Movie Review Sentiment Analysis Report

Project Overview :

This project aims to classify movie reviews into positive or negative sentiments. The dataset follows a Kaggle competition format with training and test sets. The goal is to train a model that can predict sentiment labels for unseen reviews.

Data Source :

The dataset used comes from the Kaggle challenge: 'Py-Sphere Movie Review Sentiment Challenge'. The training data contains reviews with sentiment labels, while the test data contains reviews without labels.

Workflow :

1. Data Loading:

Loaded training and test data using pandas and confirmed dataset structure.

2. Exploratory Analysis:

Checked class balance, review length distribution, and missing values to understand dataset quality.

3. Text Preprocessing:

Removed HTML tags, URLs, non-alphabetic characters, and standardized text to lowercase.

4. Feature Extraction:

Applied TF-IDF vectorization with 1000 features and removed English stopwords.

5. Baseline Model:

Established a majority-class baseline for comparison with trained models.

6. Model Training:

Trained Logistic Regression and Multinomial Naive Bayes as candidate models.

7. Model Evaluation:

Compared models on validation data using accuracy, classification report, and confusion matrix.

8. Cross-Validation:

Used 5-fold cross-validation for reliable performance estimation.

9. Final Training:

Retrained the best model on the full dataset to maximize data usage.

10. Test Prediction:

Generated predictions for test data and saved them to 'submission.csv'.

Results :

The Logistic Regression model with regularization parameter C=0.1 performed best. It outperformed Naive Bayes on validation accuracy and achieved consistent performance across folds. Final predictions were produced for the test dataset and stored in 'submission.csv'.

Conclusion :

This project demonstrates the process of text classification for sentiment analysis. Classical machine learning such as Logistic Regression are effective when combined with TF-IDF features. The workflow illustrates practical steps from data cleaning to final prediction and competition submission.