INTRODUCTION

Each of the 50,000 reviews in the dataset has been condensed to 150 words. We investigate many training sample sizes (fixed validation set of 10,000 samples): 100, 1000, 10,000, 15,000, and 30,000. The evaluation only considers the top 10,000 words. We preprocess the data before putting it into the embedding layer. After that, we assess performance using several approaches and evaluate the outcomes.

OBJECTIVE

An exercise in binary classification using the IMDB dataset aims to identify whether a movie review conveys a good or negative opinion. Finding the most efficient method to do this categorization work is the main goal.

METHODOLOGY

There are two methods available for generating word embeddings for the IMDB review dataset:

1. Customized training embedding layer: In this method, a word embedding layer that is especially made for the IMDB dataset is trained while the model is being trained.
2. Pre-trained Word Embedding Layer with the GloVe Model: Alternatively, pre-trained word embeddings generated by the GloVe model can be utilized. This offers a ready-made set of word embeddings that have been thoroughly trained on Wikipedia and news articles.

Word vector representations are produced by the unsupervised learning method known as GloVe, or Global Vectors for Word Representation. Using a corpus of aggregated worldwide word-word co-occurrence data, these representations are learned. The produced representations thereby highlight interesting linear substructures in the word vector space.

In this study, the GloVe Model 6B—which comprises 100-dimensional embedding vectors for 400,000 words or non-word tokens—was employed.

Scratch Models:

Models	Sample Size	Loss	Accuracy
1	-	0.35	0.86
2	100	0.69	0.50

Two scratch models were developed for the classification task. After training on the complete dataset, the first model produced an accuracy of 0.86 and a loss of 0.35. The second model, on the other hand, showed a lower accuracy of 0.50 and a larger loss of 0.69 after being trained on a smaller sample set of 100.

PRE-TRAINED Models:

Models	Sample Size	Loss	Accuracy
1	100	0.73	0.51
2	15000	1.05	0.49
3	30000	1.36	0.50

In the pre-trained model's evaluation, the first model trained on a sample size of 100 demonstrated a loss of 0.73 and an accuracy of 0.51. Moving to a larger sample size of 15,000, the second model experienced a loss increase to 1.05 with a slight decrease in accuracy to 0.49. Finally, for a sample size of 30,000, the third model observed a higher loss of 1.36 while maintaining a similar accuracy of 0.50.

Word embedding with Conv1D:

The performance of models varied based on the sample size. Model 1, trained on 1000 samples, achieved a loss of 0.67 and an accuracy of 0.56. As the sample size increased to 15,000 and 30,000 for models 2 and 3, respectively, there was a notable improvement in performance, with both models showcasing lower losses (0.42 and 0.40, respectively) and higher accuracies (0.80 and 0.81, respectively).

Models	Sample Size	Loss	Accuracy
1	1000	0.67	0.56
2	15000	0.42	0.80
3	30000	0.40	0.81

RESULTS:

Both models achieved around the identification test accuracy with cutoff reviews of approximately 150 words, limiting the training sample to 100, validating on 10,000 samples, and considering just the top 10,000 words. They achieved this by utilizing an embedding layer and a pre-trained word embedding.

Conclusion:

In conclusion, exploring various approaches for binary classification of IMDB movie reviews provided valuable insights. Scratch models showcased impressive performance with the full dataset but exhibited decreased accuracy with smaller sample sizes. Pre-trained models demonstrated varying performance based on sample size, while models incorporating Conv1D layers showed improved results, particularly with larger sample sizes. Overall, the study emphasizes the significance of sample size and model architecture in achieving accurate classification of IMDB movie reviews, warranting further exploration for potential enhancements.