**TEXT AND SEQUENCE**

**FALL-2023 Rajya Shree Deshmukh**

* **Executive Summary**

In our investigation, we examined modifications to the IMDB dataset. In order to reduce dimensionality and accelerate the training process, we truncated reviews to 150 words and constrained the training set to 100 examples. To mitigate overfitting on less common terms, our assessment involved 10,000 samples using only the top 10,000 words in the dataset. We conducted a performance comparison between an embedding layer and a pre-trained word embedding, finding that the latter yielded superior results. However, an interesting observation was that the optimal strategy depended on the size of the training set, with the embedding layer proving more effective for larger training sets.

* **Problem**

The IMDB dataset poses a binary classification challenge, and we assess the impact of modifying the input data while comparing the effectiveness of different methods for word embedding.

* **Technique**

This study employs an experimental research approach, involving updates to the IMDB dataset and the execution of trials to evaluate the effectiveness of different techniques. To address overfitting concerns, adjustments were made to the input data, such as truncating reviews after 150 words, restricting the training set to 100 samples, and focusing on the top 10,000 words in the dataset. Additionally, an investigation was conducted comparing the use of an embedding layer to a pretrained word embedding, with variations in the number of training samples to identify the optimal strategy. Subsequently, the outcomes of the experiments, including findings and conclusions, were thoroughly analyzed and subsequently published.

* **Conclusions**

Pretrained model with augmentation gives the highest accuracy and with the fine tuning using pretrained model got the accuracy of 97.6 %.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Embedding Technique | Training Sample Size | Accuracy | Embedding Technique | Training Sample Size | Accuracy |
| Custom-trained embedding layer | 100 | 98.7 | Pretrained word embedding layer | 100 | 100 |
| Custom-trained embedding layer | 500 | 98.5 | Pretrained word embedding layer | 500 | 100 |
| Custom-trained embedding layer | 1000 | 98.7 | Pretrained word embedding layer | 1000 | 100 |
| Custom-trained embedding layer | 10000 | 97.7 | Pretrained word embedding layer | 10000 | 100 |