Assignment 4 Summary

Using the IMDb dataset, word embedding techniques for sentiment analysis are the main focus of this study. Half of the 50,000 movie reviews in the IMDB dataset are good, and the other half are negative. Twenty-five thousand training reviews and twenty-five thousand testing reviews make up the dataset. We limited training samples to 100 and terminated reviewing after 150 words. We only considered the top 10,000 phrases after verifying 10,000 samples. The overall goal of the paper is to assess how well various models perform when given varying training samples and embedding layers. A bidirectional LSTM architecture was used to train the models in this work.

- We started by training a simple sequence model to provide a performance baseline for the task. Although the model's validation accuracy was somewhat lower than its training accuracy, this suggests that the model overfitted to the validation set.
  Then, we created a model from scratch using word

- Then, without turning on masking, we created a new model and trained it using word embedding. As a result of overfitting to the training set, the model had greater training accuracy but lower validation accuracy. In addition, we found that turning on masking can improve the model's ability to handle variable-length sequences by decreasing overfitting.

- In terms of validation accuracy, we trained a model with masking enabled, and it performed better than the previous model. It follows that masking is a crucial factor to take into account while using word embedding.
- The pre-trained model did not appear to have grasped the nuances and context of the dataset, as evidenced by the fact that the training accuracy of the model we used to train it was worse than that of any previous model. In order to achieve better performance, this highlights how important it is to experiment with different embeddings or adjust the pre-trained model.

Ultimately, we conducted experiments with various sizes of training samples in order to determine the optimal size for embedding layer training. It was shown that the model performed best with 1000 training samples, maintaining low training and validation loss while offering excellent training and validation accuracy.

Model	Train Accuracy (%)	Test Accuracy (%)
Basic Sequence	96.78	80.6
Embedding layer from Scratch	98.72	78.2
Embedding layer from Scratch with Musking	98.84	80.8
Pretrained word Embedding	81.3	77.2
1000 Training samples	98.58	83.2
5000 Training samples	98.87	82.8
10000 Training samples	98.84	83
15000 Training samples	98.72	83.1
20000 Training samples	98.2	83.2
25000 Training samples	98.60	83.1

With a test accuracy of 83.2, "1000 Training samples " is the model with the best accuracy, according to the data.
Finally, we discovered that utilizing a higher training sample size (10,000) produced the best results for all models.

Conclusion:

Our conclusion from the results is that the model's training accuracy can be raised by creating word embeddings from scratch. An appropriate regularization strategy, like masking, should be employed to prevent overfitting. Embedding with different embeddings or fine-tuning the embedding for better performance is important because pre-trained models may not always perform well on datasets. An embedding layer's performance is dependent on the amount of its training samples. The size of the training dataset, regularization strategies, and pretrained models are some of the variables that influence the efficacy of word embedding, a crucial natural language processing technique. This research provided a comprehensive understanding of word embedding methods and how sentiment analysis uses them.