# Movie Review Sentiment Analysis Report

## 1. Approach Used

### a) Data Preprocessing:

- Loaded the IMDB dataset.
- Removed special characters and stopwords.
- Tokenized and stemmed the text using NLTK.
- Converted text into numerical representation using TF-IDF vectorization.

### b) Exploratory Data Analysis (EDA):

- Visualized word frequencies using WordCloud.
- Performed sentiment distribution analysis.
- Encoded sentiment labels using one-hot encoding.

### c) Model Training & Evaluation:

- Split data into training and testing sets.
- Trained models including Naïve Bayes, Logistic Regression, and SVM
- Evaluated models using accuracy, precision, recall, and F1-score.

## 2. Challenges Faced

- **Data Imbalance:** Some sentiment classes had more samples than others, affecting model performance.
- **Text Noise:** Presence of special characters, HTML tags, and slang words required extensive preprocessing.
- **Computational Cost:** Training complex models like SVM for best parameters using GridSearch took more time and computation.

## 3. Model Performance & Improvements

- **Naïve Bayes:** Performed well on simple text features but struggled with complex patterns.
- **Logistic Regression:** Achieved balanced accuracy and worked well with TF-IDF features.
- **SVM:** Showed strong performance but was computationally expensive.
- **Improvements Made:**
    - Hyperparameter tuning improved model efficiency.
    - Experimented with different feature extraction techniques.
    - Used GridSearchCV to optimize parameters for better results.

## Conclusion

The sentiment analysis model successfully classifies movie reviews with high accuracy. Future improvements could include deep learning techniques such as LSTMs or Transformer-based models like BERT for better contextual understanding.