# Sentiment Analysis Report

## Introduction

This report summarizes the approach used for sentiment analysis of IMDb movie reviews, including preprocessing, model training, evaluation, challenges faced, and potential improvements.

## Approach

### 1. Data Preprocessing

- Loaded IMDb dataset containing labeled movie reviews.

- Converted text to lowercase.

- Removed punctuation and numbers.

- Tokenized text and removed stopwords using NLTK.

- Applied TF-IDF vectorization to convert text into numerical representation.

### 2. Model Training & Evaluation

- Experimented with three models: Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM).

- Split data into 80% training and 20% testing.

- Evaluated models using accuracy and F1-score.

- Logistic Regression performed best and was selected for final implementation.

## Challenges Faced

1. Handling Stopwords: Some stopwords contributed to sentiment; removing them affected accuracy.

2. Imbalanced Data: Certain words were more frequent in positive reviews, causing a slight bias.

3. Feature Engineering: Selecting the optimal number of TF-IDF features required experimentation.

4. Computational Complexity: SVM took longer to train compared to other models.

## Model Performance & Improvements

- Final Model: Logistic Regression

- Accuracy: 87%

- F1-score: 86%

- Confusion Matrix: Showed some false positives/negatives but overall balanced.

### Possible Improvements

- Implement n-grams to capture phrase-level sentiment.

- Use Word Embeddings (e.g., Word2Vec, GloVe) for richer text representation.

- Apply hyperparameter tuning for better model optimization.

- Experiment with deep learning (LSTMs, transformers) for advanced sentiment detection.

## Conclusion

This sentiment analysis system effectively classifies movie reviews as positive or negative. Future improvements can further enhance accuracy and robustness using advanced techniques like word embeddings and deep learning.