Project Documentation

1. Project Overview

Project Title:
"Predicting Movie Sentiments with IMDB Reviews Using Naive Bayes"

Objective:
The primary objective of this project is to predict the sentiment (positive or negative) of IMDB movie reviews. This will be achieved by analyzing the dataset through Exploratory Data Analysis (EDA) and applying machine learning algorithms, with a particular focus on the Naive Bayes classifier for sentiment prediction.

2. Dataset Description

Dataset:
The project utilizes the IMDB Movie Reviews Dataset, which consists of 50,000 reviews labeled as positive or negative. The dataset is split into training and test sets, with each containing 25,000 reviews.

Key Variables:

- **Review:** The text of the movie review.
- **Sentiment:** The label associated with the review, where 0 represents a negative sentiment and 1 represents a positive sentiment.

3. Methodology

3.1. Data Preprocessing

- **Data Cleaning:** work on converting, transform and split the data for better working.
- **Vectorization:** Converting text data into numerical data using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to represent the importance of words.

3.2. Exploratory Data Analysis (EDA)

- **Sentiment Distribution:** Analyzing the balance of positive and negative sentiments in the dataset.
- **Word Cloud Analysis:** Visualizing the most frequent words in positive and negative reviews.
- **Length Analysis:** Examining the relationship between review length and sentiment.
- **Bigram and Trigram Analysis:** Identifying common word pairs or triplets in the reviews that may indicate sentiment.

3.3. Model Selection and Training

- **Naive Bayes Classifier:**
  - Chosen for its efficiency and effectiveness in text classification tasks.
  - Applied the Multinomial Naive Bayes algorithm, which is well-suited for the discrete nature of text data.
- **Model Evaluation:**
  - **Accuracy:** Measuring the proportion of correctly predicted sentiments.
  - **Precision and Recall:** Evaluating the model's performance on positive and negative reviews.
  - **Confusion Matrix:** Providing a visual representation of the model's predictions compared to actual sentiments.

4. Results and Insights

- **Model Performance:** Achieved an accuracy of 0.91072% on the test dataset.
- **Top Features:** Identified the words and phrases most indicative of positive and negative sentiments.
- **Key Findings:**
  - Reviews with specific adjectives or exclamations are more likely to be classified as positive.
  - Negative reviews often contain words related to disappointment, plot issues, or poor acting.

5. Conclusion

This project successfully demonstrates the power of machine learning, particularly the Naive Bayes algorithm, in predicting movie review sentiments. Through thorough EDA, model selection, and optimization, this project provides accurate and actionable predictions, offering valuable insights into the sentiment trends of moviegoers.

6. References

- **IMDB Dataset:** https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data