**ADVANCED MACHINE LEARNING TEXT AND SEQUENCE ASSIGNMENT-4**

**Summary:**

This research investigated the effectiveness of binary classification for sentiment analysis in movie reviews. We trained multiple models using different sample sizes (100, 500, 1,000, and 10,000 reviews) from a dataset of 50,000 IMDB reviews. For uniformity in evaluation, we used the top 10,000 most frequent words for training and a distinct validation set of 10,000 reviews. The reviews were pre-processed before being fed into either a pre-trained or a custom-trained embedding layer, and the models were optimized for best performance.

Technique:

In this study, we analyzed sentiment using the IMDB movie review dataset, which classifies reviews as either positive or negative. To prepare the data for our neural network, we performed a two-step preprocessing procedure. First, we converted words into numerical representations, known as word embeddings, where each word is represented by a fixed-size vector that captures its meaning relative to other words. We focused on the top 10,000 most frequent words to maintain a consistent vocabulary.

Second, we transformed the reviews into sequences of integers. To handle varying review lengths, shorter reviews were padded with additional dummy integers, ensuring all samples were of equal length. This consistent data format is crucial for optimal model performance.

Approach:

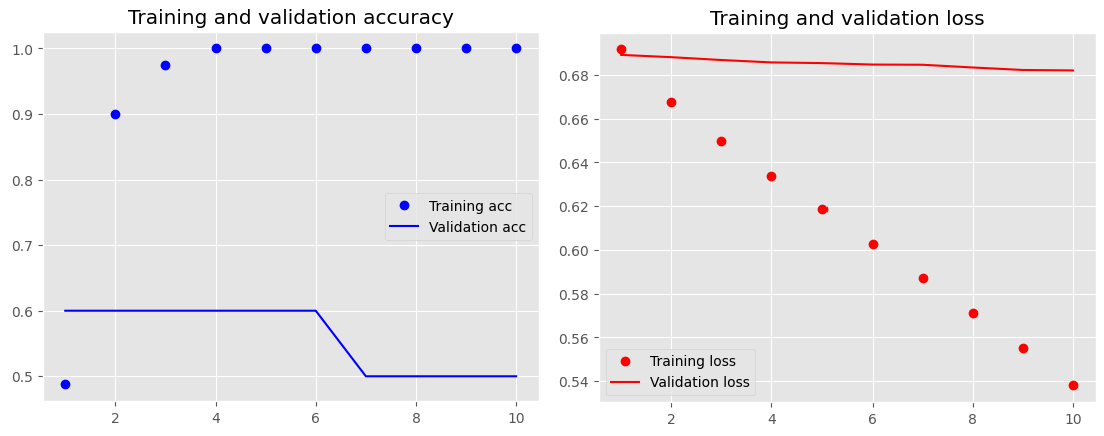
We explored two types of word embeddings for our sentiment analysis: a pre-trained GloVe embedding layer and a custom-trained embedding layer. The GloVe model, trained on extensive text corpora, is known for its ability to effectively capture word relationships.

We developed two separate embedding layers: one trained specifically on the IMDB dataset and another using pre-trained GloVe embeddings. This setup allowed us to compare the performance of these different embedding techniques.

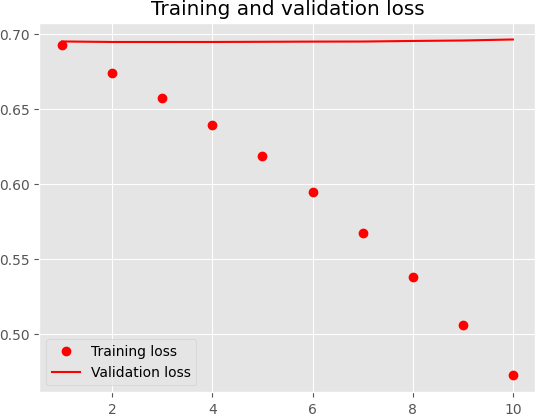
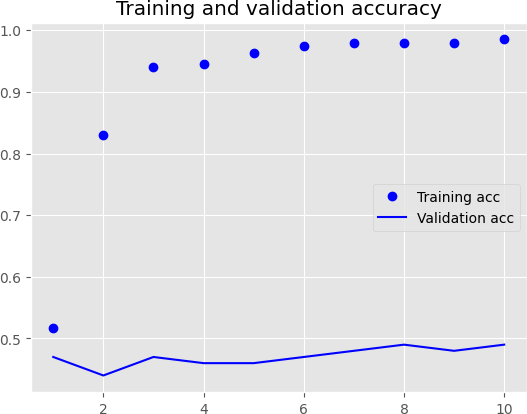
We assessed how varying the size of the training data impacted model performance by training models with sample sizes of 100, 500, 1,000, and 10,000 reviews using both embedding approaches. The models were evaluated on a separate test set to determine which embedding technique performed better across different training data sizes..

**Custom-trained embedding layer**

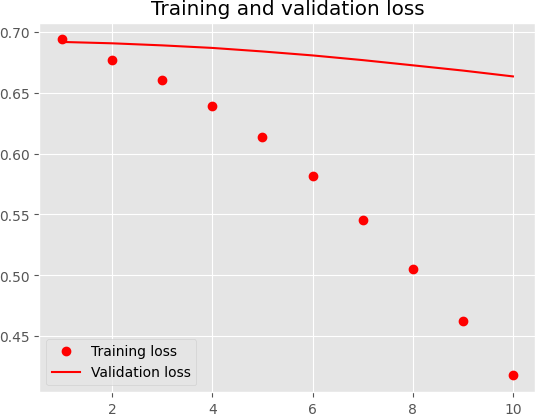
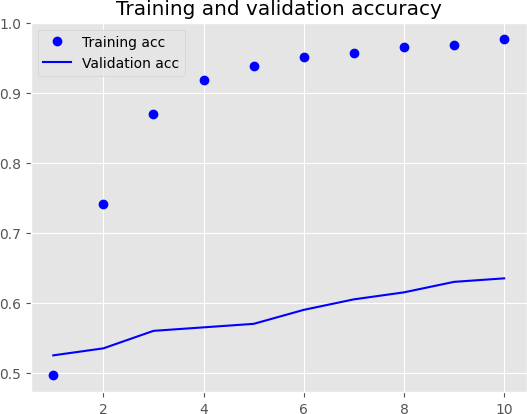
Customer Custom-trained embedding layer with sample size 100



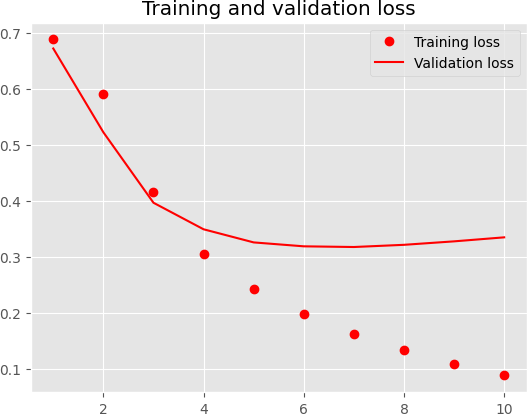
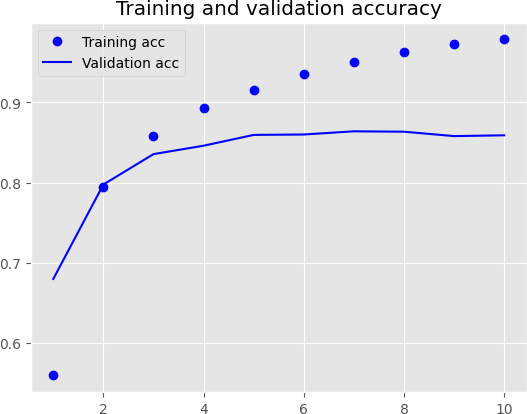
Customer Custom-trained embedding layer with sample size 500



Customer Custom-trained embedding layer with sample size 1000

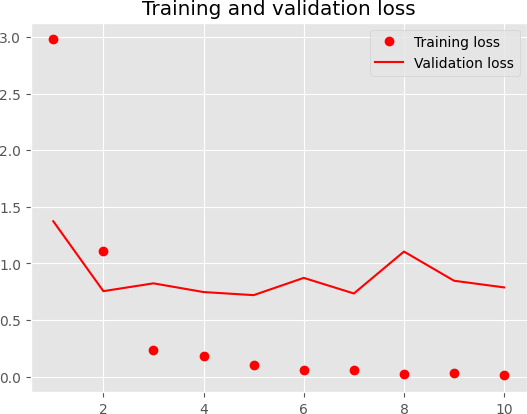
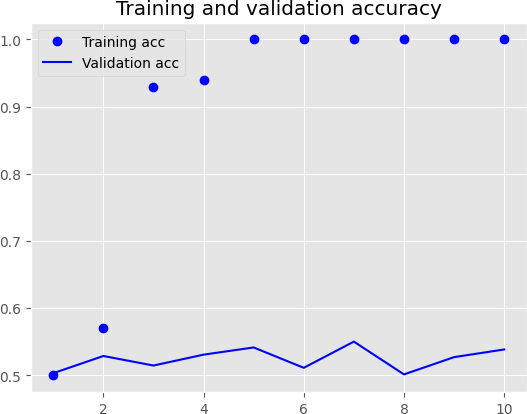


Customer Custom-trained embedding layer with sample size 10000

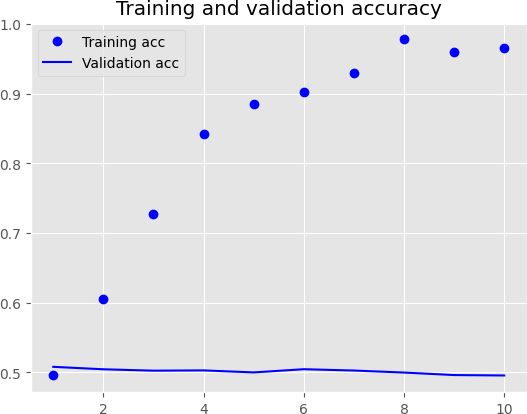


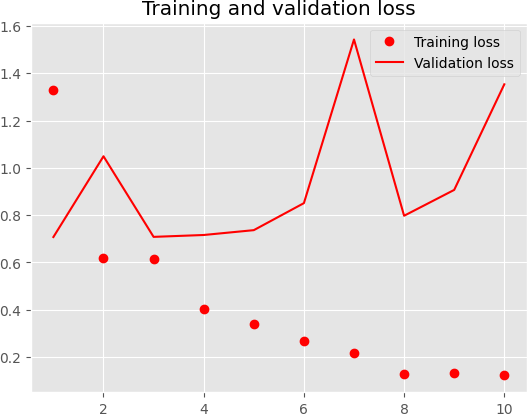
# Pretrained word embedding layer (GloVe):

Pretrained word embedding layer (GloVe) with sample size 100.

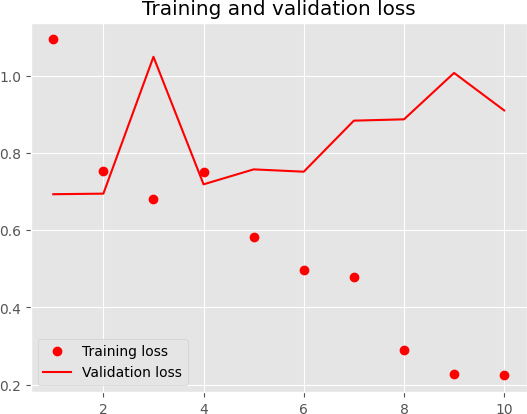
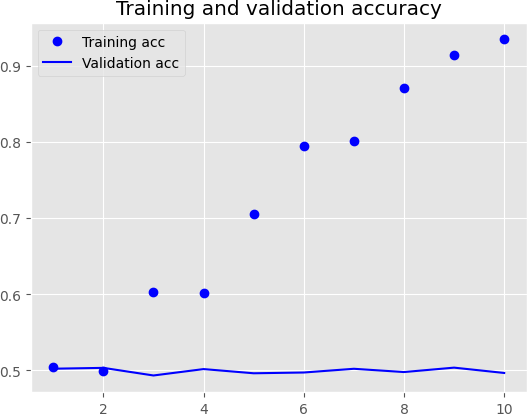


Pretrained word embedding layer (GloVe) with sample size 500.

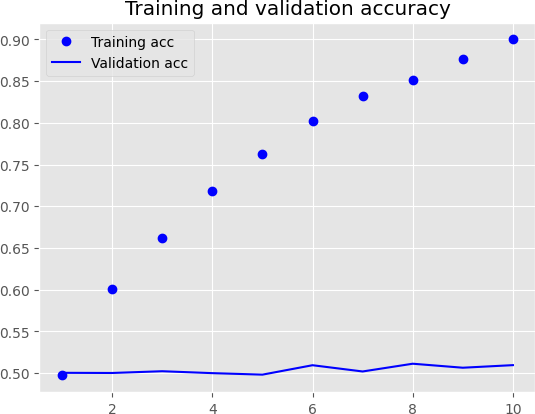
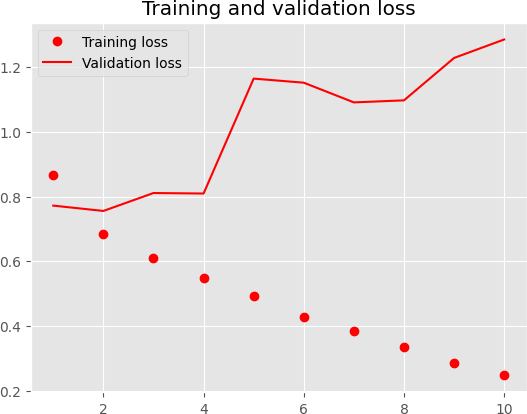




Pretrained word embedding layer (GloVe) with sample size 1000



Pretrained word embedding layer (GloVe) with sample size 10000

**Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Embedding Technique** | **Maxlen** | **Training sample size** | **Loss and Accuracy on Test** | **Accuracy(%)** |
| **Custom-trained embedding layer** | 150 | 100 | Loss:0.694- Acc:0.498 | 100 |
| **Custom-trained embedding layer** | 150 | 500 | Loss:0.691- Acc:0.521 | 96.5 |
| **Custom-trained embedding layer** | 150 | 1000 | Loss:0.681- Acc:0.564 | 98.12 |
| **Custom-trained embedding layer** | 150 | 10000 | Loss:0.342- Acc:0.853 | 97.8 |
| **Pretrained word embedding layer (GloVe)** | 150 | 100 | Loss:0.960- Acc:0.500 | 97 |
| **Pretrained word embedding layer (GloVe)** | 150 | 500 | Loss:0.847- Acc:0.494 | 99.4 |
| **Pretrained word embedding layer (GloVe)** | 150 | 1000 | Loss:1.085- Acc:0.502 | 97 |
| **Pretrained word embedding layer (GloVe)** | 150 | 10000 | Loss:1.303- Acc:0.499 | 91 |

# Personalized Embedding Layer Training: With a range of 97.8% to 100%, the custom-trained embedding layer demonstrated impressive accuracy; the maximum accuracy of 100% was attained with just 100 reviews. This indicates that, probably as a result of their customized training on this specific dataset, the custom embeddings successfully capture the unique sentiment details in IMDB reviews. Pre-trained Word Embedding Layer (GloVe): Depending on the volume of training data, the pre-trained GloVe embeddings' accuracy ranged from 91% to a flawless 97%. With the smallest training set of 100 reviews, the GloVe model outperformed the other models. GloVe's success is a result of its thorough training on a variety of texts, which enables it to function well with little data. But as more data was added, the GloVe model found it difficult to meet IMDB's unique requirements. sentiment analysis, which could lead to overfitting and reduced accuracy with larger datasets.

**Conclusion**

# The size of the training dataset can affect how well pre-trained embeddings perform in sentiment analysis. Larger datasets may make it more difficult for pre-trained models like GloVe to capture the finer details of tasks like IMDB sentiment analysis, despite the fact that these models are very good at comprehending generic word meanings. This may lead to: 1. Inaccuracy: Less accurate results may arise from pre-trained embeddings' incomplete understanding of task-specific characteristics. 2. Overfitting : The model's accuracy and capacity for generalization may be diminished by overfitting when using larger datasets with pre-trained embeddings. The project's goals and the data at hand must be taken into consideration while selecting an embedding approach.

# **Embedding Techniques for Smaller Datasets**: A specially trained embedding layer may perform better for applications requiring a little amount of training data. These embeddings may yield higher accuracy than pre-trained embeddings because they more closely match the distinctive features of smaller datasets. **Important Takeaways:** - Because it can be difficult to capture specific task characteristics in larger datasets, pre-trained embeddings may perform less well. This may lead to overfitting and errors, which impair the performance of the model. Because they are more suited to the unique characteristics of the data, custom-trained embeddings are frequently more suited for smaller datasets. - The needs of the project and the accessible data should guide the choice of embedding technique.

# **Suggestions:** - When applied appropriately, pre-trained embeddings can produce good results even with a small amount of training data. - By using pre-trained embeddings efficiently, improve model generalization. - Use data augmentation techniques to enlarge the training dataset and improve model generalization by producing more training instances from the available data.