**Advance Machine Learning**

**Assignment – 4**

**Text and Sequence Data**

**Report**

**Group 24:**

**Shiva Chaitanya Goud Gadila: 811252042**

**Praneeth Simha Mallenahalli Vijaya: 811252718**

**Introduction:**

The fundamental elements of many real-world applications like social media sentiment analysis, email spam detection, language translation and many more depend on the text and sequence data. To get useful insights and forecasts one must be able to model and analyze these kinds of data efficiently.

In this assignment our main objective is to explore the performance of different models trained on text and sequence data in various settings. We aim to investigate how factors such as a data set size and the choice of word embedding techniques impact the performance of these models.

The model’s versions work best in various settings and fully comprehend how various factors affect the model’s performance. Restricting the size of the training sets, considering the only changes were made to consider the top 10,000 words and compare the performance of a model using an embedding layer with that of a model using pre trained word embedding.

**Scratch Model’s Results**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Training**  **Sample Size** | **Loss** | **Accuracy** |
| **1** | **-** | **0.34** | **0.86** |
| **2** | **100** | **0.69** | **0.50** |

**Pre-Trained Model’s Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Training**  **Sample Size** | **Loss** | **Accuracy** |
| **1** | **100** | **0.71** | **0.51** |
| **2** | **15000** | **1.01** | **0.50** |
| **3** | **30000** | **1.27** | **0.48** |

**Embedding layer 1 and Embedding layer and conv1D Model’s Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Training**  **Sample Size** | **Loss** | **Accuracy** |
| **1** | **100** | **0.66** | **0.59** |
| **2** | **15000** | **0.43** | **0.79** |
| **3** | **30000** | **0.39** | **0.82** |

**Results:**

* After comparing the two scratch models, we found that model 1 had an accuracy of 86% whereas model 2 had an accuracy of just 50%. Both models remained unmodified.
* It suggests that it may accurately categorize the evaluations as either favorable or negative based on their wording. However, when the train evaluations were restricted to 100 samples, the model’s performance substantially decreased leading to a test accuracy of only 50%.
* When just considering the top 10,000 words the models with an embedding layer and a pre trained word embedding produced identical test results.
* The training sample size was changed to test if the embedding layer could surpass the word embedding that had been previously taught. In contrast to the pre trained word embedding at training sample 100, which only managed a test accuracy of 0.51 with 15,000 training samples, the embedding layer performed better with an acquisition of 0.79 test accuracy.
* Since an increased training sample size had very little effect on accuracy, we used conv1D in conjunction with the embedding layers and increased the training sample size to 15000 and 30000. It significantly increased test accuracy, which increased to 79% and 82%.
* When utilizing 15000 and 30000 training sample sizes, the models with the embedding layer and Conv1D outperformed the pre-trained word embedding test accuracies.

**Conclusion:**

These results directly address the introduction's goals and offer insightful information about how several elements, including model architecture, training set size, and embedding methods, impact the effectiveness of models trained on text and sequence data.

With an accuracy of 82%, the model 3 that employed the embedding layer and Conv1D performed best. In conclusion, these results suggest that the specific parameters used, such as the amount of training data, the word embedding, the maximum review period, and more, may have a substantial impact on the model’s performance. An embedding layer may be more effective when working with smaller datasets.