# Sentiment Analysis on IMDB Movie Reviews

Dr. David Raja Micheal School of Advanced Sciences Vellore Institute of Technology

Chennai, India davidraj.micheal@vit.ac.in

Mohamed Suhail School of Advanced Sciences

Vellore Institute of Technology Chennai, India [mohamed.suhail2023@vitstudent.ac.in](mailto:mohamed.suhail2023@vitstudent.ac.in)

**Abstract:**

**This notebook demonstrates how to build a sentiment analysis model using a dataset of IMDB movie reviews. The objective is to classify movie reviews as positive or negative based on the text content. By applying natural language processing (NLP) techniques such as tokenization, lemmatization, and vectorization, the notebook preprocesses the data, then uses machine learning models such as Logistic Regression and Naive Bayes for classification. The notebook also compares the performance of different models and evaluates their accuracy. The ultimate goal is to build an effective model that can predict the sentiment of a movie review based on its text.**

**Keywords:**

**Sentiment Analysis, IMDB, Movie Reviews, Natural Language Processing, Text Classification, Machine Learning, Logistic Regression, Naive Bayes, Data Preprocessing.**

**Objectives:**

The main aim of this project is to analyze and predict the sentiment of IMDB movie reviews. Through sentiment classification, the project strives to identify whether a review expresses positive or negative sentiments. By using machine learning algorithms and evaluating their performance, the objective is to build a robust model capable of accurately predicting sentiments from textual data. Additionally, the project aims to explore and implement text preprocessing steps like tokenization and lemmatization to improve the model's accuracy.

**Introduction:**

In the digital age, online reviews have become a key component for shaping opinions on movies, products, and services. Sentiment analysis, a technique in natural language processing (NLP), is used to determine whether a given text conveys positive, negative, or neutral sentiments. In this project, we use the IMDB movie reviews dataset, which contains thousands of labeled movie reviews, to train a sentiment analysis model. The goal is to build a classifier that can automatically predict the sentiment of a review as positive or negative. To achieve this, various preprocessing techniques are applied to clean and prepare the text data for training machine learning models. The analysis evaluates multiple models to determine the most effective one for this task.

**Methodology**

**Support Vector Machine (SVM):**

In this approach, a Support Vector Machine is used to classify the sentiment of movie reviews. SVM is a supervised machine learning algorithm that works by finding the hyperplane that best separates the data into different classes. In sentiment analysis, the reviews are transformed into feature vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency), and the SVM classifier is trained on these features. The trained model can then predict the sentiment (positive or negative) of new, unseen reviews. The SVM is known for its ability to work well with high-dimensional spaces, making it effective for text classification tasks.

**Decision Tree:**

Decision trees are employed as another classification model in the notebook. A Decision Tree classifier splits the data into branches based on feature values, resulting in a tree structure where each leaf node represents a class label (sentiment). In the case of sentiment analysis, the features (words or word combinations from the reviews) are used to build the tree. Each decision node tests a condition (e.g., whether a certain word appears in the review), and based on the result, the data moves down the appropriate branch. The Decision Tree model is easy to interpret but can sometimes overfit if not properly pruned.

**Naive Bayes:**

The Naive Bayes algorithm is a probabilistic classifier based on Bayes' Theorem, assuming that the features (words in the reviews) are conditionally independent. Despite this simplifying assumption, Naive Bayes is very effective for text classification tasks, especially when the data is high-dimensional. In this case, the Naive Bayes model is trained on the frequency of words in the reviews, and it calculates the probability of each class (positive or negative sentiment) based on the input features. This model is computationally efficient and works well with large datasets, often performing similarly to more complex models.

**Logistic Regression:**

Logistic Regression is a linear classifier used to model the relationship between the input features and the probability of a particular class (sentiment). In sentiment analysis, the reviews are converted into numerical feature vectors, and logistic regression is used to learn a linear decision boundary that separates positive and negative sentiments. The model estimates the probability of a review belonging to the positive class based on the input features and assigns the class label accordingly. Logistic Regression is simple, interpretable, and often performs well for binary classification tasks like sentiment analysis.

**Model Evaluation**

1. **Mean Accuracy**:

Accuracy is the proportion of correct predictions to the total number of predictions made. It gives a general idea of the model’s performance, but it can be misleading if the data is imbalanced. In the context of sentiment analysis on the IMDB dataset, the mean accuracy can be evaluated for each model.

2. F1-Score:

F1-Score is the harmonic mean of Precision and Recall. It is particularly useful when the class distribution is imbalanced, as it provides a better measure of the incorrectly classified cases. F1-Score gives a balance between precision (the accuracy of positive predictions) and recall (the ability to identify positive cases).

3. Classification Report:

A classification report provides several key metrics such as Precision, Recall, F1-Score, and Support for each class (positive/negative sentiment). These metrics give insights into how well the model is performing on each class.

Precision: Proportion of positive predictions that were actually correct.

Recall: Proportion of actual positives that were correctly identified by the model.

F1-Score: Combines Precision and Recall into a single metric.

Support: The number of true instances for each class.

4. Confusion Matrix:

A confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions, broken down by each class:

Conclusion

sentiment analysis on IMDB movie reviews using deep learning models, particularly LSTM (Long Short-Term Memory) for text classification. The approach involves data preprocessing, vectorization, and training a deep learning model to classify reviews as positive or negative. The results highlight the effectiveness of LSTM in handling sequential text data, showing high accuracy in predicting sentiment. The conclusion emphasizes the model's potential for text classification tasks and suggests improvements through model fine-tuning and hyperparameter optimization.