**ASSIGNMENT 4**

**TEXT AND SEQUENCE**

The objective of this experiment is to assess the performance of two different approaches—utilizing an embedding layer and a pre-trained word embedding—on the binary classification of movie reviews as positive or negative. Specifically, the experiment aims to:

1. Truncate reviews to a maximum of 150 words.
2. Restrict the training sample size to 100.
3. Conduct validation on a consistent set of 10,000 samples.
4. Consider only the top 10,000 words for analysis.

The primary goal is to compare the accuracy of both approaches across varying training sample sizes, ranging from 100 to a larger number, and determine at what point the embedding layer yields better performance compared to the pre-trained word embedding.

**Two Approaches:**

1. Embedding Layer
2. Pre-Trained Word Embedding

**With Different Sample Sizes:**

1. **Training Sample: 100**

**Embedding layer**

Test Embedding Accuracy: 0.49404001235961914

A graph of a graph with text

Description automatically generated with medium confidence A graph of a training and validation loss

Description automatically generated

**Pre-Trained Word Embedding**

Accuracy of Pretrained Model:0.498879998922348

A graph of a graph with text

Description automatically generated with medium confidence A graph of a graph with blue lines

Description automatically generated

In this specific scenario, the approach utilizing a pre-trained word embedding yielded slightly better performance. However, the difference in accuracy between the two approaches is relatively small, so other factors such as model complexity, computational resources, and specific task requirements should also be considered when choosing between these approaches.

Similarly accuracy for different training sample size is shown below for two different approaches

Results:

|  |  |  |
| --- | --- | --- |
| **Technique** | **Training Sample Size** | **Test Accuracy** |
| Embedding layer | 100 | 0.494 |
|  | 400 | 0.519 |
| 800 | 0. 552 |
| 5000 | 0.82 |
| 1000 | 0.85 |
| Pre-trained word | 100 | 0.498 |
|  | 400 | 0.502 |
| 800 | 0.502 |
| 5000 | 0.509 |
| 1000 | 0.5 |

Graph:

A graph with numbers and red dots

Description automatically generated

From the analysis

For Embedding Models: Performance increases with the increase in training sample size, with the highest accuracy achieved with an training sample size of 10000.

For Pre-Trained Models:

Accuracy remains relatively stable across different pre-trained model sizes, with no significant improvement observed

Embedding Models vs. Pre-Trained Models:

Embedding models generally outperform pre-trained models across all training sample sizes tested. Embedding models show a more significant improvement in accuracy with larger training sample sizes compared to pre-trained models.

**Conclusion:**

In summary, when considering accuracy and performance, embedding models, particularly those with larger sizes, outperform pre-trained models for the given task. However, it's essential to balance computational resources with performance gains, especially when working with very large training sample sizes.