Movie review Classification Model

Overview:

This project aims to develop a Machine Learning model which can accurately classify the movie reviews into positive and negative reviews.

Dataset:

The dataset was collected from Kaggle named IMDB Reviews. The dataset consists of nearly 50,000 rows having a Positive- 24884 and Negative-24698 rows.

Link: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data?select=IMDB+Dataset.csv

Steps :

1. Data Collection and Preprocessing:

   The Cleaning steps include following.

   - Clecking all reviews are of English Language
   - Removing Emojis, Hashtags, Emojis, and URL
   - Expanding the Contractions
   - Remove punctuation and numbers
   - Word Tokenization
   - Stopword Removal
   - Lemmatization

2. Exploratory Data Analysis (EDA)
3. Embedding Techniques: Testing various embedding techniques
4. Model Building : Testing various models and choosing the best performing one.
5. Model Optimization and Hyperparameter Tuning: Performing Optimization on the best model.
6. Model testing: Testing with a live example

Model Selection:

Various models were trained and tested the evaluation crietria was based on the F1 score , Precision and Recall of the model. Below give are the evaluation scores of models.

1. Logistic Regression Model:

- **Accuracy:** 0.7833
- **F1-Score:** 0.7719
- **Precision:** 0.8088
- **Recall:** 0.7383
- **AUC-ROC:** 0.7830

Logistic Regression stands out with the highest accuracy and F1-Score , making it the most balanced model. Its high precision indicates it's good at minimizing false positives, but the recall shows it might miss some true positives.

2. SVM:

- **Accuracy:** 0.7767
- **F1-Score:** 0.7616
- **Precision:** 0.8106
- **Recall:** 0.7181
- **AUC-ROC:** 0.8558

SVM has slightly lower accuracy and F1-Score compared to Logistic Regression, but it excels with the highest precision and a strong AUC-ROC , indicating it's very effective at distinguishing between classes, especially when precision is critical.

3. Random Forest Model:

- **Accuracy:** 0.7733
- **F1-Score:** 0.7655
- **Precision:** 0.7872
- **Recall:** 0.7450
- **AUC-ROC:** 0.7731

Random Forest is close to SVM in terms of F1-Score and accuracy . It offers a good balance but doesn't excel in any particular area. It's a solid, reliable model but may not capture the best of both precision and recall.

4. Naive Bayes Model:

- **Accuracy:** 0.7667
- **F1-Score:** 0.7619
- **Precision:** 0.7724
- **Recall:** 0.7517
- **AUC-ROC:** 0.7666

Naive Bayes performs consistently with accuracy and F1-Score , making it a decent choice for simpler tasks. However, it slightly lags behind SVM and Logistic Regression in terms of precision and recall.

5. Layered Neural Network:

- **Accuracy:** 0.66
- **F1-Score:** 0.6483
- **Precision:** 0.6667
- **Recall:** 0.6309
- **AUC-ROC:** 0.8549

Layered Neural Network shows significant weaknesses, with lower accuracy and F1-Score . Despite a decent AUC-ROC , it struggles with both precision and recall , indicating it's not capturing the complexity of the data well.

6. Bidirectional LSTM Model:

- **Accuracy:** 0.66
- **F1-Score:** 0.6483
- **Precision:** 0.6667
- **Recall:** 0.6309
- **AUC-ROC:** 0.5331

Bidirectional LSTM shares similar issues with the Layered Neural Network, performing the weakest overall with an AUC-ROC of 0.5331, suggesting it's not well-suited for this task or requires further tuning and optimization.