IMDb Movie Review Sentiment Analysis

Submitted by: Afsha Khan

Tools Used: Python, Google Colab, Scikit-learn, NLTK, Matplotlib

Date: 24-06-2025

# 1. Introduction

This project aims to build a machine learning model to classify IMDb movie reviews as positive or negative based on their textual content. Sentiment analysis helps in understanding user feedback, improving services, and tailoring content.  
  
The project includes steps like data exploration, text preprocessing, feature engineering, model training, evaluation, and final insights.

# 2. Data Exploration and Preprocessing

## Dataset Overview:

• Total Reviews: [50,000]  
• Labels: Positive (1), Negative (0)

## Exploratory Findings:

• No missing values.  
• Balanced classes.  
• Review length varied significantly.

## Preprocessing Steps:

• Lowercased all text.  
• Removed punctuation, special characters, and stopwords.  
• Tokenization applied using nltk.word\_tokenize.  
• Performed lemmatization.

# 3. Feature Engineering

## Text Vectorization:

• Used TF-IDF to convert text into numerical features.  
• Limited to top 5000 features.

## Additional Textual Features:

• Word count per review.  
• Character count per review.  
• Average word length.

# 4. Model Development

Trained the following 4 classification models:  
• Logistic Regression  
• Naive Bayes (Multinomial)  
• Support Vector Machine (SVM)  
• Random Forest Classifier  
  
All models were trained using the TF-IDF feature set, and train-test split.

# 5. Model Evaluation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.8831 | 0.8834 | 0.8831 | 0.8831 |
| SVM | 0.8784 | 0.8786 | 0.8784 | 0.8784 |
| Random Forest | 0.8465 | 0.8465 | 0.8465 | 0.8465 |
| Naive Bayes | 0.8453 | 0.8453 | 0.8453 | 0.8453 |

Confusion matrices and accuracy charts were generated using Seaborn and Matplotlib.

# 6. Best Model Selection

Logistic Regression was selected as the best model due to:  
• Highest F1-score (0.8831).  
• Balanced precision and recall.  
• Fast, scalable, and easy to interpret.  
• Performs well even with high-dimensional TF-IDF features.

# 7. Conclusion and Insights

• Logistic Regression is effective for text-based sentiment classification.  
• Most reviews contain strong polarity words which the model leveraged well.  
• Preprocessing (especially lemmatization and TF-IDF) significantly improved model performance.

# 8. References

• Imdb\_data Dataset (source)

• <https://scikit-learn.org/>

• <https://www.nltk.org/>

# 9. Video Link

<https://drive.google.com/file/d/18W4GEZuvIgjf4QOKU4qdLQGXBEPLDaub/view?usp=sharing>