Assignment 3

Training Models Across Different Sample Sizes:

The values are set to:

Cutoff reviews set to 150 words.

training samples = 100

Validate samples = 10,000

words= 10,000

The models have been trained using different training sample sizes from 100 to 10000 and their test loss and accuracy are recorded in below table,

sample size	one hot encoded sequence		Embedded		Embedded masked		pre trained	
	Test Loss	Test Accuracy	Test Loss	Test Accuracy	Test Loss	Test Accuracy	Test Loss	Test Accuracy
100	0.6218	0.6629	0.6708	0.5857	0.6586	0.608	0.6787	0.6134
500	0.697	0.565	0.7138	0.6067	0.7351	0.6192	0.6193	0.6696
2000	0.6596	0.5992	0.7226	0.7108	0.8353	0.7032	0.5391	0.7248
5000	0.4891	0.7961	0.5375	0.7924	0.765	0.7645	0.5137	0.7836
10000	0.4380	0.801	0.4455	0.798	0.4349	0.811	0.4573	0.783

Train sample 100, Validation 10000:

- Initial Setup:

1. IMDB review dataset has been imported for this assignment.
2. The initial setup for the model was taking 100 training samples with each review of length 150 words max and a total of 10000 words are taken as input for the model.
3. Also, this model is validated against 10000 validation samples of both positive and negative reviews.
4. The loss function "binary cross-entropy" was used as it was a classification model with optimizer "Adam".

- Models Trained:

1. There are four models trained, validated, and tested using the initial setup with performance metrics as accuracy.
2. One hot-encoded sequence model has achieved a test accuracy of 0.801 and a testloss of 0.4380.
3. The embedded model without masking gave a test loss of 0.798 and a test accuracy of 0.4455.
4. An embedded model with masking gave a test loss of 0.811 and a test accuracy of 0.4349.
5. A pre-trained model Global Vectors for word representation (GloVe) gave a test

loss of 0.4573 and test accuracy of 0.783.

The results of the analysis showed that RNNs with embedded layers performed significantly better than other word embedding techniques, such as one-hot encoded sequences, in the task of sentiment analysis. The embedded layer-based models consistently outperformed other techniques in terms of both test loss and test accuracy.

Furthermore, with comparison different types of embedded layers, including standard embedded and masked embedded layers. The standard embedded layer-based models showed slightly better performance in terms of test accuracy as compared to masked embedded layers. Although the masking technique allows the model to ignore padding tokens and focus only on the actual word embeddings, leading to more meaningful representations and improved performance here in this model implementation it can be observed that there is no impact of masking on the given IMDb dataset.

CONCLUSION:

- For the embedding layer model, the validation accuracy is higher than the test accuracy for all cutoff reviews and training sample sizes. This suggests that the model might be overfitting to the training data.
- For the pre-trained model, the validation accuracy is higher than the test accuracy for some cutoff reviews and training sample sizes, but lower for others. This suggests that the performance of the model is more variable than the embedding layer model.
- The findings indicated that the straightforward embedding layer model performed better than the pre-trained model, which goes against the common assumption that pre-trained embeddings enhance model performance. Generally, it’s important to consider that the pre-trained model here is not optimized for the task specified and didn’t fine-tune the embeddings during training. Essentially, Fine-tuning the embeddings might lead to better performance.
- Finally, we should be attentive in drawing conclusions from these results as they are based on a small number of training samples and a limited set of hyperparameters. It's possible that different hyperparameters or more training data could lead to different conclusions.