AML ASSIGNMENT 4

PURPOSE :

The purpose of this assignment is to apply RNNs or transformers to text and sequence data. Here we are considering the classification on the IMDB example to predict the movie reviews. The dataset comprises 50,000 reviews, of which we consider just the top 10,000 words, limit training samples to 100 first, and then change the values to determine model performance.

The overall aim is to compare the effectiveness of the models with different training samples and embedding layers.

APPROACH:

For this dataset, we used two methods for generating word embeddings for the data: pre-trained word embedding based on GloVe model and embedding layer. Two different layers: one with a custom-trained embedding layer and the other with a pre-trained word embedding layer to determine the model efficiency.

A range of training sample sizes (100, 500, 1000, and 10,000) were used to assess the accuracy of the two models. Initially, we created a custom-trained embedding layer using the IMDB review dataset. After training each model on different dataset samples, we measured its accuracy using a testing set. Next, we compared these precisions with a model that had a word embedding layer that had been trained beforehand and tested on different sample sizes.

RESULTS:

MODEL	TRAINING SAMPLES SIZE	TRAINING LOSS	VALID LOSS	TRAINING ACCURACY	VALIDATION ACCURACY	TEST ACCURACY
Basic Sequence model	100	0.85	57.5	97.5	77.8	80.6
Embedding layer from scratch	100	0.24	72.1	99.4	76.2	78.4
Embedding	500	0.13	84.8	99.6	81.1	80.6

layer from scratch						
Embedding layer from scratch	1000	0.08	94.2	99.8	81.2	80.6
Embedding layer from scratch	10000	0.08	1.06	99.7	79.9	81.1
Pre-trained word embedding	100	43.7	49.4	80.6	76.3	78.2
Pre-trained word embedding	500	0.08	1.12	99.6	80.8	79.4
Pre-trained word embedding	1000	0.12	88.3	99.7	80.8	79.5
Pre-trained word embedding	10000	0.11	97.2	99.7	79.7	80.7

Based on the data, depending on the training sample size the embedding layer from scratch achieved test accuracy from 78.4% to 81.1% and the pre-trained word embedding layer achieved accuracy from 78.2% to 80.7%. The “10,000” samples model performed well with high test accuracy.

Because of these results, it is difficult to determine which approach is the "best" to adopt because it relies on the demands and constraints of the particular work. However, in this study, the Embedding layer from scratch outperformed overall pre-trained word embedding, particularly when training with larger sample sizes. If computing resources are few and a modest training sample size is needed, the Embedding layer from scratch is a preferable choice, as long as overfitting is avoided. Under this situation, the "10,000 Training samples" model could be a good substitute if time and computational resources are constrained.