Sentiment analysis of IMDB Movie Reviews

ARUN R DAS

25 SEPTEMBER 2025

Introduction

- **Objective**: Build a sentiment analysis model to classify IMDB movie reviews as Positive or Negative
- **Dataset**: IMDB Movie Reviews Dataset (50,000 reviews – half positive, half negative)

Data Preprocessing

- **Imported required libraries and downloaded NLTK resources.**
- **Performed overview of the dataset.**
- **Applied text cleaning steps:**
  - Removed extra spaces (regex)
  - Removed HTML tags (BeautifulSoup)
  - Expanded contractions (contractions library)
  - Converted to lowercase
  - Removed special characters (regex)
  - Removed stopwords (NLTK corpus)
  - Lemmatized words using WordNetLemmatizer with POS tagging
- **Made a function to perform all preprocessing steps.**
- **Tested the preprocessing and applied to the dataset**
- **Saved the cleaned dataset**

Modelling

- **Imported required libraries**
- **Dataset preparation**
  - Loaded the saved cleaned version of dataset
  - Removed unwanted columns
  - Mapped outputs into numeric values (Positive: 1, Negative: 0)

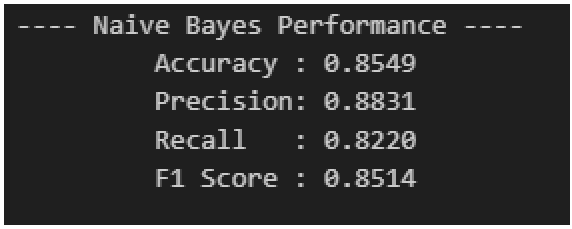
Naive Bayes Classifier

- **Performed TF-IDF Vectorization**
- **Split data into train and test sets**
- **Trained the Naïve Bayes model on training data**
- **Made predictions using the trained model**
- **Evaluated model performance**
  (Accuracy, Precision, Recall, F1 Score)

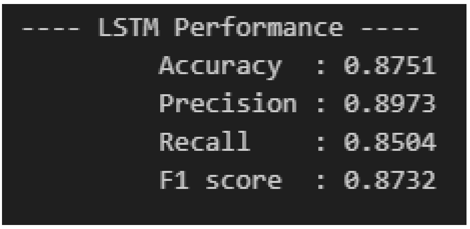
LSTM

- **Performed Tokenization**
  - Converted text into sequences
  - Padded the sequences to same length
- **Split data into train and test sets**
- **Built an LSTM model with Embedding, LSTM and Output layers**
- **Trained Model on training data**
- **Made predictions using the LSTM model**
- **Evaluated model performance**
  (Accuracy, Precision, Recall, F1 Score)

Results and Analysis



```
---- Naive Bayes Performance ----
          Accuracy : 0.8549
          Precision: 0.8831
          Recall   : 0.8220
          F1 Score : 0.8514
```



```
---- LSTM Performance ----
          Accuracy  : 0.8751
          Precision : 0.8973
          Recall    : 0.8504
          F1 score  : 0.8732
```

- **The LSTM model outperformed Naïve Bayes model across all performance metrics (Accuracy, Precision, Recall, F1 Score).**

- **Both models were tested on a new, unseen and ambiguous review:**

  “The film started off painfully slow and the acting felt wooden at times, but halfway through it unexpectedly turned into a gripping, emotional story that left me in tears by the end.”

- **Predictions:**
  - **LSTM –** Positive (Probability: 0.510175)
  - **Naïve Bayes –** Negative

  The results indicate that the LSTM model is better at understanding subtle changes in sentiment better than Naïve Bayes making it more reliable for this dataset.