Project Report: IMDB Movie Review Sentiment Analysis

Overview

This project focuses on sentiment analysis of movie reviews from the IMDB dataset, utilizing Python libraries such as Pandas, NumPy, NLTK, Seaborn, and SpaCy. The goal was to classify reviews as positive or negative sentiment using a Multinomial Naive Bayes model.

Methodology

Data Preprocessing

The initial phase involved data cleaning and preprocessing. Using Pandas, we loaded the dataset and handled duplicated reviews. NLTK and SpaCy were employed for tokenization, removing stop words, and lemmatization and stemming to ensure the text data was ready for analysis.

Text Analysis

Once the text was preprocessed, we performed exploratory data analysis (EDA) using Seaborn to visualize the distribution of sentiments. This helped identify dataset balancing.

Model Building

For the classification task, we chose the Multinomial Naive Bayes algorithm, which is well-suited for text classification. The dataset was split into training and testing sets, and the model was trained on the processed text features.

Results

The model achieved an accuracy of 0.8486, with precision and recall both around 0.85. The F1-score of 0.8489 indicates a good balance between precision and recall, confirming the model's effectiveness in sentiment classification.

Conclusion

This project successfully demonstrated the application of NLP techniques and machine learning for sentiment analysis. The results indicate that the Multinomial Naive Bayes model performs well in classifying movie reviews, providing valuable insights for further improvements and applications.