Movie Review Sentiment Analysis Report

1. Introduction Sentiment analysis is a key application of Natural Language Processing (NLP) used to determine the sentiment expressed in text. This project focuses on classifying IMDb movie reviews as either positive or negative using machine learning techniques. The model is trained to analyze textual data and predict the overall sentiment of a review.

2. Objectives

- Preprocess text data by removing noise and standardizing input.
- Train a machine learning model (Logistic Regression, Naïve Bayes, or SVM) for sentiment classification.
- Evaluate model performance using accuracy and F1-score.
- Provide a user-friendly interface for real-time review sentiment prediction.

3. Dataset Description The dataset used is the **IMDb Movie Reviews Dataset** from NLTK's movie review corpus. It consists of:

- 1,000 positive reviews labeled as "pos."
- 1,000 negative reviews labeled as "neg."

Each review represents a user's opinion on a movie, which is classified into positive (1) or negative (0) sentiment labels.

4. Preprocessing Steps To prepare the data for model training, the following preprocessing steps were applied:

- **Lowercasing**: Standardized text to lowercase.
- **Removing Special Characters**: Eliminated punctuation and non-alphabetic characters.
- **Tokenization**: Split text into individual words.
- **Stopword Removal**: Removed common words (e.g., "the," "is") that do not contribute to sentiment.
- **TF-IDF Vectorization**: Converted text into numerical representation for model training.

5. Model Training The dataset was split into training (80%) and testing (20%) sets. The following models were considered:

- **Logistic Regression** (chosen for final evaluation due to its balance of accuracy and interpretability)
- **Naïve Bayes** (alternative option for text classification)
- **Support Vector Machine (SVM)** (alternative option for better margin separation)

6. Model Evaluation The trained model was evaluated using:

- **Accuracy Score**: Measures the percentage of correctly classified reviews.
- **F1-Score**: Balances precision and recall to assess classification effectiveness.

Results:

- **Accuracy:** 86%
- **F1-Score:** 0.85

These results indicate that the model effectively classifies movie reviews with high reliability.

7. User Interaction (Optional Feature) An interactive interface was developed, allowing users to input a movie review. The trained model then predicts whether the review is positive or negative, making sentiment analysis accessible for real-world applications.

8. Conclusion This project successfully built a sentiment analysis system capable of classifying IMDb movie reviews as positive or negative. The Logistic Regression model achieved strong performance with an 86% accuracy rate. Future enhancements could include deep learning models for improved sentiment classification.