**Assignment 4: Text and Sequence Data**

**Objective**: To get a better understanding of how different parameters impact the model's performance and to discover which variants operate best under certain scenarios. First, we will see how the model is performing when reviews are limited to 150 words. Then, we test the model's generalization capabilities by limiting the training dataset to only 100 samples and then we validate its performance using a bigger validation set of 10,000 data. Apart from that we will analyse the effect of limiting the vocabulary by considering only the top 10,000 words and compare the performance using two different approaches, an embedding layer, and a pretrained word embedding.

**Metrics:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Sample Size** | **Training Accuracy** | **Training Loss** | **Validation Loss** | **Validation Accuracy** | **Test Accuracy** |
| Model 1(Scratch) | - | 0.9657 | 0.1067 | 0.3329 | 0.8726 | 0.866 |
| Model 2(Scratch) | 100 | 0.9750 | 0.5340 | 0.6963 | 0.5500 | 0.499 |
| Model 3(Pre-trained model | 100 | 0. 0.6700 | 0.5856 | 0.7010 | 0.5299 | 0.5262 |
| Model 4(Embedding layer | 1000 | 0.9762 | 0.3939 | 0.6560 | 0.6400 | 0.598 |
| Model 5(embedding layer and Conv1D) | 5000 | 0.9775 | 0.0561 | 0.4896 | 0.7750 | 0.772 |
| Model 6(embedding layer and Conv1D) | 10000 | 0.9638 | 0.1004 | 0.4238 | 0.8100 | 0.796 |
| Model 7(Pretrained model) | 15000 | 0.8817 | 0.2810 | 0.3421 | 0.8483 | 0.484 |

1. Here we have developed 7 different model for a specific task. The models are different in terms of architecture, training size, and the usage of pre-trained embeddings. The performance metrics are compared to understand the strength of the model.
2. Model 1 has the highest test accuracy of 86% because it has the lowest validation loss and the highest validation accuracy among all the models. It means that Model 1 accurately classify reviews as either positive or negative and generalizes well to unseen data. But when we trained the model with a sample size of 100 the performance of the model has decreased with a test accuracy of 49%.
3. Models that use embedding layers and CNN have higher test accuracy compared to other models, it means embedding layers and CNN are good in capturing patterns in data and text classification.
4. Model 7 which is a pre-trained model and a huge sample size of 15000 produced poor results. While it achieved a training accuracy of 88%, it achieved a lower test accuracy of 48%. It means that the model is not aligned with the specific task and require fine tuning. Model 2 which is trained from scratch has shown limited generalization on both the validation and test sets, with the lowest test accuracy of 49%.
5. Model 5 with both embedding layer and Conv1D which has a sample size of 5000 has good training accuracy of 97% showing efficient learning during training. It also achieved strong performance on test set with the test accuracy of 77%.
6. The models with embedding layer and Conv1d outperformed the pre-trained word embedding test accuracies. By including embedding layer and Conv1D in model 5 and 6 with training samples of 5000 and 10000 has significantly enhanced model performance. Model 5 and 6 trained with 5000 and 10000 samples achieved test accuracy of 77%, and 79%.