1. Introduction

This project implements a Sentiment analysis model using Long Short-Term Memory (LSTM) and DistilBERT networks. It involves classifying text into predefined categories, and in this project, reviews are classified into three sentiment categories: negative, neutral, and positive. Optuna is utilized to enhance the performance of the LSTM model and optimize hyperparameters. This report outlines the methodology, results, and conclusions of the sentiment analysis project.

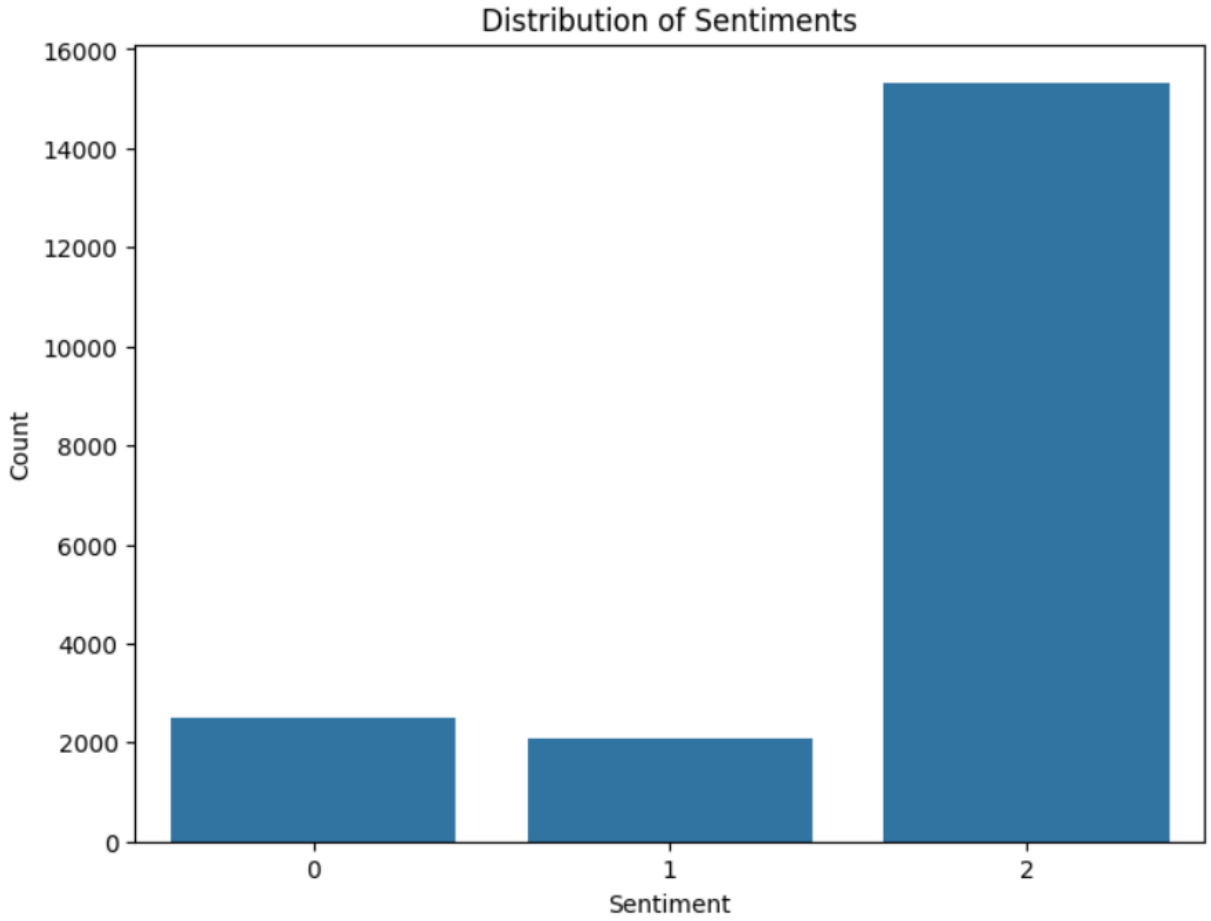
2. Dataset Description

- **Dataset Overview:**
  The dataset consists of Yelp restaurant reviews, with four columns: Yelp URL, Rating, Date, and Review Text. For this project, only the Review Text and Rating columns were used. The Rating column provides a score ranging from 1 to 5, while the Review Text contains the actual review provided by the customer.

- **Data Selection:**
  Given the computational constraints, a random sample of 10,000 reviews was selected from the dataset to train and evaluate the models.

- **Data Distribution:**

Visualization:



As observed from the histogram, the number of positive reviews (4 and 5 stars) significantly outnumbers the reviews with lower ratings (1, 2, and 3 stars). This imbalance indicates that positive sentiments are overrepresented in the dataset compared to negative and neutral sentiments.

3. Data Preprocessing Methods

- **Mapping Ratings to Sentiments:**
  The Rating column, which ranges from 1 to 5, was mapped to sentiment labels as follows:
  - ➔ 1-2 stars: Mapped to 0 (Negative sentiment)
  - ➔ 3 stars: Mapped to 1 (Neutral sentiment)
  - ➔ 4-5 stars: Mapped to 2 (Positive sentiment)

- **Cleaning and Preprocessing:**
  For the purposes of this project, only the Review Text and Rating columns were utilized.
  Missing values in the Review Text or Rating columns were dropped.

- **Text Preprocessing**:

  1. **Tokenization**: For the LSTM model, the text was tokenized into individual words.
  2. **Lowercasing**: All text was converted to lowercase to maintain uniformity.
  3. **Padding**: Sequences were padded to ensure consistent input lengths for the LSTM model.
  4. **DistilBERT Tokenization**: The DistilBERT model used a pre-trained tokenizer to handle tokenization and padding automatically.

- **Data Splitting:**
  The dataset was split into training and testing sets, with 80% used for training and 20% for testing.

4. Methodology

4.1 Model Architectures:

- **LSTM Model:**

The Long Short-Term Memory (LSTM) model is a specialized type of recurrent neural network (RNN) that excels in processing sequential data, making it highly suitable for sentiment analysis tasks such as those applied to Yelp reviews.

Training Configuration:

- **Embedding Layer:** Converts input text into dense vectors, facilitating the model's understanding of the data.
- **First LSTM Layer:**
  - Units: 64 (tunable hyperparameter)
  - Returns sequences for further processing.

- **Dropout Layer:** Applied after the LSTM layers to prevent overfitting, with a dropout rate of 0.2.
- **Second LSTM Layer:**
    - Units: 32 (tunable hyperparameter)
    - Also returns sequences for pooling.
- **Global Max Pooling Layer:** Aggregates the output from the LSTM layers, taking the maximum value across the time steps to reduce dimensionality.
- **Output Layer:**
    - Dense layer with 3 units, using the softmax activation function to output probabilities for the three sentiment classes: positive, neutral, and negative.

Optuna choose the best Learning Rate.

- **DistilBERT Model:**

The transformer-based BERT model is a general-purpose language model pre-trained on a very large corpus of unlabeled text, including the entire English Wikipedia (2500M words), and Book Corpus (800M words). The many parameters allow them to dig deep and learn how language works. However, training the pre-trained BERT model takes a longer inference time, making it difficult to deploy.

Therefore, this concept has seen the proposal of a method to pre-train a smaller general-purpose language representation model called DistilBERT. The smaller DistilBERT model has demonstrated to produce good performances similar to the larger BERT model when fine-tuned on a wide range of downstream tasks.

Training Configuration:

- **Learning Rate**: 2e-5
- **Batch Size**: 16
- **Epochs**: 5
- **Weight Decay**: 0.01

5. Training and Evaluation

The primary metrics for comparing the models were accuracy, precision, recall, and f1-score. The metrics were calculated in terms of the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The accuracy is calculated as the number of all the correct predictions divided by the total number of instances. The precision is calculated as predicted instances which were correctly divided by the size which was the predicted size of the instance. The recall, also known as sensitivity, is measured as the total correctly predicted instances divided by the actual number of instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

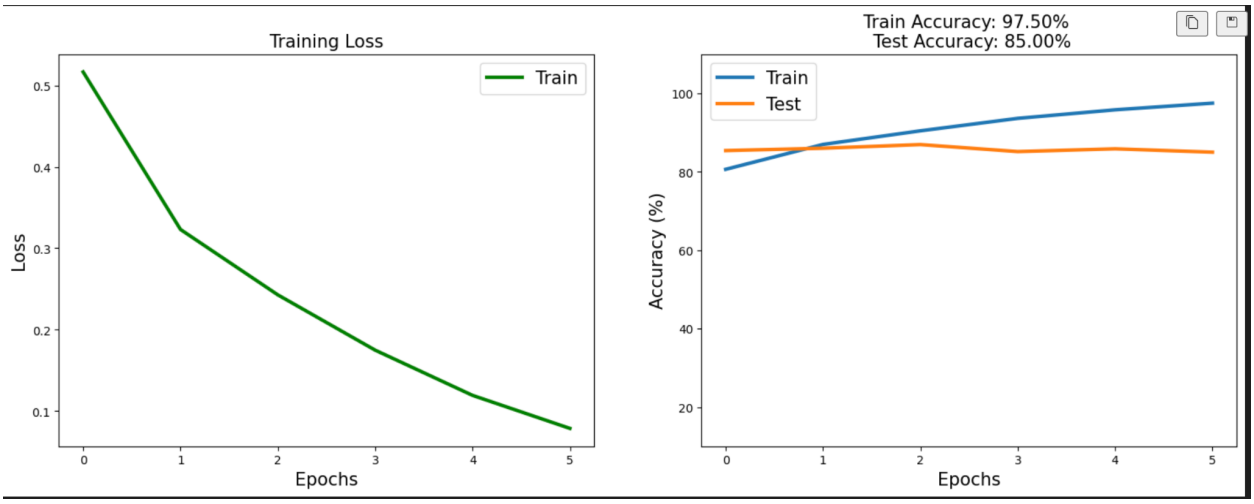
$$\text{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Results

LSTM Model:

The LSTM model was trained on the pre-processed Yelp review dataset. After hyperparameter tuning with Optuna, the final model configuration was selected based on the validation set performance. Below are the results of the LSTM model on the test set:

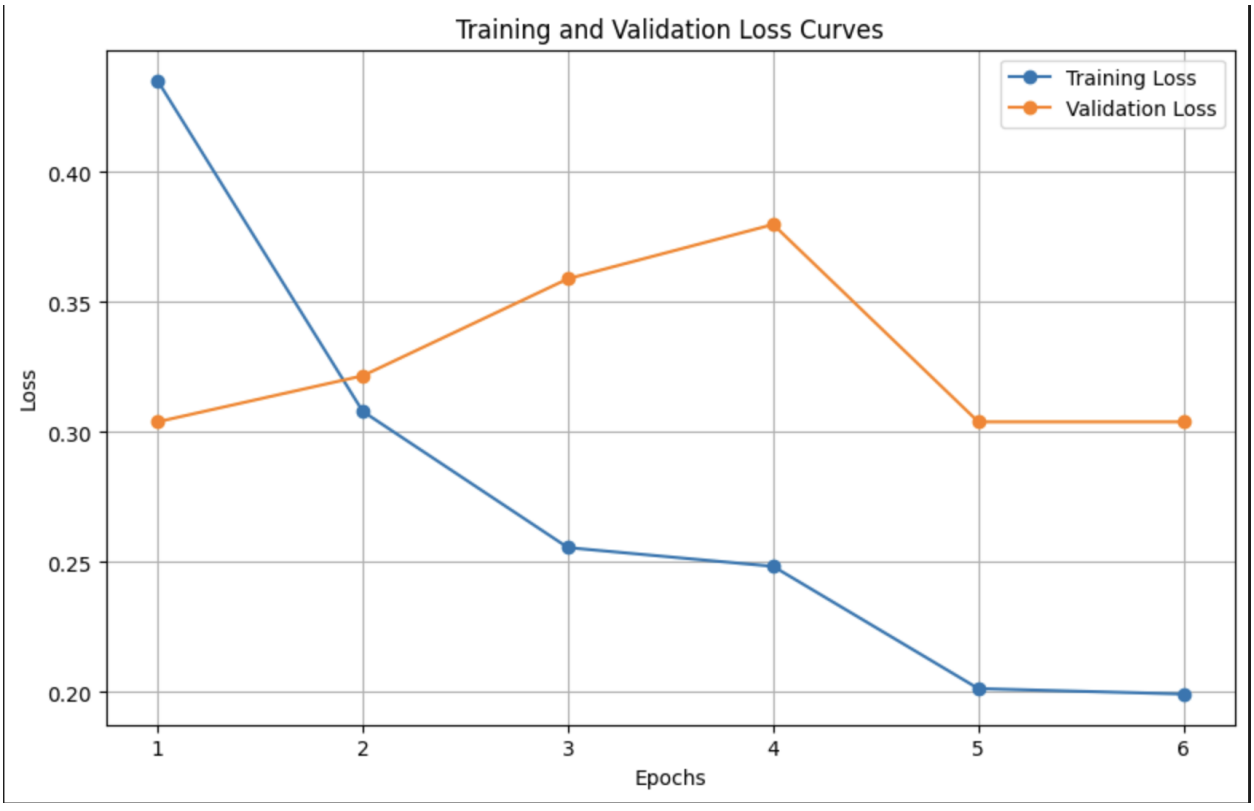
Accuracy: 86%
Precision: 67%
Recall: 66%
F1-Score:67%



DistilBERT Model:

The DistilBERT model, fine-tuned on the same dataset, yielded the following results:

Accuracy: 88.27%
Precision: 88.25%
Recall: 88.27%
F1-Score: 88.23%



Comparative Analysis

The comparative analysis between the LSTM and DistilBERT models highlights several key findings:

Performance: The DistilBERT model consistently outperformed the LSTM model across all metrics. The transformer-based architecture of DistilBERT, with its ability to capture complex dependencies within the text, provided a superior understanding of sentiment in the reviews.

Efficiency: While DistilBERT required fewer epochs to converge, the computational cost per epoch was higher due to the model's complexity. On the other hand, the LSTM model, being less computationally intensive, required more epochs and careful hyperparameter tuning to reach optimal performance.

Interpretability: The attention mechanism in DistilBERT provides a layer of interpretability by highlighting the most influential words in a review for sentiment classification. This is particularly useful for understanding the model's decision-making process. The LSTM model lacks this feature, making it harder to interpret its predictions.

Review Length: DistilBERT demonstrated better performance on longer reviews with more complex language structures, whereas the LSTM model performed comparably on shorter reviews. This suggests that the transformer-based model is more robust to variations in review length and complexity.

Conclusion

In this project, we developed and compared two deep learning models—LSTM and DistilBERT—for sentiment analysis on Yelp reviews. The DistilBERT model outperformed the LSTM model in terms of accuracy, precision, recall, and F1-score, demonstrating its superiority for this particular task. The transformer-based architecture's ability to capture intricate language patterns and the attention mechanism's interpretability were significant advantages. However, the LSTM model's lower computational requirements may make it a more viable option in resource-constrained environments.

Future work could involve exploring more advanced versions of transformer models, such as full-scale BERT or GPT, and experimenting with more diverse datasets to generalize the findings further. Additionally, integrating these models into a real-time sentiment analysis system could provide valuable insights for businesses leveraging user reviews.