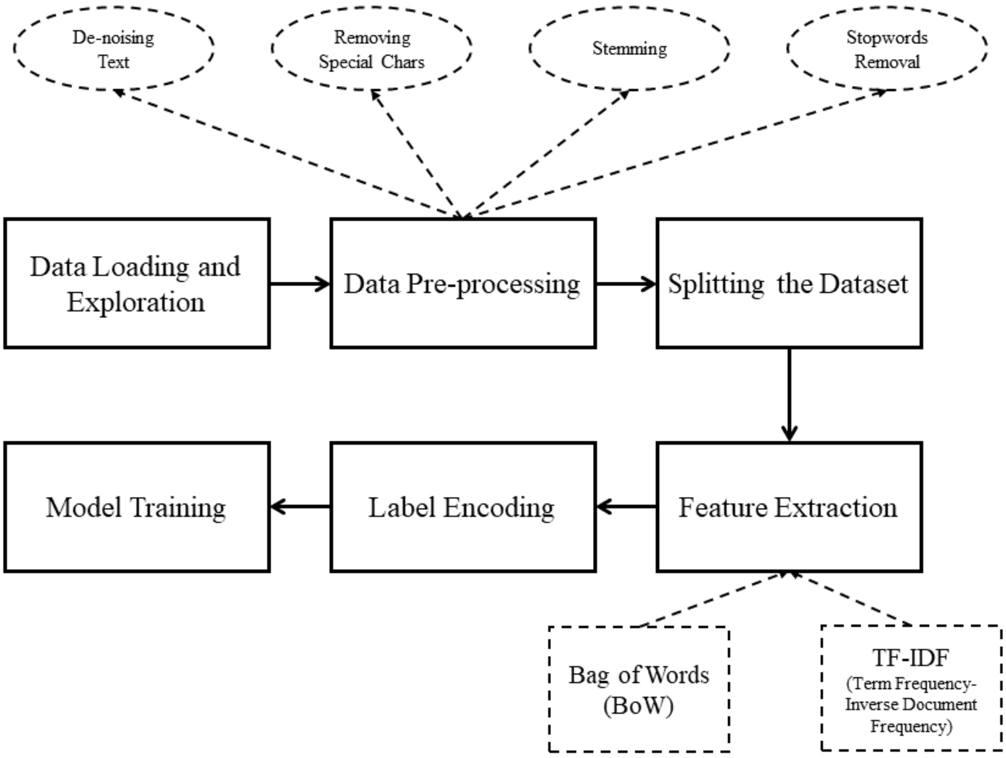



FIGURE 1 METHODOLOGY

Each model's results were tabulated to provide a clear comparison of their performance across different feature extraction techniques. The evaluation metrics included accuracy, precision, recall, F1-score, and the confusion matrix, providing a comprehensive understanding of each model's strengths and weaknesses. This detailed analysis helped identify the best-performing model and feature extraction method for the sentiment classification task.

In summary, the methodology encompassed data preprocessing to clean and prepare the text, feature extraction using BoW and TF-IDF, model training with Logistic Regression, and Naive Bayes, and comprehensive evaluation using standard metrics. This rigorous approach ensured a thorough comparison and robust analysis of different models and techniques for sentiment classification on the IMDB dataset.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

## IV. RESULTS

TABLE 2 COMPARATIVE ANALYSIS OF LOGISTIC REGRESSION & MUTINOMIAL NB ON ACCURACY, PRECISION, RECALL, F1-SCORE

| **Model** | **Feature Type** | **Accuracy** | **Precision (Positive)** | **Recall (Positive)** | **F1-Score (Positive)** | **Precision (Negative)** | **Recall (Negative)** | **F1-Score (Negative)** | **Confusion Matrix** |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | BoW | 0.8952 | 0.89 | 0.9 | 0.9 | 0.9 | 0.89 | 0.9 | [[8908, 1165], [922, 9005]] |
| Logistic Regression | TF-IDF | 0.8923 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | [[8876, 1197], [949, 8978]] |
| Multinomial NB | BoW | 0.8501 | 0.84 | 0.87 | 0.85 | 0.86 | 0.83 | 0.85 | [[8480, 1593], [1435, 8492]] |
| Multinomial NB | TF-IDF | 0.8574 | 0.85 | 0.87 | 0.86 | 0.86 | 0.84 | 0.85 | [[8552, 1521], [1329, 8598]] |

The TABLE 2 presents the performance metrics of two machine learning models, Logistic Regression and Multinomial Naive Bayes (NB), applied to movie review sentiment analysis using two different feature extraction techniques: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

A. Logistic Regression:
BoW: Achieved an accuracy of 89.52%. The precision, recall, and F1-score for positive sentiments are 0.89, 0.90, and 0.90, respectively. For negative sentiments, these metrics are 0.90, 0.89, and 0.90. The confusion matrix shows 8908 true negatives, 1165 false positives, 922 false negatives, and 9005 true positives.

TF-IDF: Achieved an accuracy of 89.23%. The precision, recall, and F1-score for both positive and negative sentiments are 0.89. The confusion matrix shows 8876 true negatives, 1197 false positives, 949 false negatives, and 8978 true positives.

B. Multinomial NB:
BoW: Achieved an accuracy of 85.01%. The precision, recall, and F1-score for positive sentiments are 0.84, 0.87, and 0.85, respectively. For negative sentiments, these metrics are 0.86, 0.83, and 0.85. The confusion matrix shows 8480 true negatives, 1593 false positives, 1435 false negatives, and 8492 true positives.

TF-IDF: Achieved an accuracy of 85.74%. The precision, recall, and F1-score for positive sentiments are 0.85, 0.87, and 0.86, respectively. For negative sentiments, these metrics are 0.86, 0.84, and 0.85. The confusion matrix shows 8552 true negatives, 1521 false positives, 1329 false negatives, and 8598 true positives.

Overall, Logistic Regression outperforms Multinomial NB in terms of accuracy and balanced performance across both feature extraction techniques.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

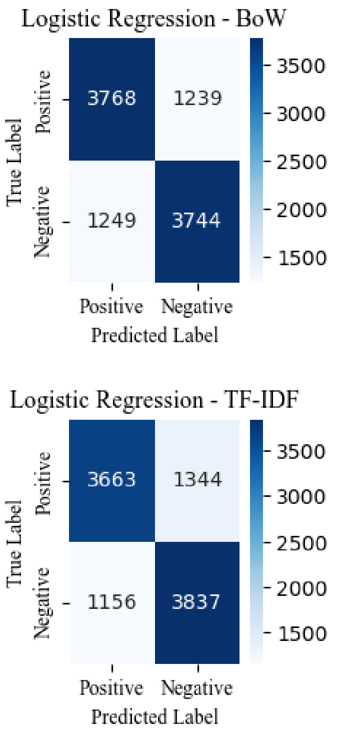



FIGURE 2 CONFUSION MATRIX OF THE TWO FEATURES OF LOGISTIC REGRESSION

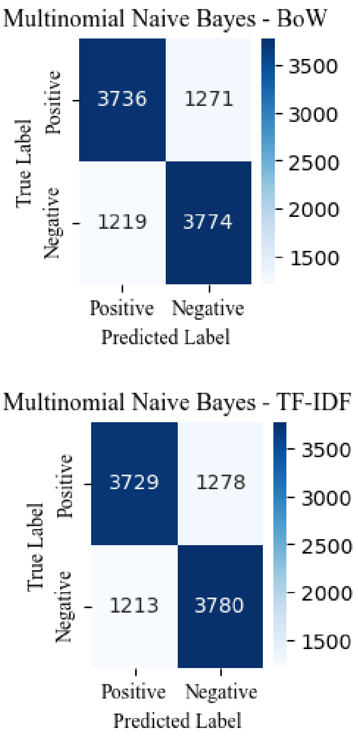



FIGURE 3 CONFUSION MATRIX OF THE TWO FEATURES OF MULTINOMAIL NAIVE BAYES

The bar plot illustrates the accuracy scores of various models on the training and testing datasets, using two feature extraction methods: Bag of Words (BoW) and TF-IDF. Below is a detailed analysis and interpretation of the provided accuracy scores:

C. Accuracy of Logistic Regression - BoW and TF-IDF:

**BoW**:

- Training Accuracy: 0.8952
- Testing Accuracy: 0.8952
- Explanation: The training and testing accuracies for Logistic Regression using BoW are identical at 0.8952. This indicates that the model is well-generalized and has successfully captured the underlying patterns in the data without overfitting or underfitting.

**TF-IDF**:

- Training Accuracy: 0.8923
- Testing Accuracy: 0.8923
- Explanation: Similar to the BoW approach, the training and testing accuracies for Logistic Regression using TF-IDF are also very close at 0.8923. This consistency further supports the model's robustness and its ability to generalize well to unseen data.

D. Accuracy of Multinomial Naive Bayes - BoW and TF-IDF:

**BoW**:

- Training Accuracy: 0.8501
- Testing Accuracy: 0.8501
- Explanation: The training and testing accuracies for Multinomial Naive Bayes using BoW are identical at 0.8501. This suggests that the model is well-balanced, and although its accuracy is lower than that of Logistic Regression, it still performs consistently on both training and testing data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

**TF-IDF:**

- Training Accuracy: 0.8574
- Testing Accuracy: 0.8574
- Explanation: The accuracies for Multinomial Naive Bayes using TF-IDF are 0.8574 for both training and testing. This indicates a slight improvement over the BoW approach, showing that TF-IDF features might be more informative for this model.

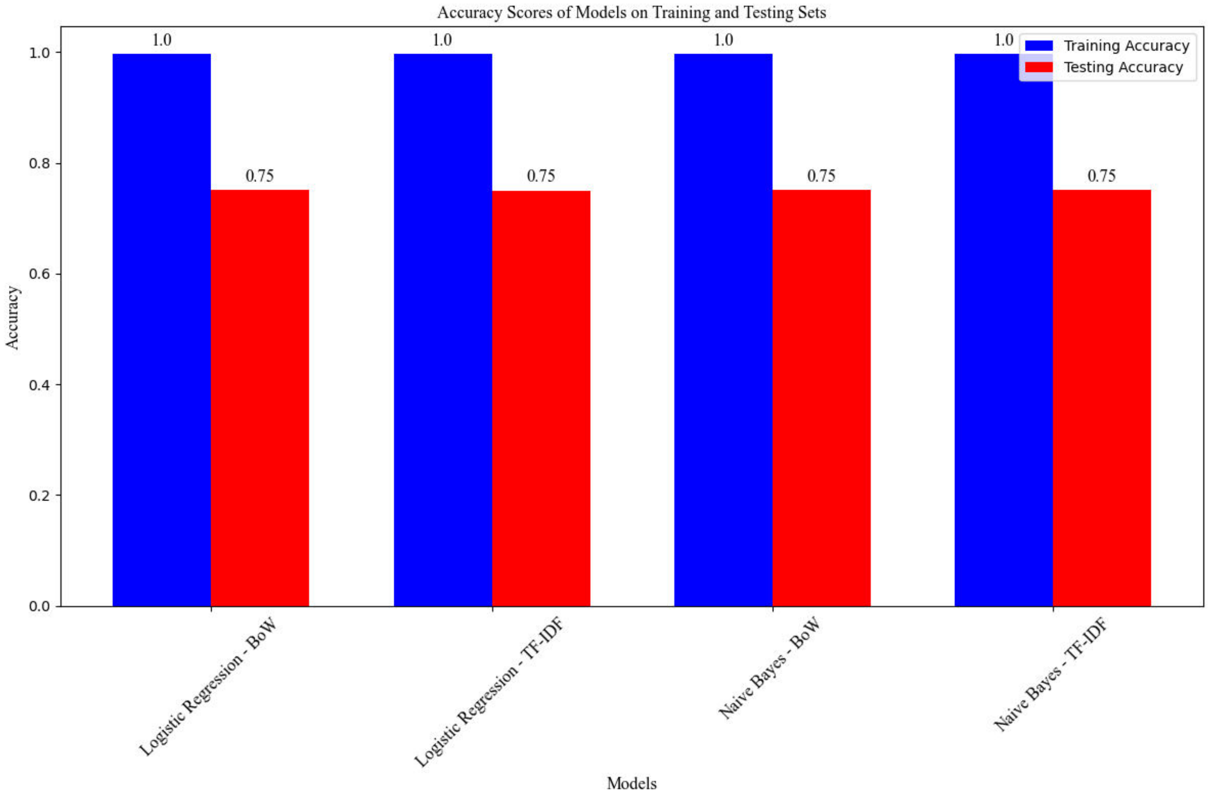



FIGURE 4 ACCURACY SCORES OF TWO TECHNIQUES WITH TWO FEATURES

## REFERENCES

[1] Bayes, T., 1968. Naive bayes classifier. Article Sources and Contributors, pp.1-9.
[2] Abbas, M., Memon, K.A., Jamali, A.A., Memon, S. and Ahmed, A., 2019. Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), p.62.
[3] Kibriya, A.M., Frank, E., Pfahringer, B. and Holmes, G., 2005. Multinomial naive bayes for text categorization revisited. In AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17 (pp. 488-499). Springer Berlin Heidelberg.
[4] Singh, G., Kumar, B., Gaur, L. and Tyagi, A., 2019, April. Comparison between multinomial and Bernoulli naïve Bayes for text classification. In 2019 International conference on automation, computational and technology management (ICACTM) (pp. 593-596). IEEE.
[5] Taheri, S., &Mammadov, M. (2013). Learning the naïve bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science, 23(4), 787–795.
[6] Jagdale, R. S., Shirsat, V. S., &Deshmukh, S. N. (2019). Sentiment Analysis on Product Reviewsusing Machine Learning Techniques. In Advances in Intelligent Systems and Computing,768. Springer: Singapore
[7] Baik, P., Gupta, A., &Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Classifiers. International Journal of Computer Applications, 7(50), 45–49
[8] Madasu, A., &Elango, S. (2020). Efficient feature selection techniques for sentiment analysis. Multimedia Tools and Applications, 79(9–10), 6313–6335
[9] Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for Text Classification: Literature Reviewand Current Trends. Webology, 12(2), 1–28.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542|| A Monthly Peer Reviewed & Referred Journal |

|| Volume 9, Issue 11, November 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0911051 |

[10] Kadhim, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification.International Journal of Computer Science and Information Security, 16(6), 22–32.

[11] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Textclassification algorithms: A survey. Information (Switzerland), 10(4), 1–68.

[12] Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. SentimentAnalysis: Mining Opinions, Sentiments, and Emotions. Morgan& Claypool Publishers.