
MOVIE
REVIEW

SENTIMENT ANALYSES OF MOVIE REVIEWS

Team Id: T45

T.A: Micheal Mansour

Team Members:

Name	Id	Section
Yasmine Khaled atta	20201701154	Sec 9
Ahmed Esmail Mohamed	20201700024	Sec 1
Haidy Ashraf Mondy	20201700953	Sec 9
Nada Abdallah Mahdy	20201700921	Sec 8
Abdelhamid Magdy Salem	20191700885	Sec 4

Overview

- **Introduction**
- **Data Preprocessing**
- **Data Visualization**
- **Feature Extraction**
- **Model Training and Testing**
- **Models Evaulation**
- **Conclusion**

Introduction

This project focuses on sentiment analysis of movie reviews, utilizing natural language processing techniques to classify reviews as positive or negative. The project aims to develop a sentiment analysis model capable of accurately classifying movie reviews, covering data collection, preprocessing, feature extraction, model training and evaluation.

Data Preprocessing

1. Label Encoding:

- Label encoding is a technique used to convert categorical labels (in this case, sentiment labels: positive or negative) into numerical format.
- 'positive' encoded as 1, and 'negative' as 0.

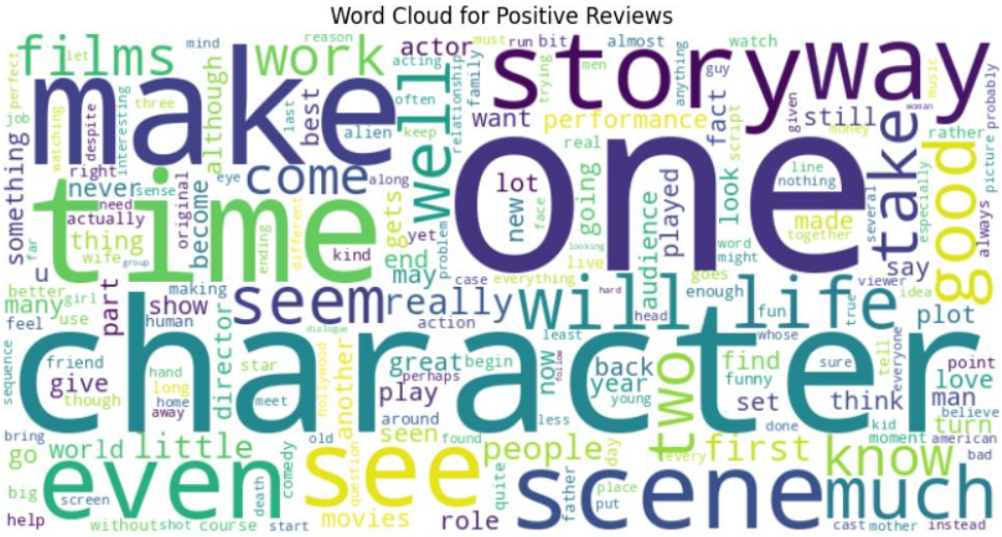
2. Text Cleaning:

- **Lowercasing:**
    - All text data is converted to lowercase to ensure uniformity and avoid duplication of words with different cases.
- **Removing Contractions:**
    - Contractions like "wasn't", "weren't", etc., are expanded to their full forms (e.g., "was not", "were not") for consistency and better analysis.
- **Tokenization:**
    - Text is tokenized into individual words to facilitate further processing.
- **Removing Punctuation:**
    - Punctuation marks are removed to focus on meaningful words only.
- **Stopword Removal:**
    - Common stopwords (e.g., 'the', 'is', 'are') are removed as they do not contribute much to sentiment analysis and may introduce noise.

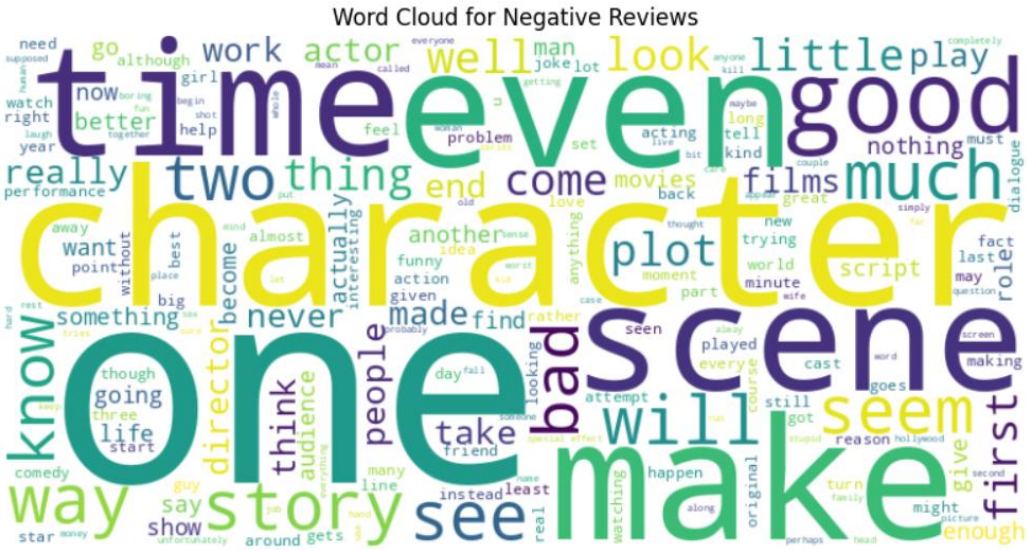
- **Stemming:**
    - Words are stemmed using the Porter stemming algorithm to reduce them to their base or root form.
    - For example, 'running', 'ran', 'runs' would all be stemmed to 'run'.
    - This helps in reducing the dimensionality of the feature space and capturing the essence of words regardless of their variations.

Data Visualization

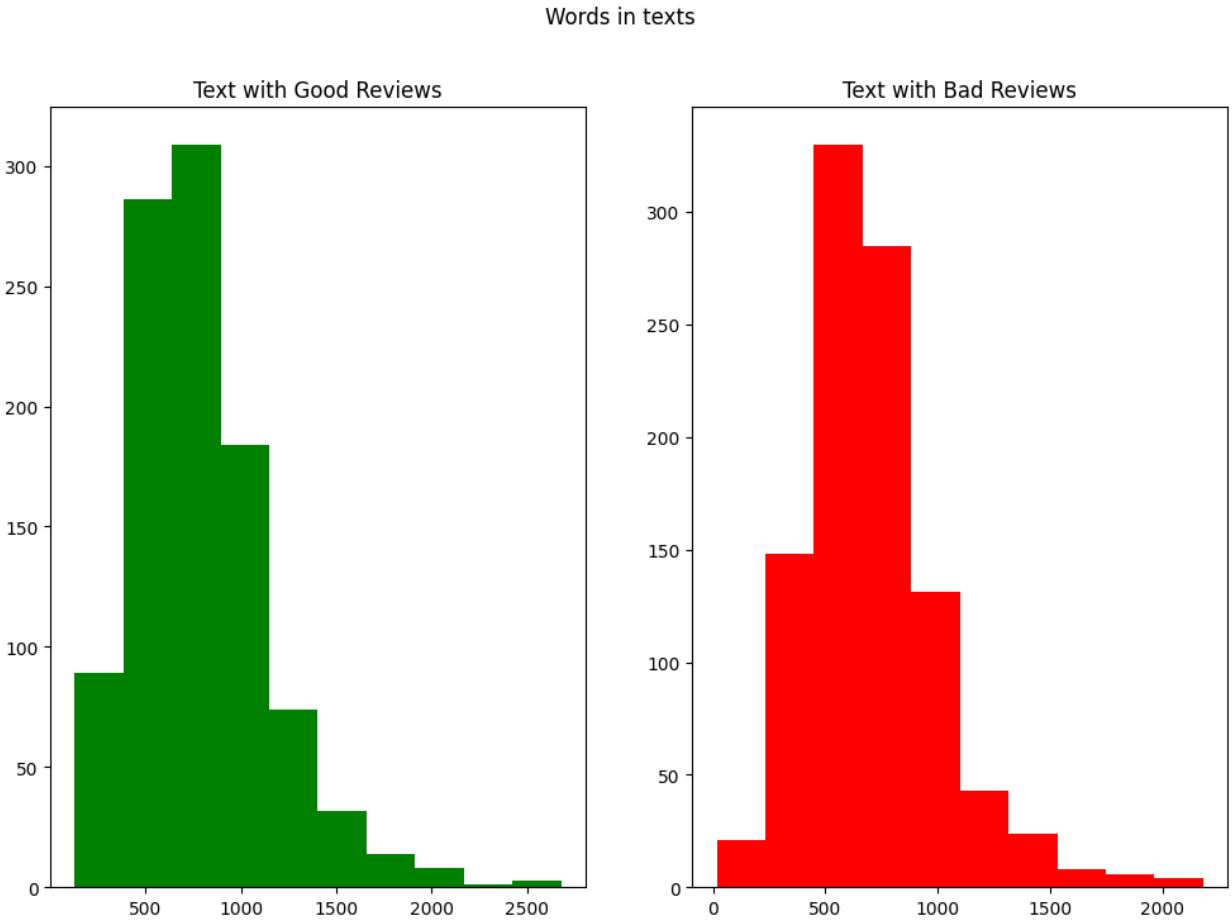
- **Generate wordcloud**
    - word cloud for positive reviews:



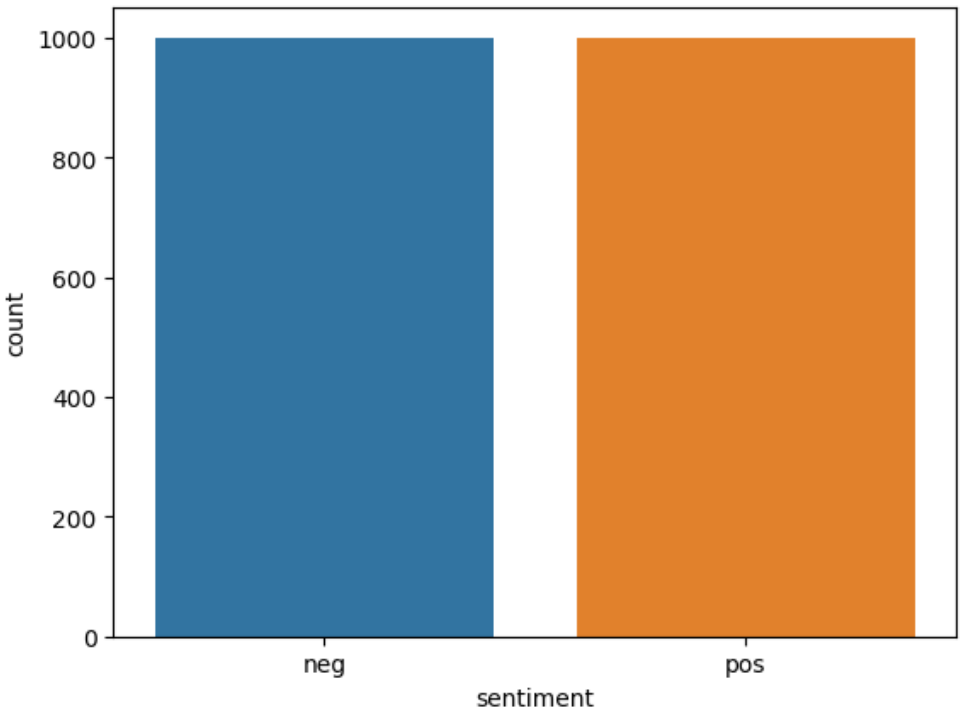
    - word cloud for negative reviews:



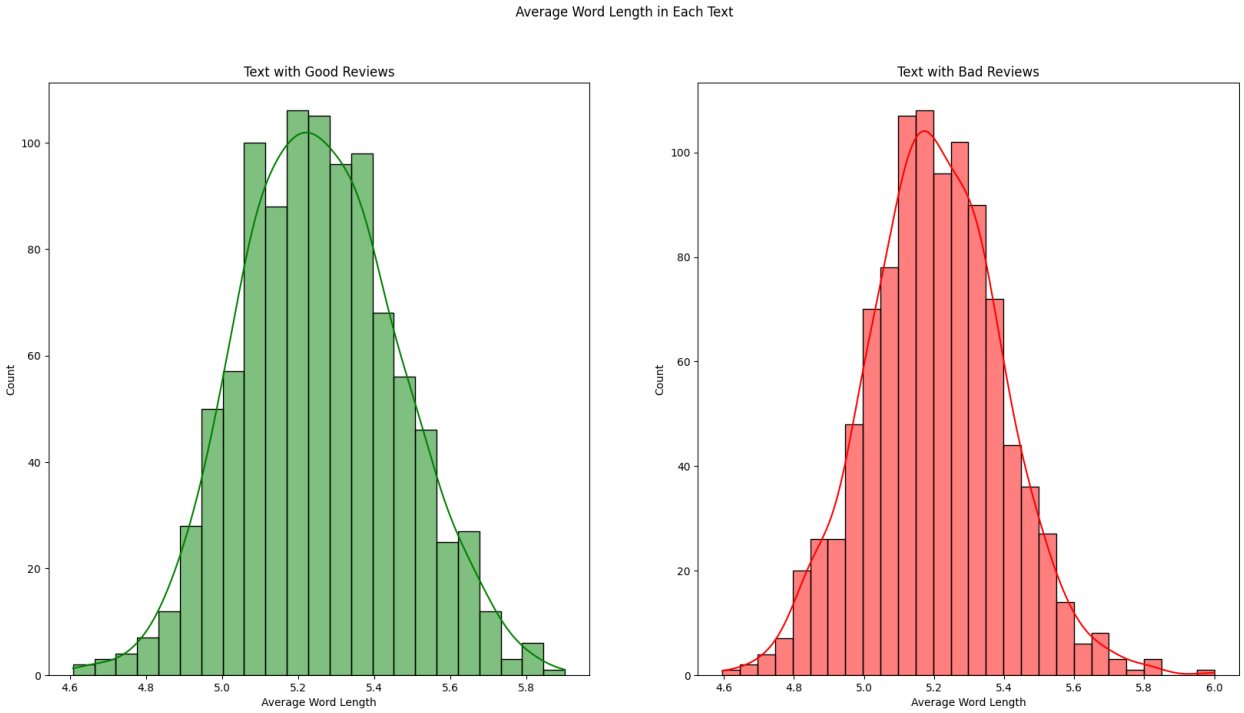
- **histogram of the word counts**



- **Count plot**



- **average word length in reviews**



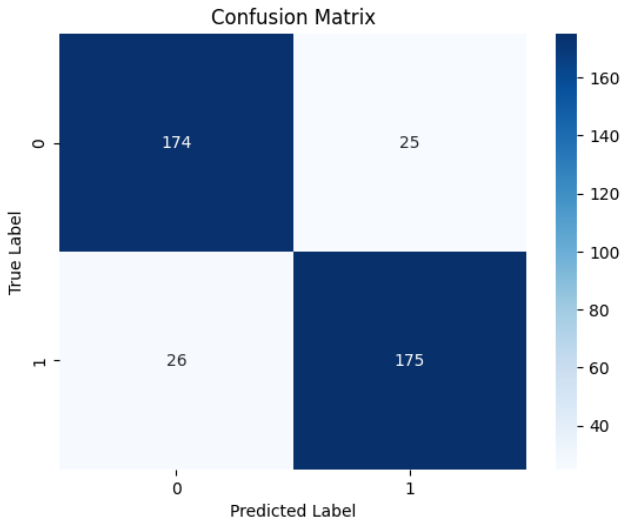
Feature Extraction

TF-IDF: is a common technique in text analysis tasks, including sentiment analysis. It converts textual data into numerical features that can be fed into machine learning models.

Model Training and Testing

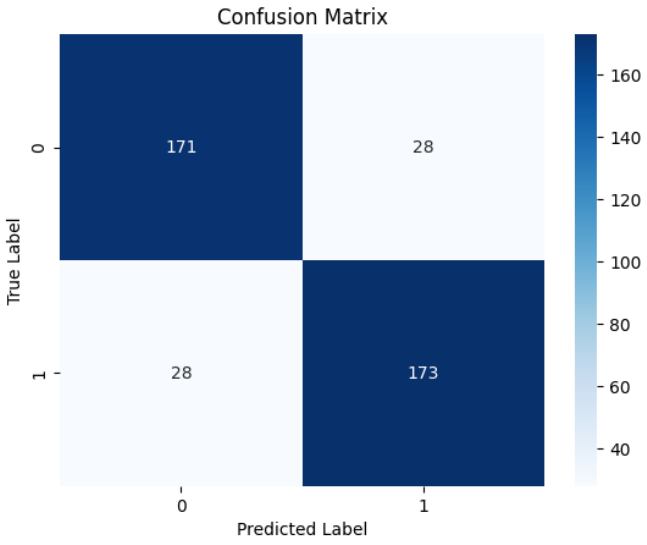
1. Support Vector Machine (SVM):

- SVM is trained with different hyperparameters using a grid search approach (GridSearchCV).
- The best performing SVM model with hyperparameters (C=1, gamma=1, kernel='linear') achieves an accuracy of 87.25% on the testing data.



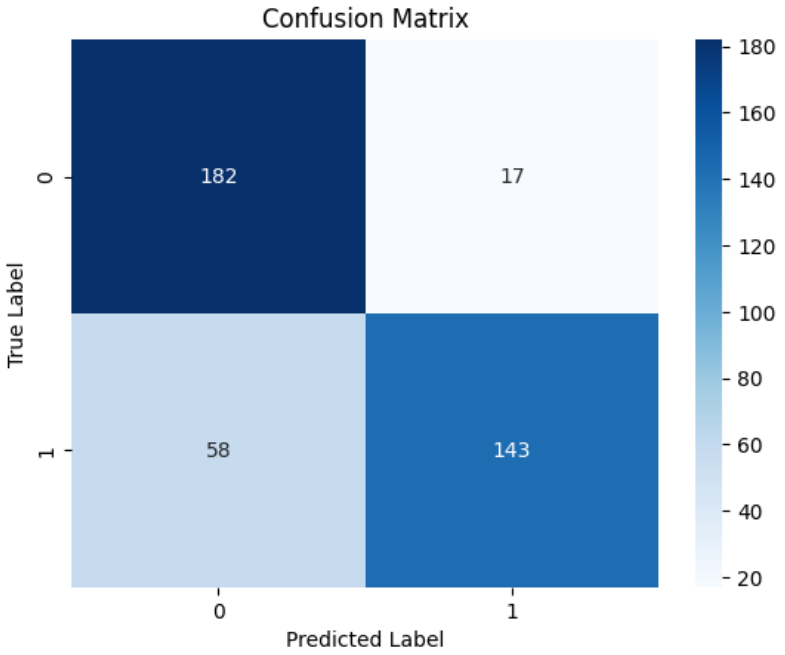
2. Logistic Regression:

- Logistic Regression is trained with various hyperparameters (solver, penalty, and C) using grid search (GridSearchCV).
- The best logistic regression model with hyperparameters (C=10, penalty='l2', solver='newton-cg') achieves an accuracy of 83.52% on the testing data.



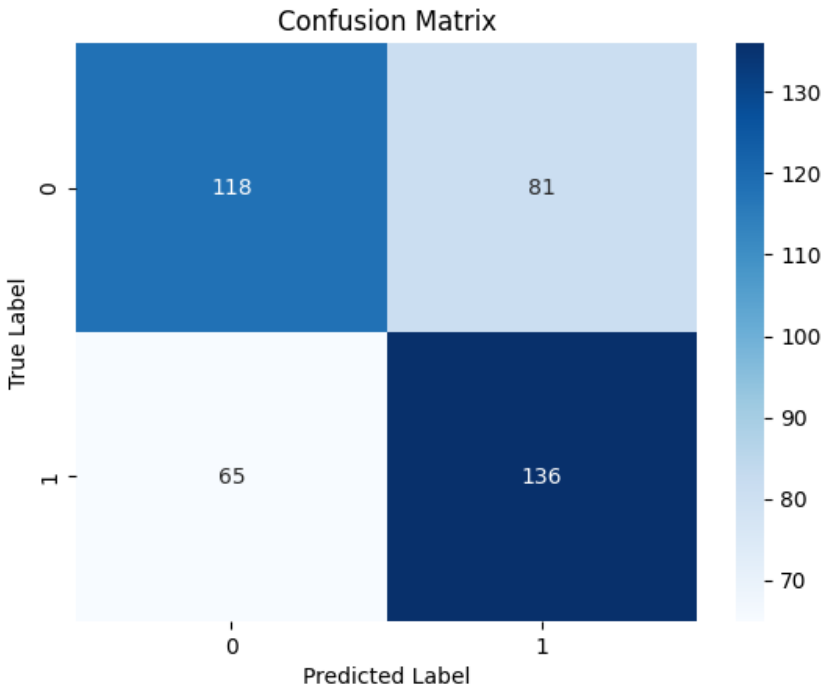
3. Random Forest:

- Random Forest classifier with 100 estimators is trained without hyperparameter tuning.
- The accuracy obtained on the testing data is 81.25%.



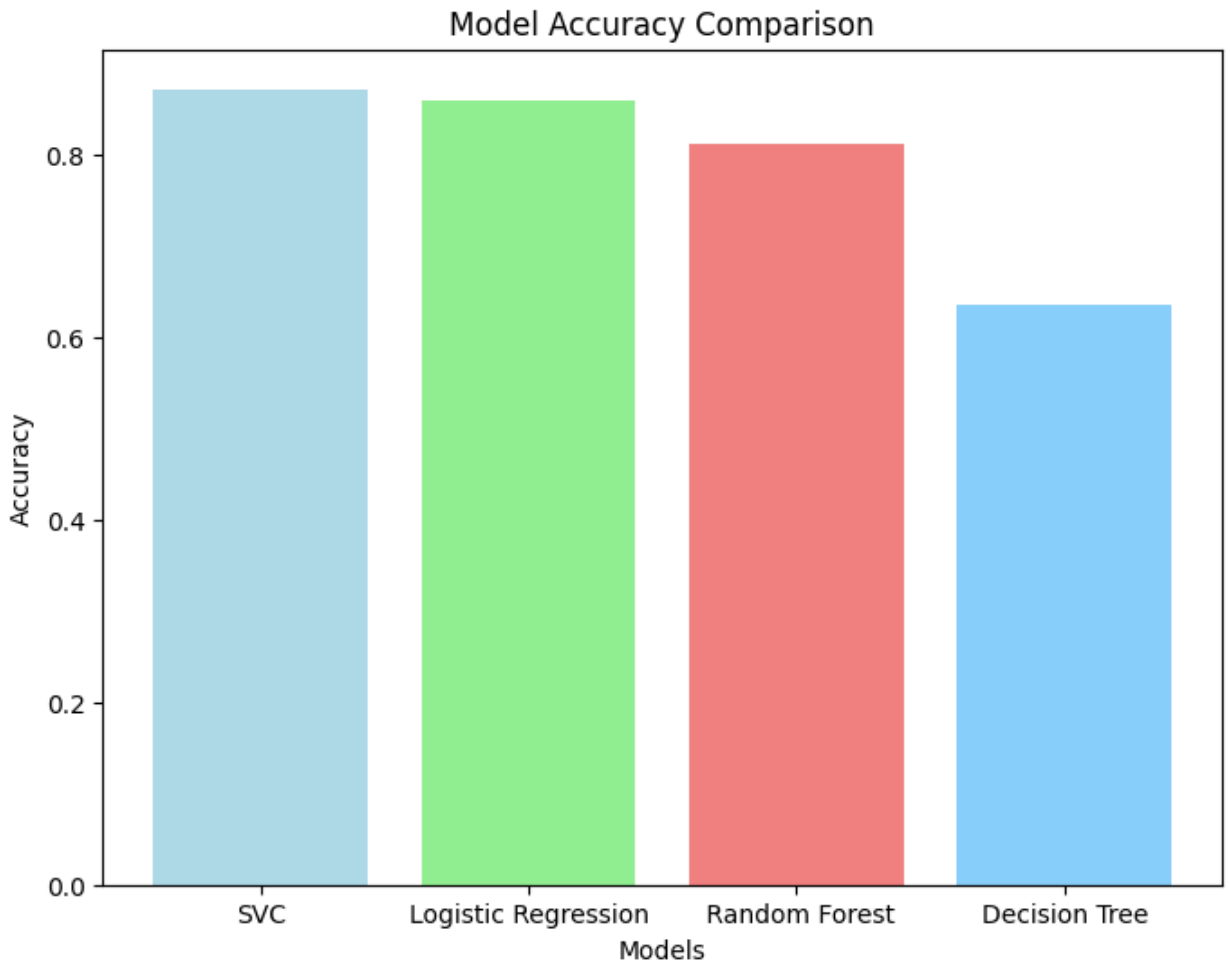
4. Decision Tree:

- Decision Tree classifier is trained without hyperparameter tuning.
- The accuracy obtained on the testing data is 63.5%.



Models Evaulation

- Among the classifiers tested, SVM with linear kernel achieved the highest accuracy of 87.25%, followed by logistic regression with an accuracy of 83.52%.
- Random Forest classifier performed moderately well with an accuracy of 81.25%, while Decision Tree classifier showed lower performance with an accuracy of 63.5%.
- **Model Accuracy Comparison:**



Conclusion

In our project, various machine learning models were explored for sentiment analysis of movie reviews using TF-IDF vectorization. After thorough evaluation, the Support Vector Machine (SVM) classifier with a linear kernel and TF-IDF feature extraction emerged as the best model, achieving the highest accuracy of 87.25% on the testing data, making it the optimal choice for classifying movie reviews into positive and negative sentiments.