**Sentiment Analysis of IMDb Movie Reviews -- Report**

**Introduction**

In the digital age, online platforms are filled with user-generated content, including movie reviews. Understanding whether a review is positive or negative is essential for businesses, filmmakers, and recommendation systems. This project focuses on classifying IMDb movie reviews as positive or negative using machine learning techniques.

**Approach**

The dataset used was the IMDb Movie Reviews Dataset (50,000 reviews) from Kaggle, balanced evenly between positive and negative sentiments. The workflow followed these steps:

1. **Data Preprocessing** – Reviews were cleaned by converting text to lowercase, removing HTML tags, punctuation, and stopwords. This step ensured that noise did not affect model learning.
2. **Feature Extraction** – The cleaned reviews were transformed into numerical representations using TF-IDF vectorization, capturing the importance of words relative to all reviews.
3. **Model Training** – Three machine learning models were trained and compared:
   * Logistic Regression
   * Multinomial Naïve Bayes
   * Support Vector Machine (LinearSVC)
4. **Evaluation** – Each model was evaluated using accuracy and F1-score to ensure balanced performance on both positive and negative reviews.
5. **Input Testing** – A terminal-based interface was added, allowing users to type custom reviews and instantly see the predicted sentiment.

**Challenges**

* Text Preprocessing: Reviews often contained HTML tags, special characters, and lengthy descriptions. Designing a preprocessing function that removed noise while keeping meaningful words was critical.
* Class Imbalance Risks: Although the dataset was balanced overall, certain words (like “good,” “bad,” “film”) appeared too frequently, making it challenging for models to learn deeper patterns.
* Model Selection: Different algorithms excel in different scenarios. For instance, Naïve Bayes is fast and simple but sometimes less accurate compared to SVM or Logistic Regression. Balancing speed and accuracy was an important consideration.

**Outcomes**

* Performance: All models performed well, with Logistic Regression and SVM achieving the highest accuracy (around 88–90%), while Naïve Bayes was slightly lower but computationally lighter.
* Interpretability: Logistic Regression provided insights into which words most strongly influenced predictions (e.g., “excellent,” “worst”).
* Practical Application: The manual input function allowed real-time testing. For example, the review “This movie was absolutely wonderful! The acting and story were perfect.” was classified as Positive, while “The plot was boring and the acting was terrible.” was classified as Negative.