Machine Learning Assignment

PROJECT REPORT

TEAM ID : 33

PROJECT TITLE : Classification for Sentiment Analysis of IMDb Reviews

Name	SRN
DARSHAN N	PES2UG24CS808
GAGAN SS	PES2UG23CS193

Problem Statement

In today's digital age, online movie reviews contain valuable insights into audience opinions that can influence box office trends and film production decisions. Manually analyzing thousands of text reviews is time-consuming and prone to bias.

This project builds an automated sentiment analysis system using the **Stanford Sentiment Treebank (SST-3)** dataset, which classifies movie reviews as Positive, Neutral, or Negative. The system demonstrates how machine learning models can efficiently identify sentiment patterns and provide insights into audience reactions.

Objective / Aim

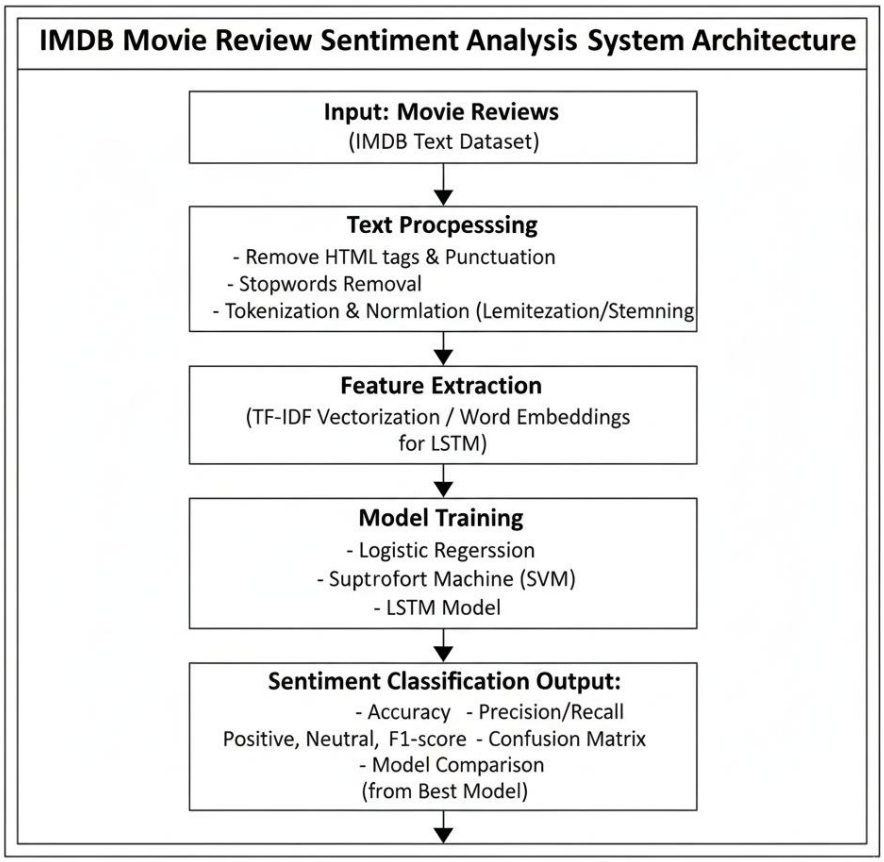
The goals of this project are to:

- Develop a machine learning model to classify movie reviews into **Positive, Neutral, or Negative** sentiment categories using the SST-3 dataset.
- Implement **text preprocessing and feature extraction techniques** for natural language data, including tokenization, stop word removal, and TF-IDF vectorization.
- Train and evaluate **Logistic Regression** and **SVM** models to compare their performance on multi-class sentiment classification.
- Visualize results using **confusion matrices** and **word clouds** to interpret model predictions and key features of each sentiment category.
- Provide a foundation for future enhancements, such as using **deep learning models (LSTM, BERT)** to capture context and improve accuracy on neutral or mixed-sentiment reviews.

Dataset Details

- **Source:** Stanford Sentiment Treebank (SST-3)
- **Size:** ~11,855 reviews total (8,544 for training, 1,101 for validation, 2,210 for testing)
- **Key Features:**
    - review: Text of the movie review
    - sentiment: Label representing review polarity (0 = Negative, 1 = Neutral, 2 = Positive)
- **Target Variable:** Sentiment (Negative / Neutral / Positive)

Architecture Diagram



Methodology

1. **Data Loading:** Load SST-3 train, validation, and test sets.
2. **Parse Tree Format:** Extract review text and corresponding sentiment labels from SST-3 tree format.
3. **Text Preprocessing:**
    - Remove special characters and numbers
    - Convert text to lowercase
    - Tokenize and remove stopwords
4. **Feature Extraction:** Use **TF-IDF vectorization** to convert text into numerical feature vectors.
5. **Model Training:** Train **Logistic Regression** and **SVM** classifiers on the training data.

6. **Evaluation:** Evaluate models on the test set using:
    - Accuracy
    - Precision, Recall, F1-score (per class)
    - Confusion matrix
7. **Visualization:**
    - Confusion matrices for model predictions
    - Word clouds for Positive, Neutral, and Negative reviews
8. **Prediction on New Reviews:** Test model predictions on sample review texts.

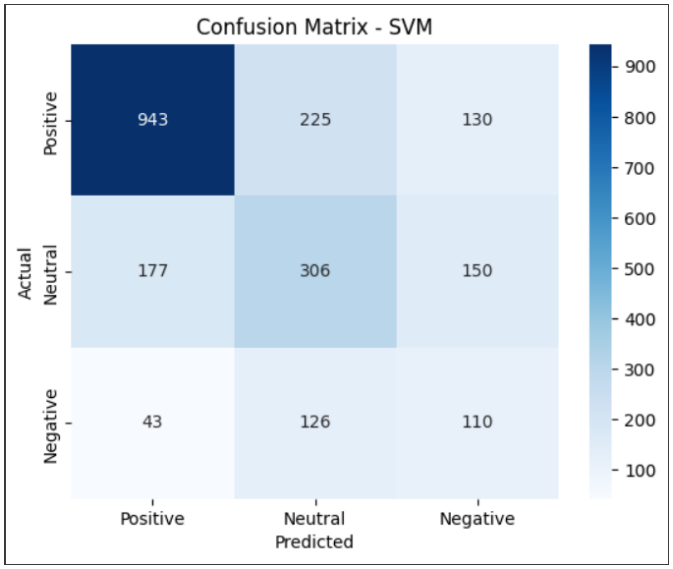
Results & Evaluation

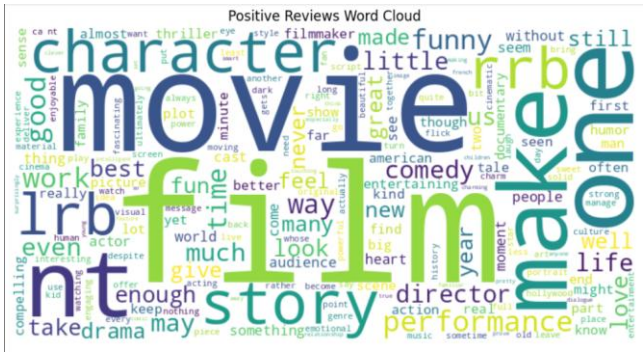
```
Training samples: 9645
Test samples: 2210
```

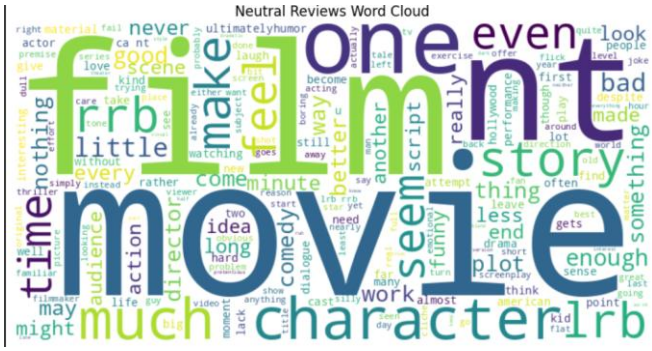
	review	sentiment
0	The Rock is destined to be the ...	Positive
1	The gorgeously elaborate continuation ...	Positive
2	SingerVcomposer Bryan Adams contributes ...	Positive
3	You 'd think by now America would ...	Positive
4	Yet the act is still charming here .	Positive

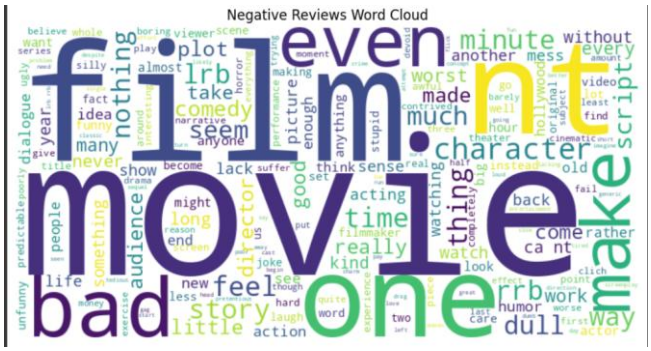
```
=== SVM Results ===
Accuracy: 0.6149321266968326
              precision    recall  f1-score   support

    Negative       0.28      0.39      0.33       279
     Neutral       0.47      0.48      0.47       633
    Positive       0.81      0.73      0.77      1298

    accuracy                           0.61      2210
   macro avg       0.52      0.53      0.52      2210
weighted avg       0.65      0.61      0.63      2210
```

```
=== Logistic Regression Results ===
Accuracy: 0.6257918552036199
              precision    recall  f1-score   support

    Negative       0.31      0.38      0.34       279
     Neutral       0.46      0.45      0.46       633
    Positive       0.79      0.76      0.78      1298

    accuracy                           0.63      2210
   macro avg       0.52      0.53      0.53      2210
weighted avg       0.64      0.63      0.63      2210
```









- **Logistic Regression:**
  - Accuracy: ~62.6%
  - Macro F1-score: ~0.53

    - Strengths: Captures Positive reviews well; struggles with Negative and Neutral balance

- **SVM (Linear Kernel):**
    - Accuracy: ~61.5%
    - Macro F1-score: ~0.52
    - Strengths: Performs similarly to Logistic Regression; slightly better on Neutral class

Evaluation Metrics Used:

- Accuracy
- Precision, Recall, F1-score (per class)
- Confusion Matrix visualization to identify misclassification patterns

Insights:

- Positive reviews dominate the dataset, leading to higher accuracy for Positive class.
- Neutral reviews are harder to predict due to overlapping expressions with Positive/Negative sentiments.
- Word clouds help visualize common words in each sentiment category, providing interpretability of models.

Conclusion

- Successfully built a **multi-class sentiment analysis system** using SST-3 dataset.
- Implemented **Logistic Regression and SVM** classifiers with TF-IDF features for text representation.
- Achieved reasonable accuracy for multi-class sentiment detection, with the best performance for Positive reviews.
- Learned the importance of preprocessing, feature extraction, and evaluation for NLP tasks.
- The project lays the foundation for **future improvements** such as deep learning approaches (LSTM, BERT) and handling imbalanced data to improve performance on Neutral and Negative reviews.