Assignment 4

Introduction:

Recurrent neural networks (RNNs), a type of deep learning model, are great for sequential data processing tasks like natural language processing (NLP) tasks. Due to their outstanding ability to understand the sequential dependencies in data, RNNs do exceptionally well at tasks involving sequences of words, phrases, or paragraphs. In this study, we used RNNs with embedded layers to perform sentiment analysis on the IMDb movie review dataset.

Train sample 100, Validation 10000, Test 25000:

- **Initial Setup:**
- 1. For this task, an IMDB review dataset has been imported.
- 2. The model was initially configured to accept 100 training samples, with Each review lasting no more than 150 words, for a total of 10000 words as input.
- 3. This model has also been tested against 10000 validation samples of positive and negative evaluations.
- 4. Because it was a classification model using the optimizer "Adam," the loss function "binary cross-entropy" was applied.
- **Models Trained:**
    1. There are two models that have been trained, verified, and tested using the basic setup with accuracy as the performance parameter.
    2. The embedded model yielded a test loss of 0.6707 and a test accuracy of 0.5857 without masking.
    3. A pre-trained Global Vectors for Word Representation (Glove) model produced a test loss of 0.6787 and a test accuracy of 0.6134.

Analysis:

Models Across Different Sample Sizes:

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	**Embedded**		**pre trained**	
	TestLoss	**Test Accuracy**	**TestLoss**	**Test Accuracy**
100	0.6707	0.5857	0.6787	0.6134
500	0.7138	0.6067	0.6193	0.6696
2000	0.7226	0.7108	0.5391	0.7248
5000	0.5375	0.7924	0.5137	0.7836
10000	0.5413	0.8024	0.4452	0.8065

Summary:

The investigation's findings demonstrated that when it came to sentiment analysis, RNNs with embedded layers greatly outperformed alternative word embedding methods like In terms of test loss and test accuracy, embedded layer-based models consistently outperformed competing strategies.

Additionally, it was shown that as the sample size was increased, the performance of the RNN-based models improved. The test accuracy of the RNN-based models grew from about 60% to over 80% as the sample size increased from 100 to 10,000 samples. This shows that bigger sample sizes provide the model access to more training data, which improves performance.

Also included in the comparison are standard embedded and masked embedded layers, among other types of embedded layers. Using pre-trained word embeddings, specifically Glove embeddings, produced more efficient and effective models than training embedded layers from scratch, according to the IMDb movie review dataset. When trained on 10,000 samples, the pre-trained model outperformed both masked and conventional embedded layers in terms of test accuracy, with a test accuracy of 0.8065.

Conclusion:

Based on the results and analysis of the experiment, it is possible to infer that bigger sample sizes often result in greater performance. As the sample size grows, the model has more data from which to learn and is more likely to generalize to previously unknown data.

These sample sizes consistently outperformed other embedding approaches, including typical embedded layers and pre-trained Glove embeddings.

Furthermore, the performance of the pre-trained Glove embeddings was found to be consistently superior across different sample sizes, with higher accuracy and lower loss when compared to alternative embedding approaches. This implies that the advantages of pre-trained Glove embeddings are especially apparent when the available training data is restricted, since they offer a strong initialization for the model and use the information obtained from massive quantities of data during pre-training.