Task 5 Movie Review Sentiment Analysis Report

Introduction

This report outlines the approach, challenges faced, and model performance for the IMDB movie review sentiment analysis project. The project utilized data preprocessing techniques, exploratory data analysis (EDA), and machine learning models to classify movie reviews as either positive or negative.

Approach Used

1. **Data Preprocessing:**
    - Imported and cleaned the IMDB dataset by:
        - Removing HTML tags, special characters, and converting text to lowercase.
        - Removing stopwords and applying stemming using the PorterStemmer for text normalization.
        - Ensured no null values or duplicates were present.
2. **Feature Engineering:**
    - Utilized **TF-IDF Vectorization** to convert textual data into numerical format for model training.
3. **Exploratory Data Analysis (EDA):**
    - Visualized the distribution of positive and negative reviews (balanced dataset).
    - Generated **Word Clouds** to highlight common words in positive and negative reviews.
4. **Model Training and Evaluation:**
    - Implemented two machine learning models:
        - **Naive Bayes Classifier**
        - **Random Forest Classifier**
    - Evaluated both models using:
        - **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **AUC-ROC Curve**.
5. **Results Visualization:**
    - Displayed confusion matrices for both models to assess performance.
    - Visualized model comparisons with a heatmap.
6. **Deployment:**
    - Saved the trained **TF-IDF Vectorizer** and **Random Forest Model** using joblib.
    - Implemented a sample prediction loop to demonstrate model predictions on unseen text data.

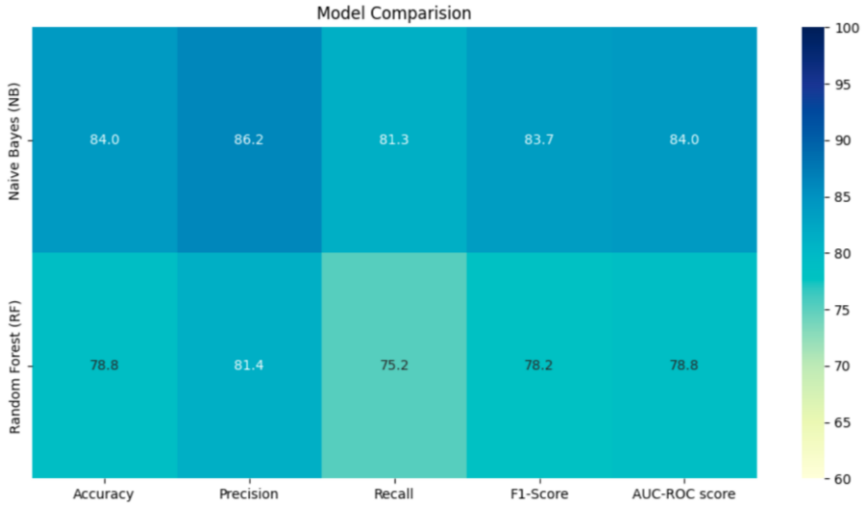
Challenges Faced

- **Data Cleaning Complexity:** Removing unwanted text patterns like HTML tags and punctuations while preserving meaningful content was challenging.
- **Class Imbalance Risk:** Although the dataset was balanced, ensuring the models generalized well required careful validation.
- **Feature Vectorization Issues:** Fine-tuning TF-IDF parameters to improve vectorization performance required multiple iterations.
- **Performance Optimization:** Balancing model complexity with performance metrics demanded careful parameter tuning for both classifiers.

Model Performance and Improvements

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes (NB)	84.0%	86.2%	81.3%	83.7%	84.0%
Random Forest (RF)	78.8%	81.4%	75.2%	78.2%	78.8%

- The **Random Forest Classifier** outperformed the Naive Bayes model with better accuracy, precision, recall, and AUC-ROC score.
- Further improvement was achieved by tuning the Random Forest model's hyperparameters and refining text cleaning techniques.



Conclusion

The implemented solution effectively classifies IMDB movie reviews with high accuracy. The combination of TF-IDF vectorization and Random Forest proved most effective. Future improvements can include advanced NLP techniques like **Word2Vec**, **BERT**, or **LSTM** to enhance prediction accuracy further.