ASSIGNMENT 4

TEXT AND SEQUENCE

INTRODUTION:

The dataset contains 50000 reviews, with reviews cut off at 150 words, training samples limited to 100, 1000,10000,15000 and 30000 validations on 10000 samples, and only the top 10000 words evaluated. The data undergoes pre-processing. Later, we input data to the embedding layer as well as a pre-trained embedding model and evaluate performance by experimenting with various ways.

OBJECTIVE:

The IMBD dataset is a binary classification problem in which the goal is to predict whether a move review is positive or negative and which approach performed best.

METHODOLOGY:

There are two possibilities for creating word embedding for the IMDB review Dataset.

- A Customed-Training embedding Layer
- A pre- Trained word embedding Layer using The GloVe Model.

A GloVe is an approach for generating word vector representations that use unsupervised learning. The representations are trained using a corpus of aggregated global word-word co-occurrence information, and the resulting representations highlight fascinating linear substructures of the word vector space.

In this study, I applied the GloVe Model 6B containing 100-dimensional embedding vectors for 400,000 words (or non-word tokens).

CUSTOMED – TRAINING EMBEDDING LAYERS:

MODELS	TRAINING	TEST LOSS	TEST ACCURACY
1	100	0.323	0.756
2	1000	0.658	0.685
3	10000	0.455	0.753
4	15000	0.345	0.789
5	30000	0.323	0.723

PRE- TRAINED WORD EMBEDDING LAYER (GLOVE MODEL):

MODELS	TRAINING	TEST LOSS	TEST ACCURACY
1	100	0.331	0.756
2	1000	0.665	0.834
3	10000	0.356	0.814
4	15000	0.413	0.805
5	30000	0.432	0.867

RESULTS:

Both models using an embedding layer and a pre-trained word embedding accomplished roughly the identical test accuracy with cutoff reviews after 150 words, restricting the training sample to 100, validating on 10,000 samples, and considering just the topmost 10,000 words.

CUSTOMED – TRAINING EMBEDDING LAYERS**:**

- The accuracy obtained using the custom- trained layer ranged between 68.5% to 78.9% using different Training sample sizes (100,1000,10000,15000 &30000).
- Model 3 with the Training sample size of 10000 has obtained the highest accuracy among the all the five Models.
- The accuracy did not continue to improve significantly beyond the model 3 with Training sample size of 10000.

PRE- TRAINED WORD EMBEDDING LAYER (GLOVE MODEL):

- The accuracy obtained using the pre-Trained word embedding layer ranged between 75.6% to 86.7% using different Training sample sizes (100,1000,10000,15000 &30000).
- The Model 5 with the training sample size of 30000 has obtained the highest accuracy among all the five models.

CONCLUSION:

Overall, the results above suggest that the effectiveness of the model can be highly dependent on the specific parameters used, like size of training samples, the word embedding used, the maximum review duration, and other factors. When dealing with smaller datasets, using an embedding layer may be more advantageous.