Sentiment Analysis on IMDB Movie Reviews using Machine Learning (LSTM Networks)

Aryan Gahlaut (3rd year)
Delhi Technological University
`Machine Learning Intern, Cantilever`

August 3, 2025

Abstract

This document provides details of a sentiment-analysis project based on IMDB movie reviews, which utilizes natural language processing (NLP) techniques and a deep learning model to classify sentiments as either a positive or negative sentiment. After going through the necessary preprocessing, tokenisation of the text, and feeding the data to a LSTM-based neural network the model was able to classify similar data well in new unseen data.

Introduction

Sentiment classification of text has many applications: recommendation engines, social media, customer feedback, and other areas. This project looks at an LSTM (Long Short Term Memory) neural network for sentiment classification, a type of neural network particularly good for text data of sequential form.

Dataset

IMDB dataset having 50K movie reviews for natural language processing or Text analytics. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. It provides a set of 25,000 highly polar movie reviews for training and 25,000 for testing.

Preprocessing

- HTML tags and non-alphabetic characters removed using regular expressions.
- Tokens were lemmatized using `WordNetLemmatizer`.
- Stopwords were removed.
- Labels were encoded (positive = 1, negative = 0).

Model Architecture

An LSTM-based deep learning model was designed:

- **Tokenizer:** 10,000 frequent words retained.
- **Embedding layer:** Converts tokens to dense vectors of embedding dimension 128.
- **LSTM layer:** 64 units with dropout of 0.2 and recurrent dropout of also 0.2.
- **Dense output layer:** Sigmoid activation for binary classification.

Training

- Training epochs: 5
- Batch size: 64
- Loss function: Binary cross-entropy
- Optimizer: Adam
- Validation set: 20% split from training data

Results

Model Evaluation

The model was evaluated on unseen test data using accuracy, precision, recall, F1 score, and a confusion matrix.

Classification Report:

- Accuracy: **86.7%**
- Precision: **87%**
- Recall: **87%**
- F1 Score: **87%**

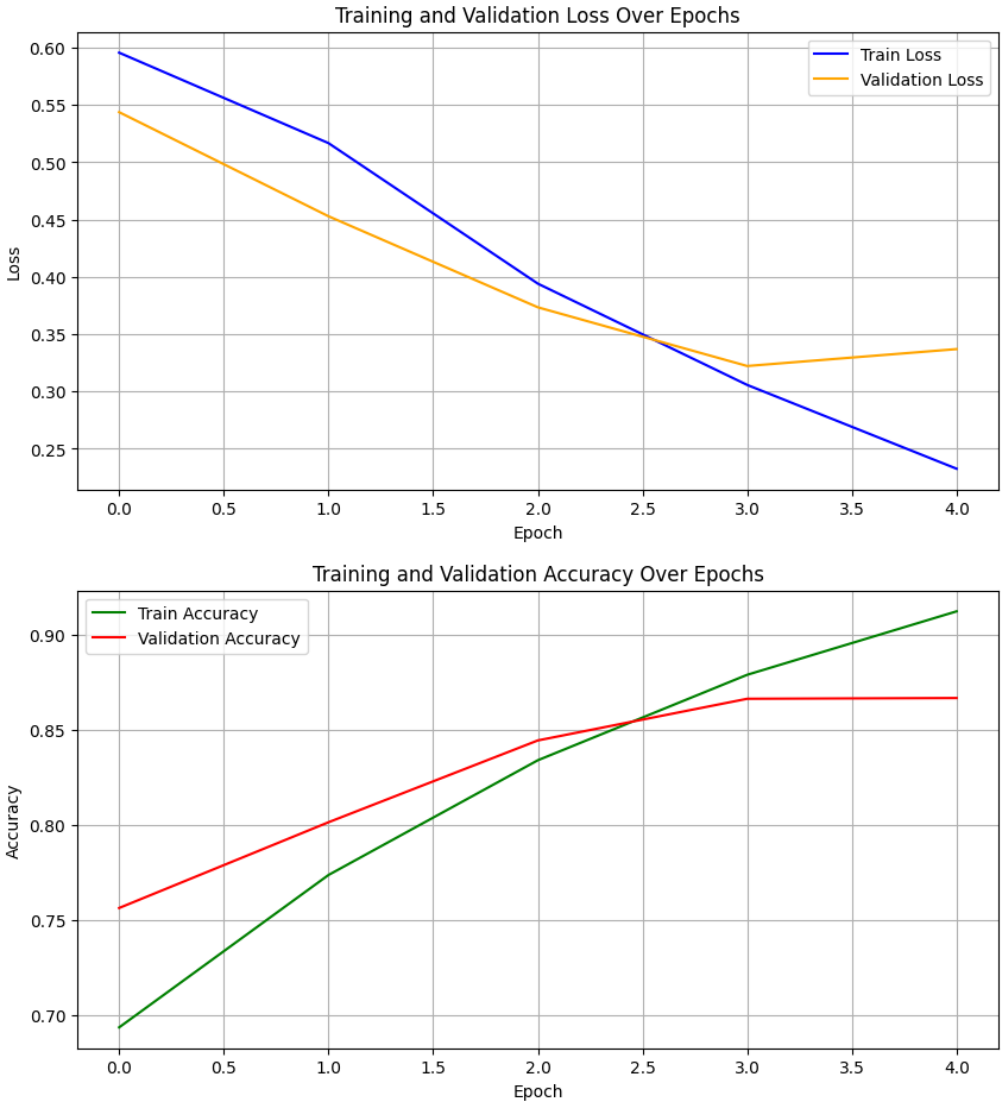

Figure 1: Training and Validation Accuracy/Loss per Epoch

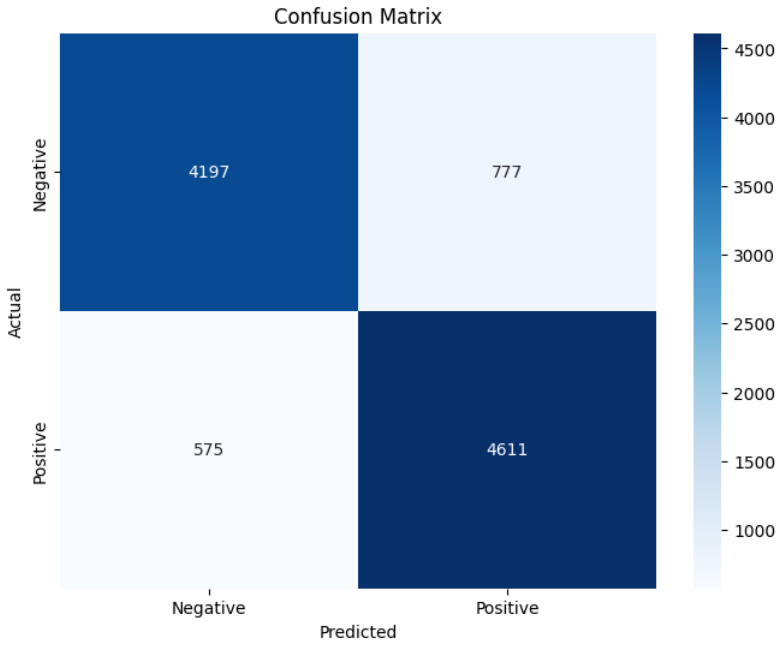

Figure 2: Confusion Matrix on Test Data

Conclusion

The sentiment classifier based on learnable long short-term memory gave great results on IMDB reviews, having the ability to further improve will include utilizing pre-trained word embeddings (e.g., GloVe), using bidirectional LSTM or upgrading to transformer-based models like BERT. UI is also made using Gradio based to input review texts and getting the prediction easily.

References

- IMDB Dataset: `https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k`
- TensorFlow/Keras Documentation: `https://www.tensorflow.org/`
- Jurafsky and Martin. Speech and Language Processing. Pearson.