Sentiment Analysis Report

Problem Statement:

The objective of this project is to build a sentiment analysis model that classifies text data into two categories: **positive** and **negative**. This classification is crucial for understanding customer opinions, enhancing decision-making, and improving product and service quality. For this purpose, the IMDB movie review dataset was used as a benchmark.

Approach:

1. **Data Collection and Preprocessing:**
   - The dataset consisted of 50,000 movie reviews, balanced across positive and negative sentiments.
   - Steps involved:
     - Lowercasing text to ensure uniformity.
     - Removing punctuation and special characters using regex.
     - Tokenization of words for structured processing.
     - Removal of stopwords to focus on meaningful words.
     - Lemmatization to reduce words to their base forms.
2. **Feature Extraction:**
   - Utilized **TF-IDF Vectorization** to convert textual data into numerical features while preserving important term frequencies.
3. **Model Selection:**
   - Selected **Multinomial Naive Bayes** as the classifier due to its effectiveness in handling text-based data and high dimensionality.
4. **Data Splitting:**
   - The dataset was split into **80% training** and **20% testing** sets to evaluate model performance.
5. **Model Training:**
   - The TF-IDF-transformed training data was fed into the Multinomial Naive Bayes classifier.
6. **Model Evaluation:**
   - The testing data was used to evaluate the model using accuracy, precision, recall, F1-score, and a confusion matrix.

Problems Faced:

1. **Imbalanced Predictions:**
   - Initially, the model showed a slight bias towards predicting the majority class. This was resolved by ensuring balanced training data and tuning hyperparameters.
2. **Text Preprocessing Challenges:**

        - Cleaning text data required significant effort to handle edge cases like special characters and combined words.
3. **Overfitting on Training Data:**
    - Regularization techniques and stratified splitting helped mitigate this issue.

Learned Outcomes:

1. **Effective Preprocessing:**
    - The importance of thorough text preprocessing for high-quality predictions.
2. **TF-IDF Insights:**
    - Learned how TF-IDF prioritizes relevant terms by reducing the impact of frequent but less informative words.
3. **Evaluation Metrics:**
    - Gained deeper insights into evaluating model performance using precision, recall, F1-score, and confusion matrices.

Model Accuracy:

- **Accuracy:** 86.64%

Detailed Metrics:

Classification Report:

```
              precision    recall  f1-score   support

    negative       0.85      0.88      0.87      4961
    positive       0.88      0.85      0.86      5039

    accuracy                           0.87     10000
   macro avg       0.87      0.87      0.87     10000
weighted avg       0.87      0.87      0.87     10000
```

Confusion Matrix:

```
[[4384  577]
 [ 759 4280]]
```

Key Observations:

1. **False Negatives and Positives:**

    - The confusion matrix revealed 759 false positives and 577 false negatives, suggesting room for improvement in capturing nuances in certain reviews.
2. **Balanced Performance:**
    - The precision, recall, and F1-scores are well-balanced between positive and negative classes, indicating good generalization of the model.

Conclusion:

The sentiment analysis model achieved a commendable accuracy of 86.64% with balanced precision and recall across both classes. The project demonstrated the significance of robust preprocessing and feature engineering for text-based machine learning models. Future improvements could involve using advanced models like **BERT** or **LSTMs** to capture deeper contextual relationships in the text.