**Assignment 4: Text Data**

**Introduction:**

The IMDb movie review dataset has been analyzed for sentiment using RNNs with embedded layers in this study, making use of their prowess with sequential data processing. Assessing the psychological tone of material, like movie reviews, involves sentiment analysis. RNN-embedded layers learn to represent words and other discrete input data as continuous vectors to capture their semantic value. The IMDb dataset, consisting of critiques of movies and sentiment labels, was used to train the RNNs. To identify long-range dependencies, the RNNs modified their hidden state variables at each successive step during training. On a different test set, the efficiency of the RNNs was assessed, and it is likely that the hyperparameters were modified for the best performance. The study's conclusions may have consequences for applications like social media, market research, and movie recommendation systems.

**Dataset and Initial Setup:**

For this work, the IMDb review dataset was imported.

The model was initially set up to receive 100 samples for training, with a 150-word maximum for each review, for a total input of 10,000 words.

10,000 validation samples of both good and negative reviews were used to evaluate the model.

For binary classification, the algorithm used the "Adam," the optimizer, and the "binary cross-entropy" loss function.

**Models Trained and Performance Analysis:**

With accuracy serving as the performance metric, two models were trained, verified, and evaluated using the aforementioned methodology.

With no masking, the embedded model achieved a test loss of 0.6707 and a test accuracy of 0.5857.

Test accuracy was 0.6134 and test loss was 0.6787 for a pre-trained Global Vectors for Word Representation (GloVe) model.

**Performance across Different Sample Sizes:**

The models' test loss and accuracy are reported in the table below. The models were trained using various sample sizes ranging from 100 to 10,000.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SAMPLE SIZE** | **EMBEDDED MODEL  TEST LOSS** | **EMBEDDED MODEL  TEST ACCURACY** | **GLOVE MODEL  TEST LOSS** | **GLOVE MODEL 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 results of this study indicate that for sentiment analysis tasks, RNNs with embedded layers perform better than alternative word embedding techniques.

Additionally, the performance of the RNN-based models improved steadily as the sample size increased, with test accuracy rising from about 60% to over 80% as the sample size rose from 100 to 10,000 samples. This shows that more training data is available from higher sample sizes, which enhances model performance.

Additionally, across a range of sample sizes, pre-trained GloVe embeddings consistently outperformed alternative embedding strategies, including normally embedded layers. The GloVe model outperformed alternative embedding techniques in terms of accuracy and loss, demonstrating the advantages of using pre-trained embeddings, particularly when training data is few.

**Conclusion:**

According to the experiment's findings, greater model performance in sentiment analysis tasks is typically correlated **with bigger sample sizes**. GloVe embeddings that have already been trained are especially efficient and perform better even with little training data. GloVe embeddings' capacity to draw knowledge from sizable external corpus aid in reducing the drawbacks of sparse training data. Pre-trained GloVe embeddings can be a potent tool for enhancing sentiment analysis model performance. The results of this study have practical ramifications for sentiment analysis's real-world applications in market research, consumer feedback analysis, and social media monitoring.