Summary Report

Jhansi Bussa
Hilda Gandu

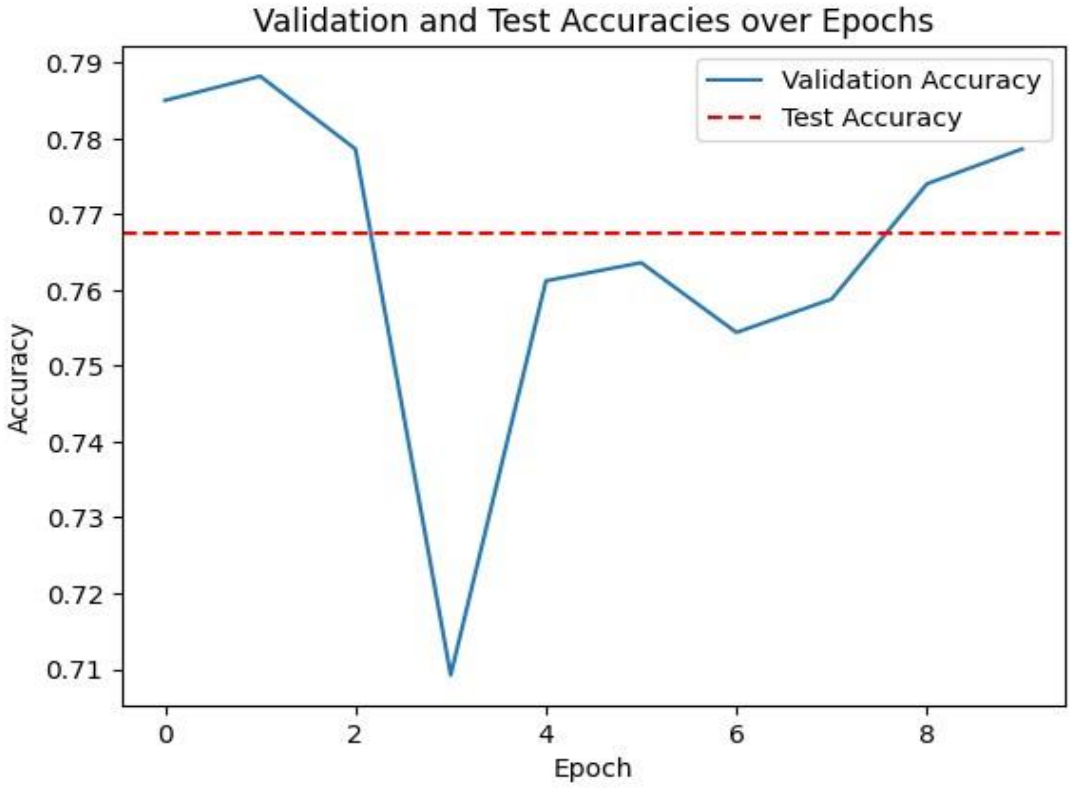
Introduction:

In this assignment, I first started downloading the IMDb movie review dataset, this includes reviews labeled into two categories: positive or negative. After extracting the dataset, I focused on the labeled reviews and eliminated any unsupervised training data.

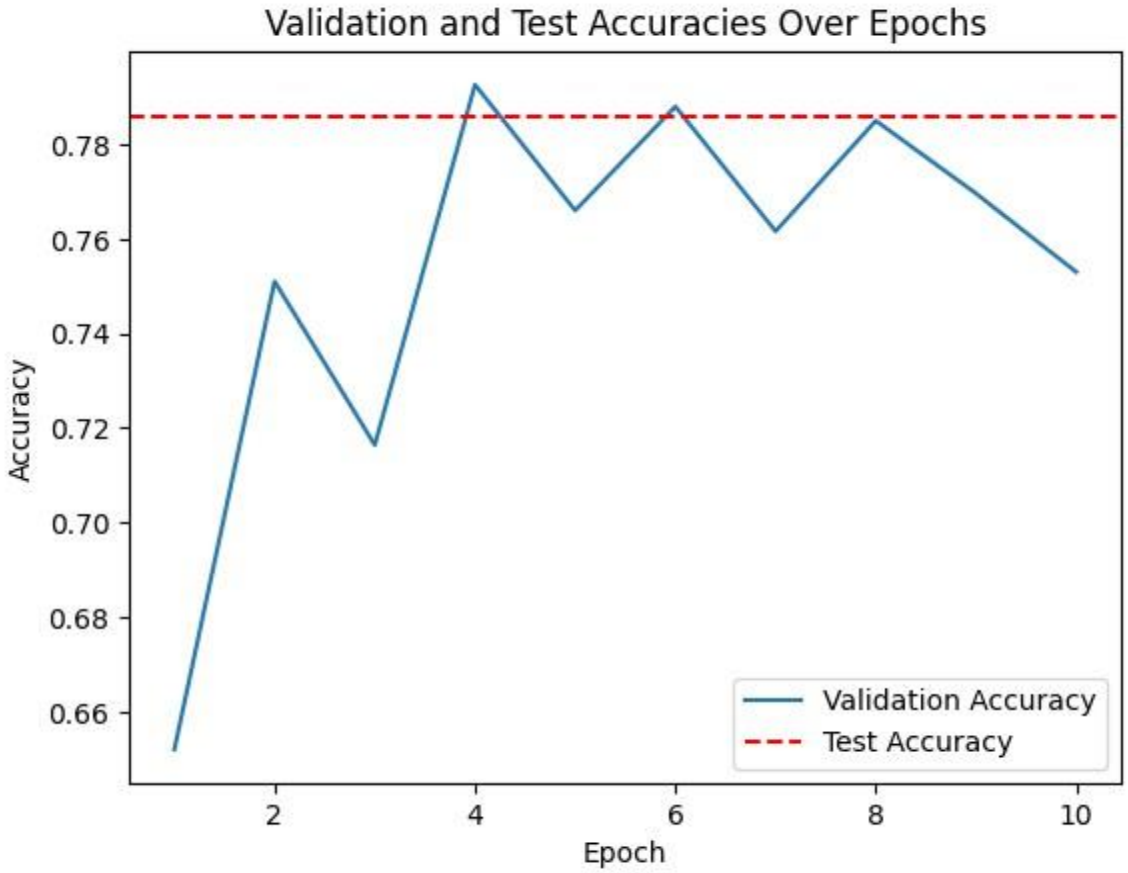
I partitioned the dataset into three categories: training, validation, and test. Using an initial split ratio, files were moved from the original training directory to the validation directory during this step. The data can possibly be divided such that the model is evaluated on one set of data to assess its performance, validated on another set, and trained on a different set.

Modifications:

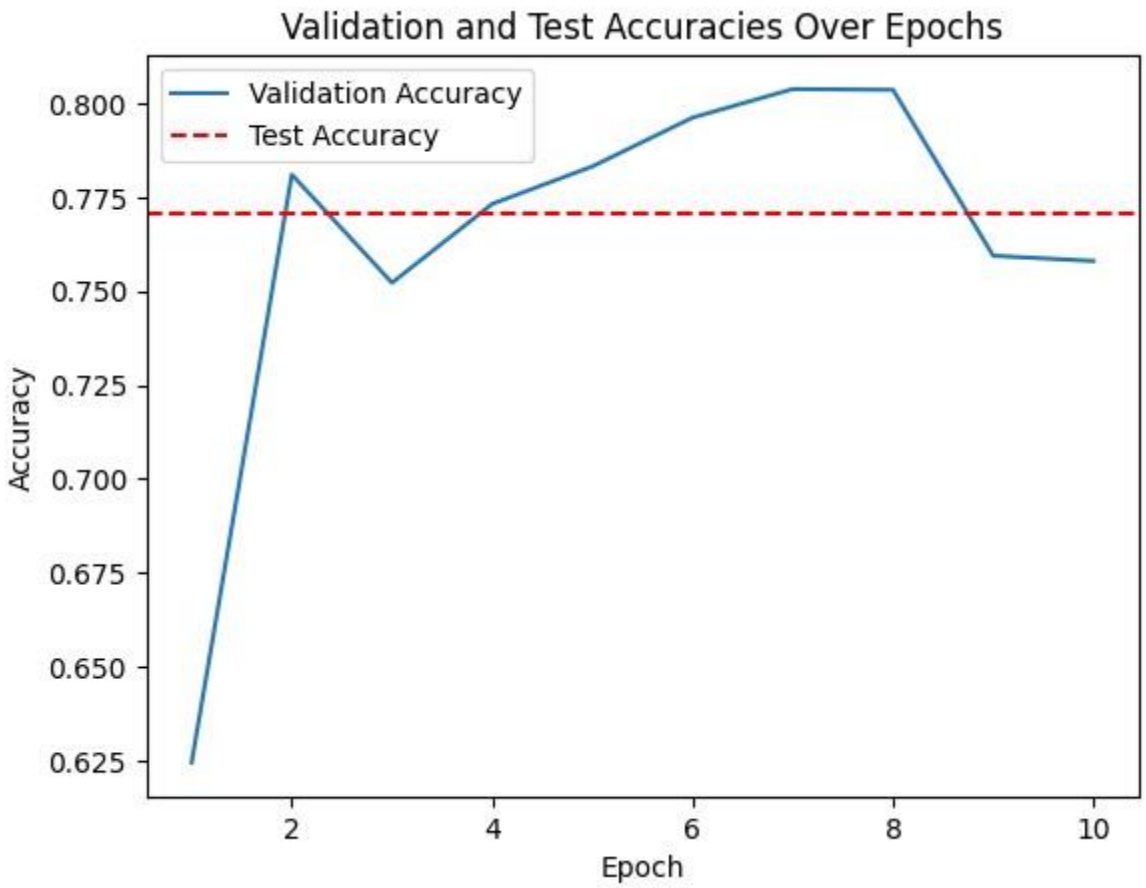
1) After 150 words and 100 training samples, the model's validation accuracy is around 77.9% and test accuracy is 76.7% for the cutoff. This shows that even with short review times and small sample sizes, sentiment analysis jobs can still deliver acceptable levels of accuracy.



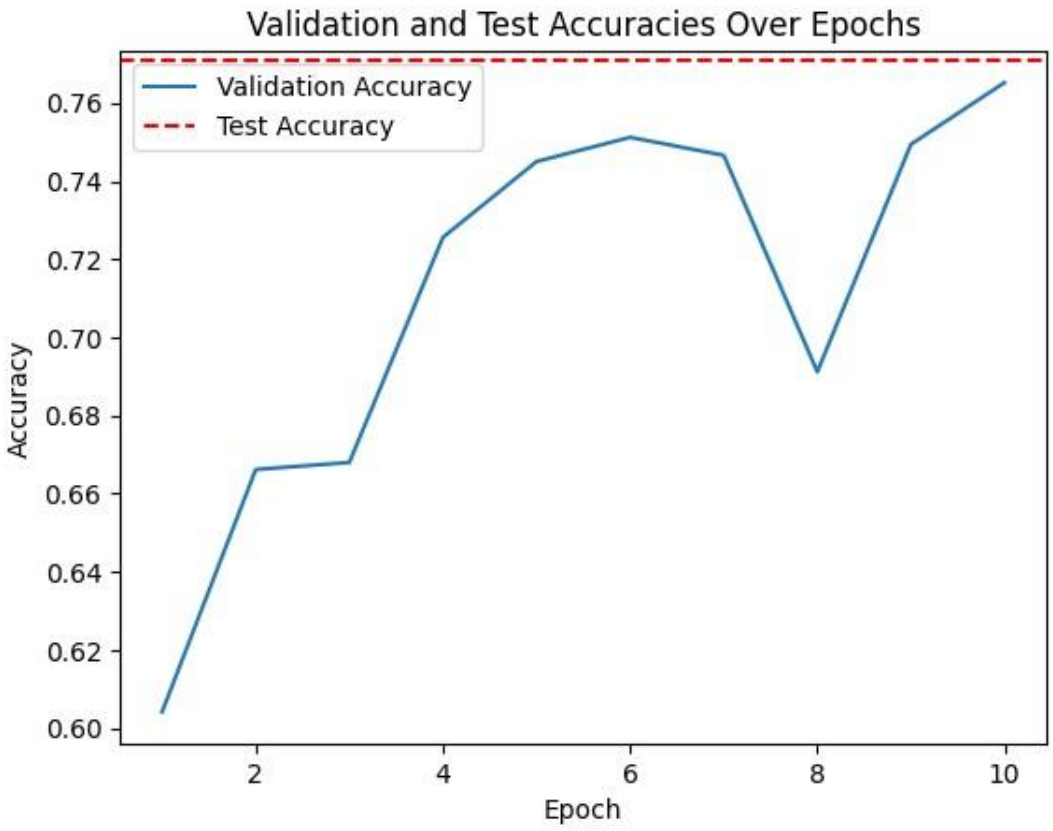
2) The best results weren't achieved when the training samples were limited to 100. The model performed poorly, with a validation accuracy of around 50%, which is equivalent to wild guess. The model was underfit and unable to identify any significant patterns in the data due to the lack of training samples. Larger training datasets may typically result in more effective training of deep learning models, particularly for challenging applications like sentiment analysis.



3) While creating the val_ds dataset, the validation of 10,000 samples was enabled by setting validation_split=0.2 and subset="validation". Through this procedure, the original dataset was split into subsets for training and validation (the validation subset including 10,000 samples). According to the validation_split option, 20% of the data were used for training and the remaining 80% for validation. The model was trained on an important amount of data and had a big validation set. The validation process not only assisted with evaluating the model's ability to adapt to new, untested data, but also offered insights into its overall performance and potential for practical applications.



4) The top 10,000 words will be considered by setting the TextVectorization layer's max tokens = 10000 definition. This selection indicated that only 10,000 most common terms in the dataset should be considered for tokenization, with all other words to be treated as out-of-vocabulary tokens. Limiting the vocabulary to 10,000 words allowed the model to focus on the most relevant terms in the dataset, possibly improving performance by eliminating noise from fewer frequent words. This method might have increased the model's capacity for generalization, improved training efficiency, and simplified the tokenization process.



Training samples	Validation accuracy
100	77.9 %
1000	78.59%
5000	77.09%
10000	77.09%

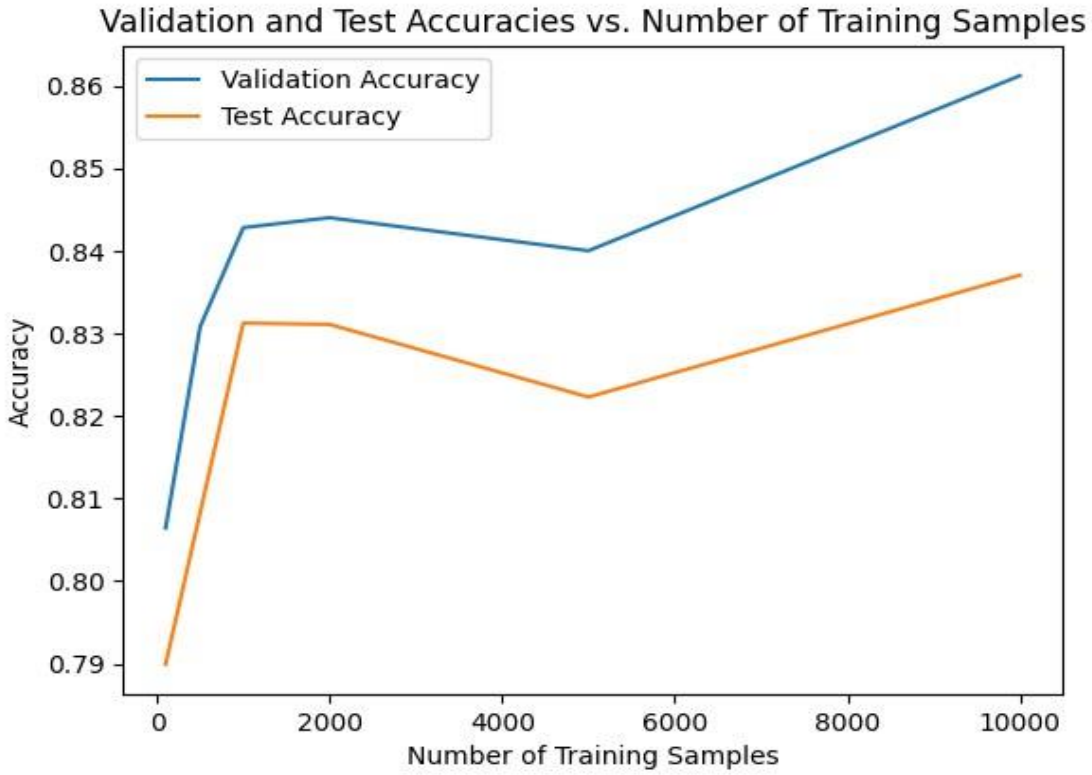
5) The model includes a bidirectional embedding layer added to it prior to the layer. This layer transforms input integer sequences into dense vectors of fixed size, allowing the model to analyze the context and meaning of words in the reviews.
Pretrained embeddings from an extensive body of text would need to be loaded and assigned as the embedding layer's initial weights, if pretrained word embeddings were to be employed. This approach can enhance the model's performance by using pretrained word representations, which frequently capture semantic relationships more correctly than embeddings learned during training.

- Both strategies have a similar general plan and method of operation. The implementation of the GloVe embeddings and the data splitting are the primary differences. Strategy 1 seems to provide an easier method for the design and execution of data separation. Approach 2 might provide greater flexibility in

loading GloVe embeddings and data splitting. The findings of sentiment analysis are enhanced by this method since it makes it possible to understand word context and meaning more precisely.

- To identify the point at which the embedding layer performs better, the validation accuracy of the models with varying amounts of training samples can be evaluated. The outcome shows the validation accuracy increases as the number of training samples does, indicating that the embedding layer functions better with more data. For different numbers of training samples, the validation accuracy values are as follows:

Training samples	Validation accuracy	Test accuracy
100	80.6%	79.0%
500	83.1%	80.8%
1000	84.3%	83.1%
2000	84.4%	83.1%
5000	84.4%	82.2%
10000	86.1%	83.7%



The results show that the embedding layer tends to perform better with more training data, with the validation accuracy improving up to 5000 training samples.
Subsequently, the performance gain appears to have decreased, suggesting that more training samples could not have an important effect on the embedding layer's performance.

Conclusion:

In conclusion, I used a variety of techniques and adjustments for performing sentiment analysis on the IMDb movie review dataset during this assignment. However, Pretrained Word Embedding showed potential for improvement with a greater validation set and pretrained embeddings, Embedding layer displayed reasonable accuracy despite its limitations. Based on the study, it is suggested to use pretrained word embedding with a larger training dataset for better performance. This method's accuracy in understanding word significance and context improves sentiment analysis results.