IMDb Movie Review Sentiment Classification

Mini-Project II

Mahira Ibnath Joytu

1. Introduction:

Sentiment analysis is a critical task in natural language processing (NLP) that involves classifying text data based on sentiment polarity. In this project, we focus on the sentiment classification of IMDB movie reviews, categorizing them as positive or negative. Given the large volume of online movie reviews, automated sentiment analysis can help businesses, filmmakers, and audiences gain insights into public opinion efficiently.

The dataset consists of 50,000 IMDB movie reviews, evenly split into positive and negative sentiments. The primary objective is to develop and evaluate machine learning models for sentiment classification, ensuring optimal accuracy and generalization. We implement three models: Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression, and analyze their comparative performance.

This report details the data preprocessing steps, modeling approaches, performance evaluation, and key findings.

2. Methodology:

2.1. Data Collection:

The dataset was sourced from Hugging Face which comprises 50,000 IMDB movie reviews, pre-labeled as either positive (1) or negative (0). Initially, the dataset was structured such that all positive reviews appeared first, followed by negative reviews.

2.2 Data Preprocessing:

To ensure data quality and improve model performance, some preprocessing steps were applied:

2.2.1 Shuffling the dataset

The dataset was shuffled to ensure an even distribution of sentiment labels across training and test sets since the dataset was initially distributed in such a way that the positively labelled reviews appeared first and then the negative reviews followed. Without shuffling the train and test datasets would be heavily biased towards one label which is not ideal for training.

2.2.2 Text Cleaning

Lowercasing text: Converting all characters to lowercase ensures uniformity and prevents the model from treating words with different cases as distinct entities.

Removing punctuation and stopwords: Eliminating unnecessary characters and frequent words (such as "the," "and," "is") helps reduce noise and improve model efficiency by focusing on meaningful content.

2.2.3 TF-IDF Vectorization:

Text data was transformed into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency) with 5,000 features. This method helps to capture the importance of words while reducing noise from operational patterns, balancing short-term variability and data sufficiency

2.3 Model Training

Each model was trained on TF-IDF-transformed text features, and their performance was evaluated using key metrics. Three machine learning models were trained to classify the movie reviews.

2.3.1 Naïve Bayes

A probabilistic classifier based on Bayes' theorem, particularly effective for text classification. Multinomial Naïve Bayes (MNB) was used, as it works well with TF-IDF-transformed data.

2.3.2 Support Vector Machine (SVM)

A powerful classification model that finds the optimal hyperplane for separating classes. Linear SVM was used due to its efficiency in handling high-dimensional text data.

2.3.3 Logistic Regression

A widely used linear classifier for binary classification. The model was trained with L2 regularization to improve generalization.

3. Results

After training the Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression models, their performance was evaluated using few key metrics as these metrics provide insights into how well each model correctly classifies the review based on sentiment.

3.1 Accuracy

The primary metric used is accuracy, which indicates the proportion of correctly classified reviews. It measures overall correctness but does not differentiate between false positives and false negatives.

Model	Accuracy
Naïve Bayes	85.46%
SVM	88.32%
Logistic Regression	88.40%

3.2 Precision, Recall, and F1-Score

Precision: The proportion of true positive predictions out of all positive predictions. Higher precision means fewer false positives.

Recall: The proportion of actual positives that were correctly identified. A higher recall indicates fewer false negatives.

F1-Score: The harmonic mean of precision and recall, providing a balanced measure when both are equally important. The following table summarizes the evaluation results for the three models:

Model	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	0.85	0.85	0.85
SVM	0.89 (-ve) 0.87 (+ve)	0.87 (-ve) 0.89(+ve)	0.88
Logistic Regression	0.90 (-ve) 0.87 (+ve)	0.87 (-ve) 0.90 (+ve)	0.88

3.3 Confusion Matrix

The following confusion matrices were generated for the three models:

3.3.1 Naïve Bayes

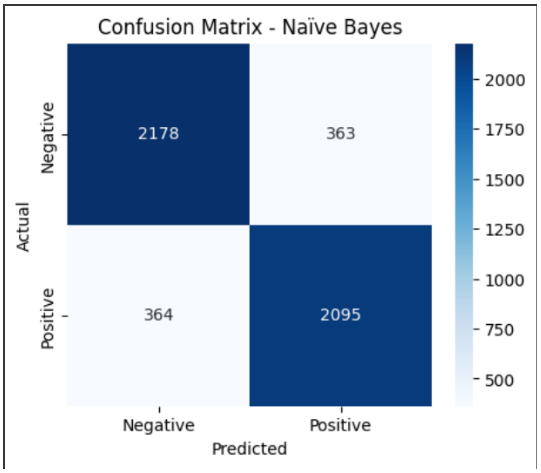

Figure 1: Naïve Bayes confusion matrix

The confusion matrix for Naïve Bayes shows:

- True Negatives: 2178 negative reviews were correctly classified
- True Positives: 2095 positive reviews were correctly classified
- False Positives: 363 negative reviews were misclassified as positive
- False Negatives: 364 positive reviews were misclassified as negative

The balanced misclassification between false positives and negatives suggests that Naïve Bayes has a consistent classification pattern but lacks higher precision compared to other models. The errors could be attributed to the assumption of feature independence, which is a limitation of Naïve Bayes when applied to natural language data.

3.3.2 Support Vector Machine (SVM)

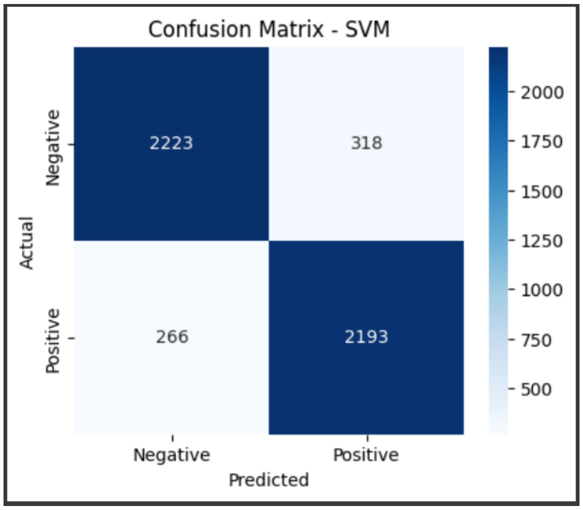

Figure 2: SVM confusion matrix

- True Negatives: 2223 negative reviews were correctly classified.
- False Positives: 318 negative reviews were misclassified as positive.
- False Negatives: 266 positive reviews were misclassified as negative.
- True Positives: 2193 positive reviews were correctly classified.

The SVM model demonstrates strong performance with a relatively low misclassification rate. The number of false positives and false negatives is lower than in Naïve Bayes, indicating better generalization and a more refined decision boundary for sentiment classification.

3.3.3 Logistic Regression

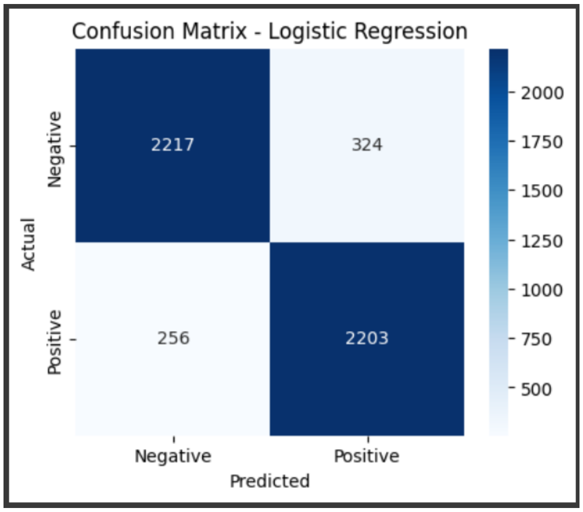

Figure 3: Logistic Regression confusion matrix

- True Negatives: 2217 negative reviews were correctly classified.
- False Positives: 324 negative reviews were misclassified as positive.
- False Negatives: 256 positive reviews were misclassified as negative.
- True Positives: 2203 positive reviews were correctly classified.

Logistic Regression demonstrates strong classification performance, with slightly fewer false negatives and false positives compared to SVM. The model maintains a balanced classification with a minimal misclassification rate, indicating that it effectively captures sentiment patterns in movie reviews. The slight edge over SVM suggests that Logistic Regression generalizes well while maintaining high accuracy.

3.4 Learning Curve of Models

A learning curve helps visualize how a model's performance evolves as more training data is provided. It demonstrates whether a model is underfitting, well-fitted, or overfitting to the training data.

3.4.1 Naïve Bayes

Unlike SVM and Logistic Regression, a traditional learning curve cannot be applied to Naïve Bayes in the same way since it does not learn iteratively from data. Instead, it computes probabilities based on word occurrences and assumes feature independence.

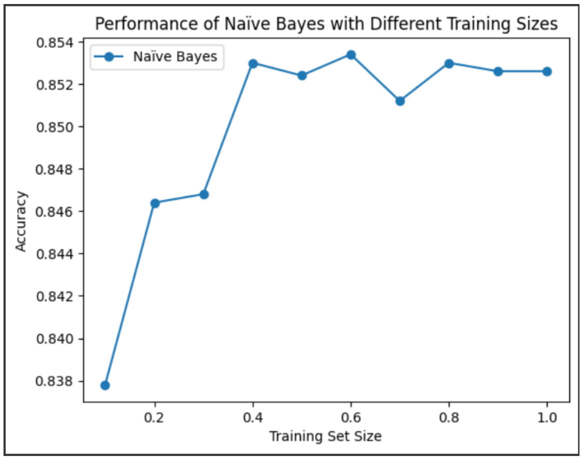

Figure 4: Naïve Bayes learning curve

The model shows rapid improvement with small training data and reaches stability around 85% accuracy. However, beyond this point, additional training samples do not contribute significantly to performance improvement, as the model relies on predefined probability distributions rather than iterative optimization.

3.4.2 Support Vector Machine (SVM)

SVM exhibits a steady improvement in accuracy as training data increases, unlike Naïve Bayes, which stabilizes early.

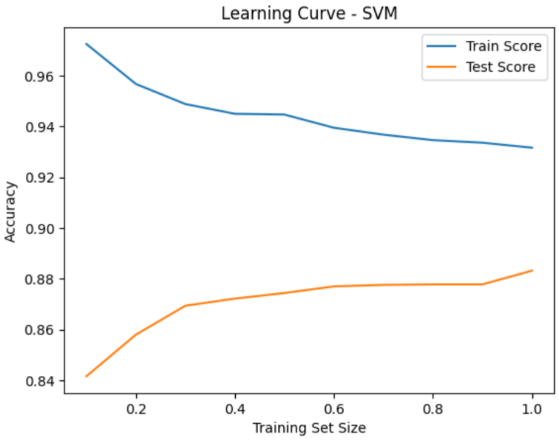

Figure 5: SVM learning curve

The model shows a slight gap between training and test accuracy, indicating mild overfitting but maintaining strong generalization ability. While it performs well with relatively less data, adding more training samples continues to enhance its performance, showing its scalability and robustness for sentiment classification.

3.4.3 Logistic Regression

Logistic Regression follows a similar pattern to SVM but achieves slightly better test accuracy, suggesting that it generalizes slightly more effectively. The model improves consistently with increasing data, demonstrating strong adaptability. The learning curve indicates that Logistic Regression benefits from larger datasets, making it a reliable choice for sentiment classification.

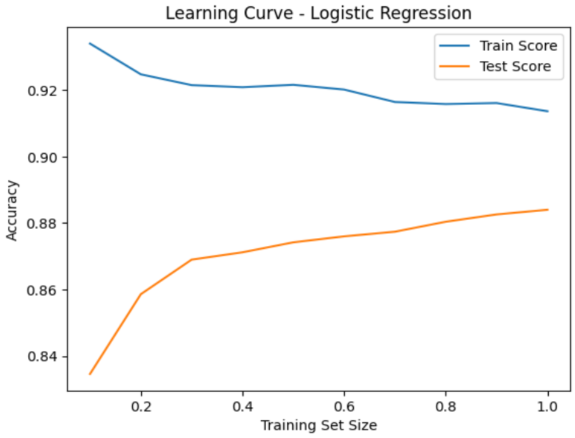

Figure 6: Logistic Regression learning curve

The learning curves for all three models reveal distinct behaviors. Naïve Bayes reaches peak performance quickly, showing little improvement with additional data due to its probabilistic nature. In contrast, both SVM and Logistic Regression show continuous improvements, with Logistic Regression demonstrating the best generalization. These trends suggest that while Naïve Bayes serves as a strong baseline, SVM and Logistic Regression are better suited for handling large-scale sentiment classification tasks due to their scalability and ability to refine decision boundaries with more data.

3.5 Analysis of Model Performance

The results indicate that Logistic Regression achieved the highest accuracy at 88.40%, followed closely by SVM at 88.32%, while Naïve Bayes lagged at 85.46%. The slight edge of Logistic Regression over SVM suggests that it was able to better model the decision boundary for sentiment classification while maintaining generalization.

In terms of precision, recall, and F1-score, SVM and Logistic Regression exhibited balanced scores across both positive and negative sentiment classes, reflecting their robustness in distinguishing between sentiments. Naïve Bayes, although performing decently, had slightly lower scores due to its assumption of feature independence, which does not hold well in natural language data. The learning curves further support these findings, where Naïve Bayes quickly reaches its peak performance, while SVM and Logistic Regression continue to improve with more data, demonstrating better scalability. These results suggest that while Naïve Bayes is a strong baseline model, SVM and Logistic Regression are better suited for sentiment classification tasks due to their superior generalization ability and adaptability to larger datasets.

3.6 Key Takeaways

The key takeaways from this project and evaluation of model performance are:

- Logistic Regression slightly outperformed SVM (88.40% vs. 8.32%), making it the best-performing model.
- Naïve Bayes performed well (85.46%), proving to be a strong baseline for text classification.
- SVM and Logistic Regression are better choices for generalization, as they demonstrated higher accuracy and balanced precision-recall values.
- TF-IDF worked effectively, but deep learning models could further improve accuracy.

5. Conclusion

This study evaluated three machine learning models—Naïve Bayes, SVM, and Logistic Regression—for sentiment classification on IMDB movie reviews. The results indicate that Logistic Regression performed best with an accuracy of 88.40%, closely followed by SVM at 88.32%, while Naïve Bayes achieved 85.46%. The learning curves and confusion matrices demonstrated that Naïve Bayes, despite being a strong baseline model, reached its peak performance quickly and lacked scalability. On the other hand, SVM and Logistic Regression showed continuous improvement with increasing training data, highlighting their ability to generalize well. While TF-IDF effectively transformed text data into numerical features, advanced NLP models such as Word2Vec or BERT could potentially improve classification performance further. The results underscore the importance of selecting models based on both accuracy and scalability, with SVM and Logistic Regression emerging as the most effective options for sentiment analysis in large datasets.