Sentiment Analysis of Movie Reviews : IMDB Dataset

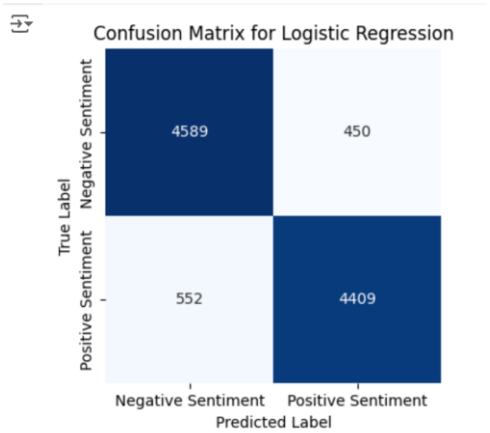
After EDA and preprocessing steps, the dataset is ready to be fed into the models. Two models are chosen to investigate : Logistic Regression and Naïve Bayes models.

After fitting each model on the train data, the predicted results are obtained over the test data. In order to check the performance of the models, classification reports and confusion matrices are extracted.

```
Logistic Regression Accuracy:  0.8998
Naive Bayes Accuracy:  0.8893
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.91      0.90      5039
           1       0.91      0.89      0.90      4961

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.88      0.89      5039
           1       0.88      0.90      0.89      4961

    accuracy                           0.89     10000
   macro avg       0.89      0.89      0.89     10000
weighted avg       0.89      0.89      0.89     10000
```

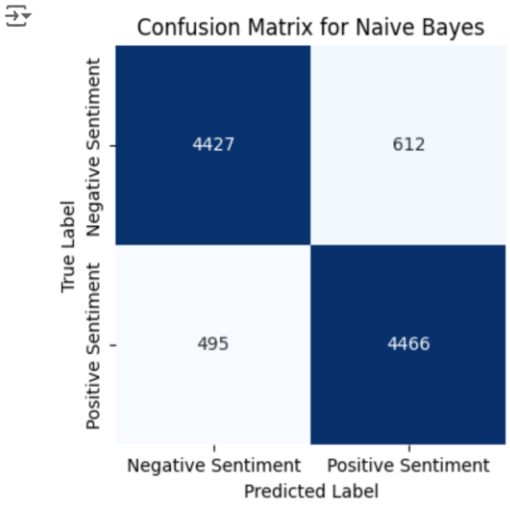
As can be seen, the f1-score metric for Logistic Regression model is a bit higher than for Naïve Bayes model. This shows that Logistic Regression with linear boundary performs slightly better over this dataset.

The confusion matrix for Logistic Regression is as follows:



The matrix shows that almost 4500 positive reviews and almost 4400 negative ones were correctly predicted. More than 10% of positive reviews and less than 10% of negative reviews were precited incorrectly.

The confusion matrix for Naïve Bayes model is as follows:



The matrix shows that almost 4500 positive reviews and almost 4400 negative ones were correctly predicted. More than 10% of negative reviews and less than 10% of positive reviews were precited incorrectly.

Comparing two confusion matrices, shows that Logistic Regression model does better prediction over “negative” reviews, while Naïve Bayes model performs better over “positive” reviews.

The following diagrams are ROC curves for Logistic Regression and Naïve Bayes models. They are almost the same, because both models perform with almost the same quality over the dataset. But based on the AUC factor, the Logistic Regression model performs slightly better than the Naïve Bayes model, which is in accordance with what we discussed earlier based on f1-score metric.