**IMDb Movie Review Sentiment Analysis**

**1. Introduction**

**Problem Statement**

The primary objective of this project is to build a machine learning classification model that can predict the sentiment of IMDb movie reviews. The dataset contains a collection of movie reviews, and each review is labeled as either positive or negative.

**Dataset Information**

The IMDb dataset contains a large number of movie reviews, each labeled with either a positive or negative sentiment.

* **Text of the review**: The actual review provided by the user.
* **Sentiment label**: The sentiment of the review, either "positive" or "negative."

**2. Data Preprocessing**

**2.1 Data Cleaning**

* Removed special characters, punctuation, and stopwords.
* Converted all text to lowercase for uniformity.
* Applied tokenization and lemmatization for text normalization.

**2.2 Data Splitting**

* The dataset was split into training and testing sets using an 80-20 ratio.

**3. Model Implementation**

**3.1 Feature Engineering**

* Used TF-IDF (Term Frequency-Inverse Document Frequency) and Bag of Words vectorizations to convert textual data into numerical features with max feature of 5000 to train our models.

**3.2 Textual Features**:

* Additional features such as word count, character count, average word length, and sentence count were extracted from the preprocessed text.

**4.Machine Learning Models**

Training the model with various Machine Learning models to test our datasets to select the best models for our Sentiment Analysis.

* **Logistic Regression**
* **Naive Bayes**
* **Support Vector Machine (SVM)**:

**5. Model Evaluation**

**5.1 Performance Metrics**

* **Logistic Regression**: Achieved an accuracy of 88.46% with TF-IDF features and 87.29% with BoW features.
* **Multinomial Naive Bayes**: Achieved an accuracy of 85.18% with TF-IDF features and 84.55% with BoW features.
* **Random Forest**: Achieved an accuracy of 84.8% with TF-IDF features and 84.69% with BoW features.
* **Support Vector Machine (SVM)**: Achieved an accuracy of 87.78% with TF-IDF features and 86.47% with BoW features.

**5.2 Results Comparison**

* Logistic Regression has given best results for our model So we are proceeding with this model.
* SVM also provided the nearly same accuracy score but due to computational time constrain we are proceeding only with the Logistic Regression.

**Classification Report Analysis**:

* Our Logistic Regression Model's Overall Accuracy = 88%.
* The model correctly classifies 88% of reviews (both positive & negative).
* This is good performance for a basic logistic regression model.
* F1-Score (Balanced Performance Measure.
* F1-score for both classes ~88-89% → Indicates a well-balanced model.
* No major bias towards positive or negative reviews.

**6. Findings & Conclusion**

* Machine learning techniques effectively classify movie reviews into positive and negative sentiments.
* TF-IDF vectorization with Logistic Regression provided competitive results for classical machine learning models.

**6.1 Key Takeaways**

* **Text Preprocessing**: Proper cleaning and preprocessing of text data are crucial for achieving good model performance.
* **Feature Engineering**: Both TF-IDF and BoW features were effective, with TF-IDF slightly outperforming BoW.
* **Model Selection**: Logistic Regression proved to be the most effective model for this Sentiment Analysis Classification Task.