### **Name:Angothu Muni Enrollment Number:21CS002361**

### **College:Spsu Group:Batch1**

### **Sentiment Analysis of IMDB Movie Reviews**

**Problem Statement**

The objective is to predict the sentiment (positive or negative) of IMDB movie reviews using various machine learning models. This involves preprocessing the text data, vectorizing it, training different classifiers, and evaluating their performance.

**1. Importing Necessary Libraries**

We imported essential libraries such as NumPy, Pandas, NLTK, Matplotlib, and scikit-learn modules for text preprocessing, feature extraction, model building, and evaluation.

**2. Importing the Training Dataset**

We loaded the IMDB movie reviews dataset from a CSV file using Pandas. The dataset includes reviews and corresponding sentiment labels (positive/negative).

**3. Exploratory Data Analysis**

We performed initial data exploration to understand the structure and content of the dataset. This involved examining the shape of the dataset, displaying the first few rows, and summarizing the sentiment distribution.

**4. Splitting the Dataset**

We split the dataset into training and testing sets. The first 40,000 reviews were used for training, and the remaining for testing.

**5. Text Normalization**

Text normalization steps included:

* **Removing HTML Strips and Noise Text:** We used BeautifulSoup and regular expressions to remove HTML tags and noise text from the reviews.
* **Removing Special Characters:** Special characters were removed using regular expressions to clean the text further.
* **Text Stemming:** Porter stemming was applied to reduce words to their base or root form.
* **Removing Stopwords:** Common English stopwords (e.g., 'and', 'the', 'is') were removed using NLTK's stopwords list.

**6. Feature Extraction**

Two methods were employed for feature extraction:

* **Bag of Words (BoW) Model:** Implemented using CountVectorizer to convert text into numerical vectors.
* **Term Frequency-Inverse Document Frequency (TF-IDF) Model:** Used TfidfVectorizer to transform text into TF-IDF features, which reflect the importance of words in documents.

**7. Model Training and Evaluation**

We trained multiple classification models:

* **Logistic Regression**
* **Linear Support Vector Machines (SVM)**
* **Multinomial Naive Bayes**

Each model was trained on both BoW and TF-IDF representations of the text data.

**8. Model Performance Evaluation**

Model performance was evaluated using:

* **Accuracy Score:** Measures how well the model predicts sentiment correctly.
* **Classification Report:** Provides precision, recall, and F1-score for each sentiment class.
* **Confusion Matrix:** Summarizes the number of correct and incorrect predictions.

**9. WordCloud Visualization**

We visualized frequently occurring words in positive and negative reviews using WordClouds, providing insights into the most common sentiments expressed in the reviews.

**Conclusion**

In conclusion, logistic regression and multinomial naive bayes models demonstrated better accuracy compared to linear SVM for sentiment analysis of IMDB movie reviews. Further improvements could involve advanced preprocessing techniques and leveraging lexicon models for sentiment analysis.