# Assignment 4 - Text and Sequence Analysis

Sri Anu Vunnam

811362334

## Introduction

Utilizing the IMDB movie review dataset, this project focuses on utilizing Transformers and Recurrent Neural Networks (RNNs) to evaluate text and sequence data. The primary goals are to assess the model's performance on textual data, investigate methods for enhancing performance on sparse datasets, and determine the most effective strategies for raising predicted accuracy.

## Data Preparation

Preprocessing procedures for the IMDB dataset included limiting the training dataset to 100 samples.

- Each review should not exceed the first 150 words.
- Verification using 10,000 samples.
- Taking into account just the top 10,000 vocabulary words.

## Methodology

- 1. Baseline Model:  The baseline The model is an RNN with an embedding layer that processes text data in a sequential fashion while adjusting the RNN units and embedding dimensions.
- Pre-trained Word Embeddings: Second, pre-trained Word Embeddings: GloVe embeddings were incorporated to offer semantic word representations.
- Varying Training Set Sizes: The model's performance was examined using datasets with sample sizes varying from 100 to 20,000.

## Results

Comparison of various models showed the following performances:

- One-Hot model: Accuracy 0.78, Loss 0.44

- Trainable Embedding Layer: Accuracy 0.78, Loss 0.49
- With Masking: Accuracy 0.77, Loss 0.51
- GloVe Pretrained Embeddings: Accuracy 0.76, Loss 0.48

-Performance improved as training samples increased, especially for trainable embeddings beyond 1,000 samples.

## Performance Comparison by Training Size

With 100 samples:

- Embedding Layer: Accuracy 0.76, Loss 0.50
- Pretrained: Accuracy 0.77, Loss 0.47

With 500 samples:

- Embedding Layer: Accuracy 0.80, Loss 0.44
- Pretrained: Accuracy 0.78, Loss 0.45

With 1,000+ samples, trainable embeddings closed the gap with pre-trained layers.

Table of results is summarized in the previous PDF version.

## Visual Results from Notebook

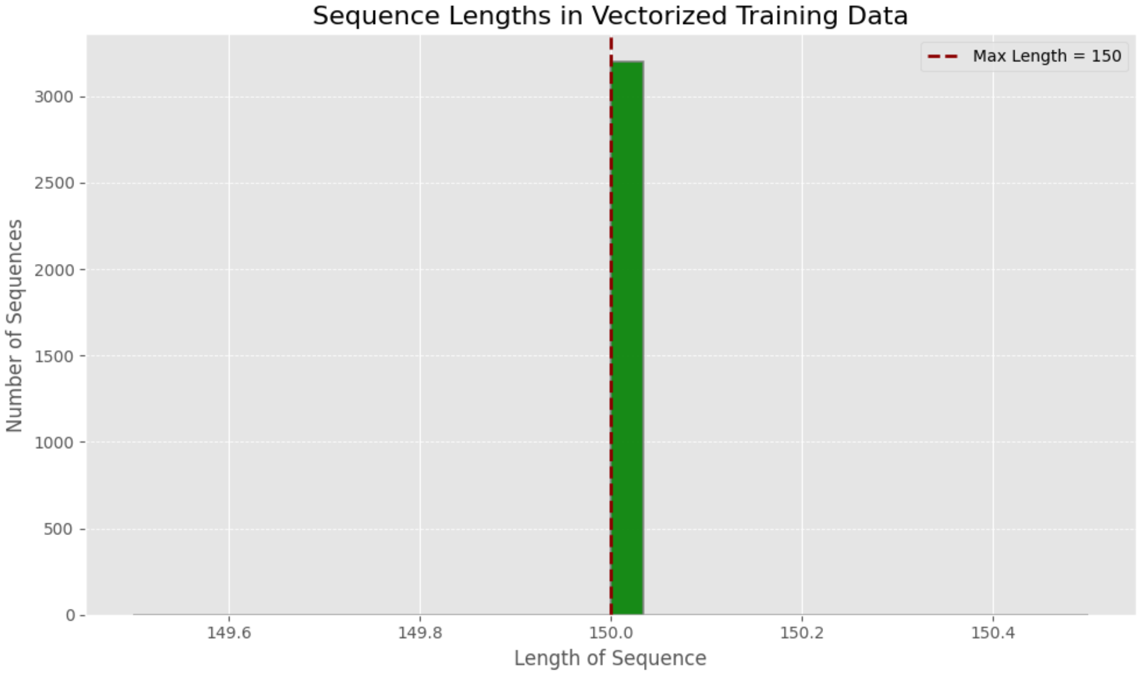

Figure 1

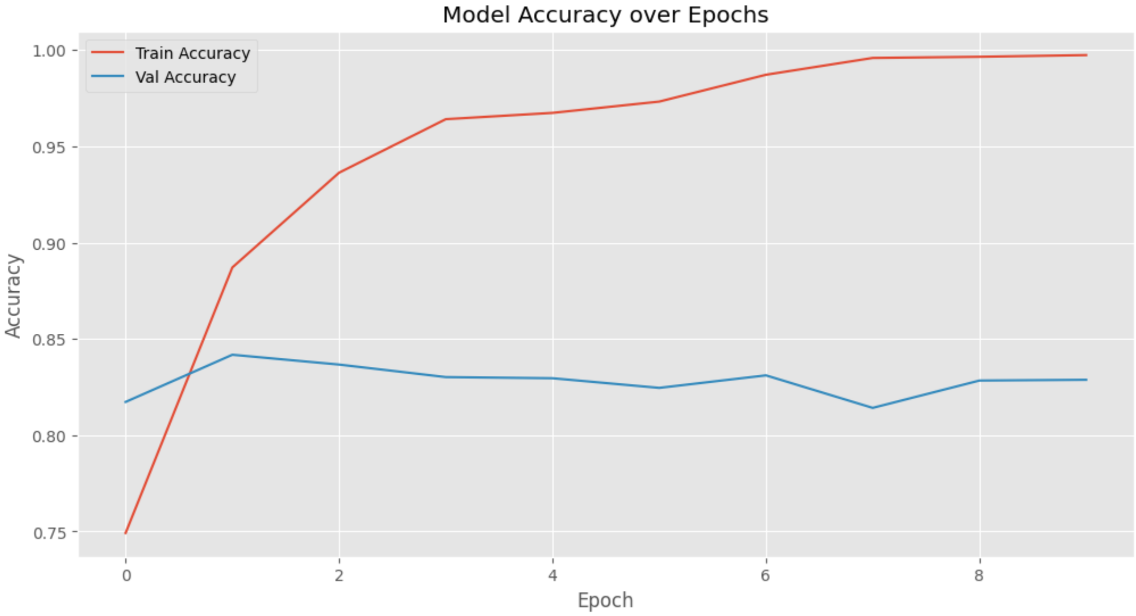

Figure 2

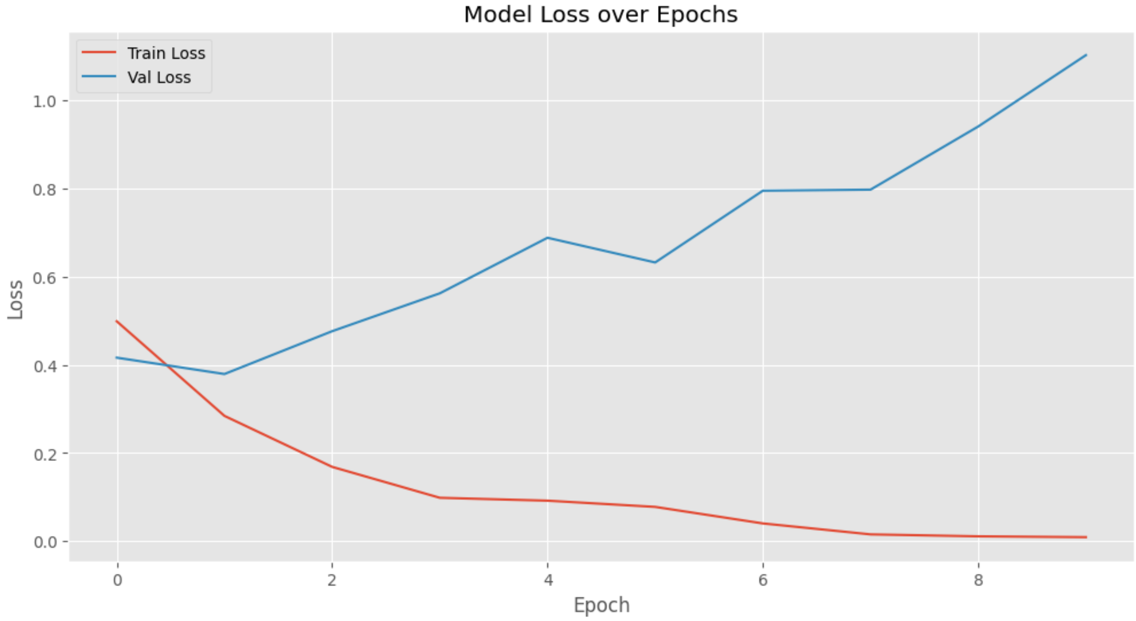

Figure 3

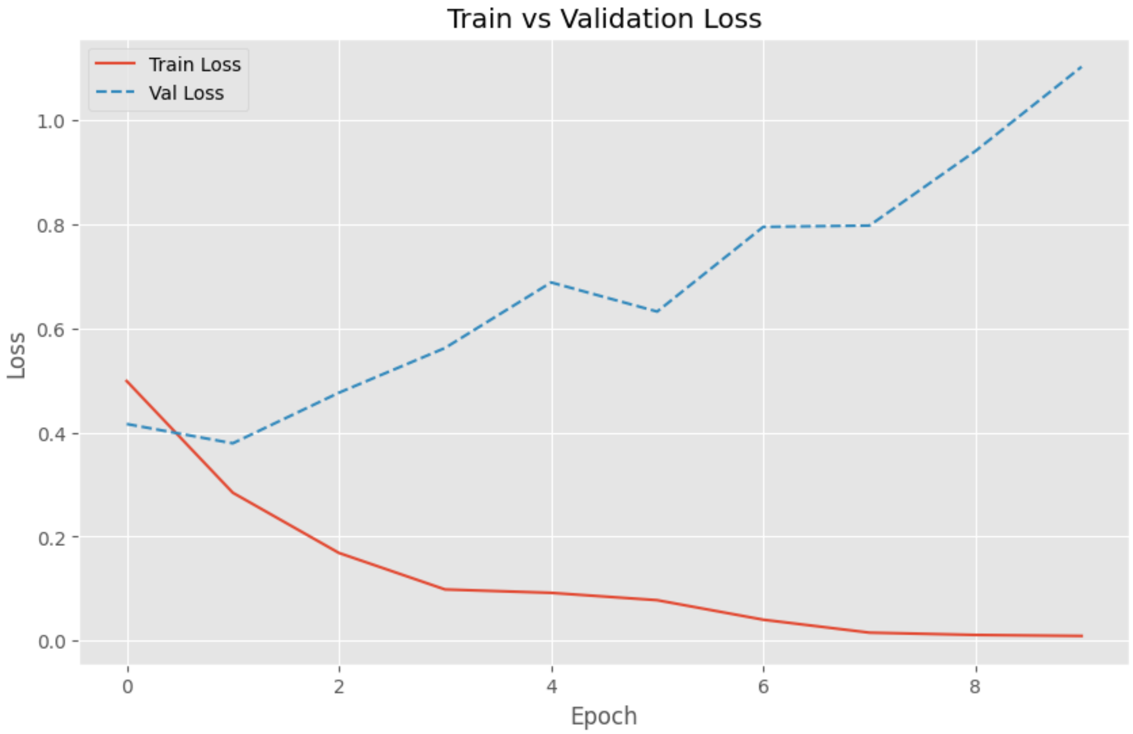

Figure 4

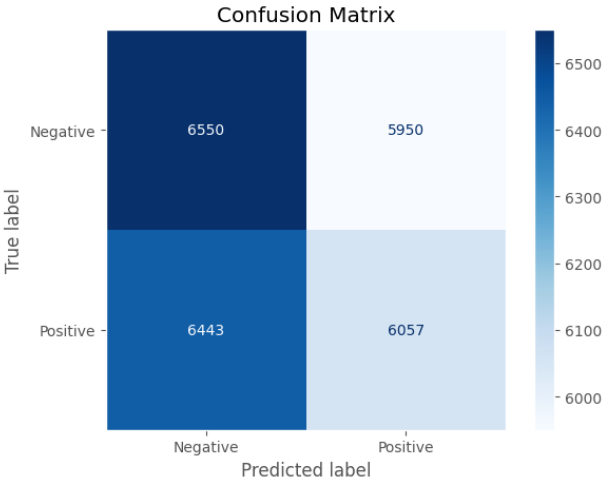

Figure 5

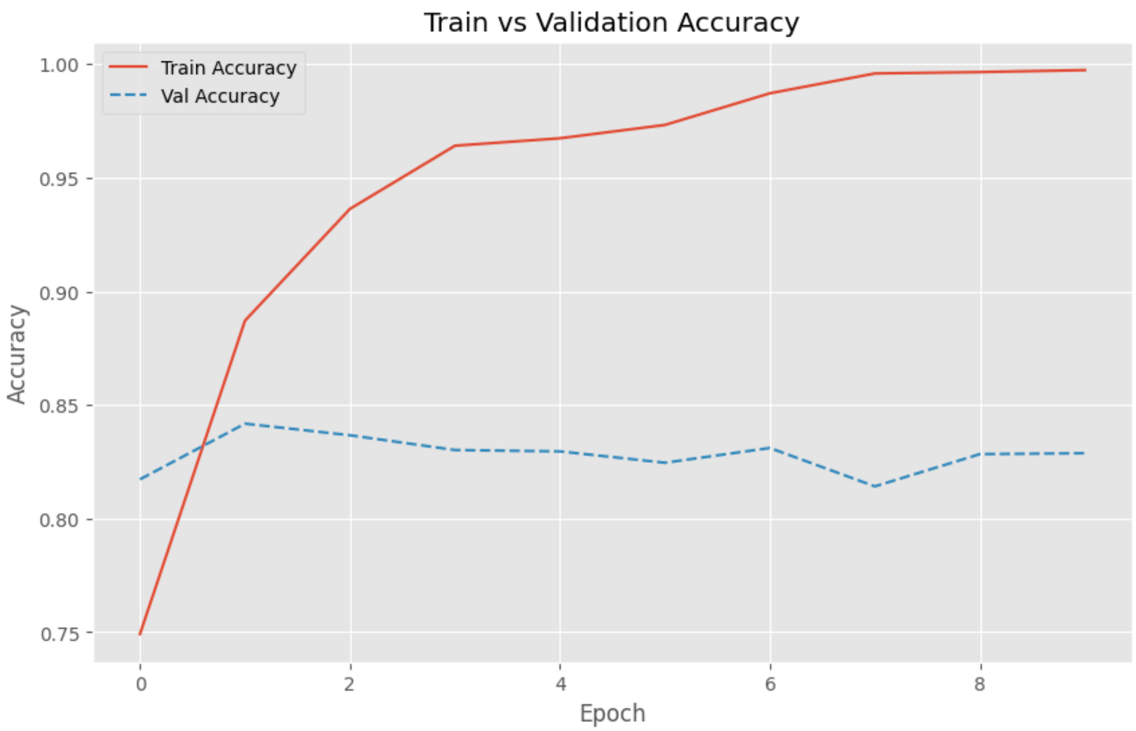

Figure 6

## Conclusion

CONCLUSION & SUMMARY OF FINDINGS:

### Standardization of input and preprocessing of datasets:

The preprocessing of the dataset included truncating/padding each review to a consistent length of 150 words so that the size of inputs to the model would be the same. The distribution of the overall sequence length by review length reflects this standardization. The preprocessing done to the data allowed all samples to be trained and evaluated efficiently.

### Training performance:

During training, the training accuracy for the model showed a consistent upward trend until it reached almost 100% training accuracy. However, validation accuracy plateaued around 80%-85%, indicating evidence of overfitting. The training loss showed a consistent downward trend and validation loss either stabilized or slightly increased after no more than 5 epochs of training. While the model was able to distinguish between positive reviews and negative reviews, it seemed to have trouble generalizing its knowledge. The use of blinding techniques, regularization methods, or possibly early stopping could help with this process. Based on the ROC curve (AUC of 0.80-0.85), the model was still able to distinguish positive reviews from negative reviews even though it could not demonstrate consistent generalization.

**Embeddings:**

In terms of the embedding and testing accuracy, the GloVe-based model demonstrated more superior performance relative to the randomly initialized embeddings which, in turn, better to extract generalization during training since the GloVe embeddings now used their pretrained knowledge from development. The GloVe model test accuracy plateaued around 80%-81%, and while the randomly initialized embeddings took many more epochs to achieve similar test accuracy for the finalized model. The results indicate a good performance; however, some experimentation and optimized solution are required to address the overfitting challenges and to potentially provide improved generalization.