Texts and Sequences

The study aimed to analyse sentiment on the IMDB dataset, which contains 50,000 movie reviews, using a binary classification approach. Two techniques for embedding textual data into numerical representations were compared: custom-trained embedding layers and pretrained word embeddings (GloVe). Evaluation was conducted on different training sample sizes to assess their impact on accuracy and test loss.

Data Preprocessing

Text Conversion: Reviews were transformed into numerical sequences, where each word was represented as an integer index. Padding was applied to ensure uniform sequence lengths across all samples.

Embedding Techniques: Custom-trained Embedding Layer: A layer trained specifically for the dataset. Pretrained Embedding Layer (GloVe): Used pre-trained word embeddings trained on extensive corpora.

Procedure

Custom-trained Embeddings:

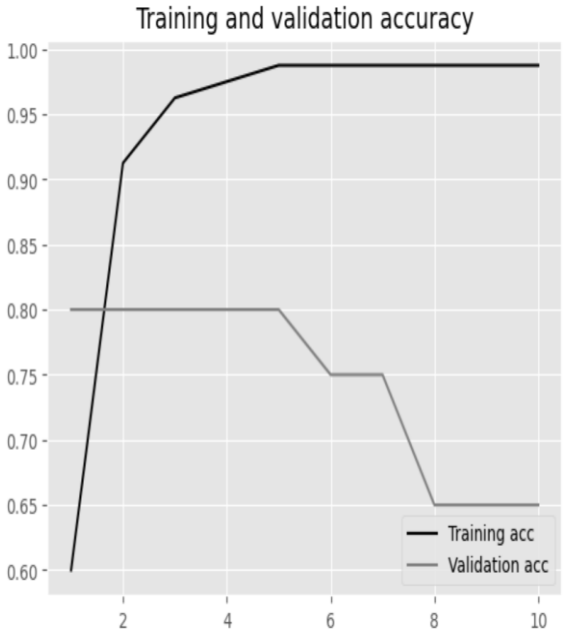
Each model was trained on subsets of the dataset (100, 1,000, 5,000, and 10,000 samples). Training accuracy and test loss were recorded after testing on a fixed validation dataset.

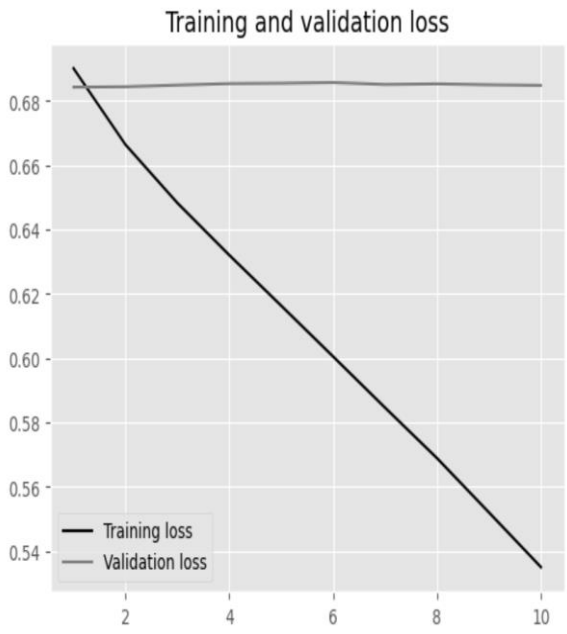
Pretrained Embeddings (GloVe):

Similar to the custom-trained setup, models were trained using the same subsets and evaluated on the validation set.

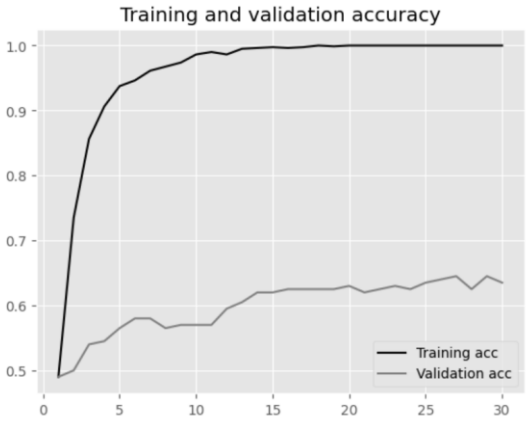
CUSTOM-TRAINED EMBEDDING LAYER

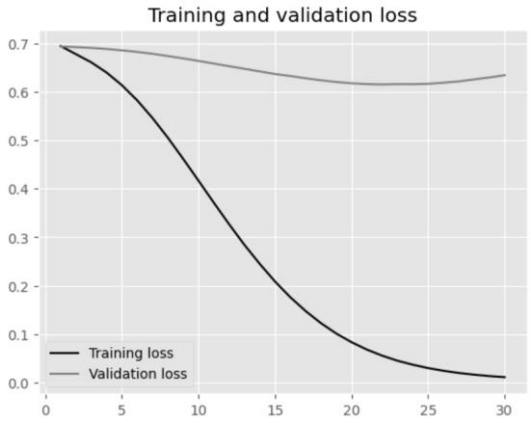
1. Custom-trained embedding layer with training sample size = 100



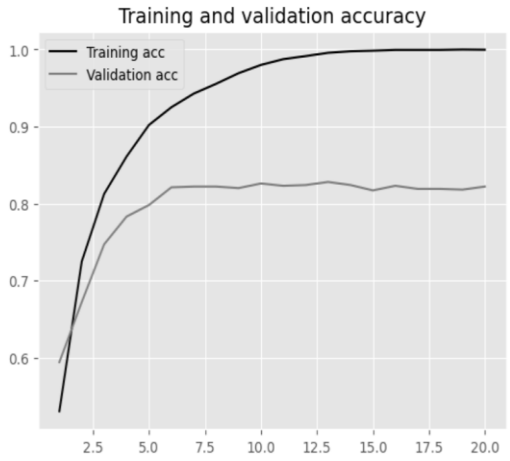


2. Custom-trained embedding layer with training sample size = 1000

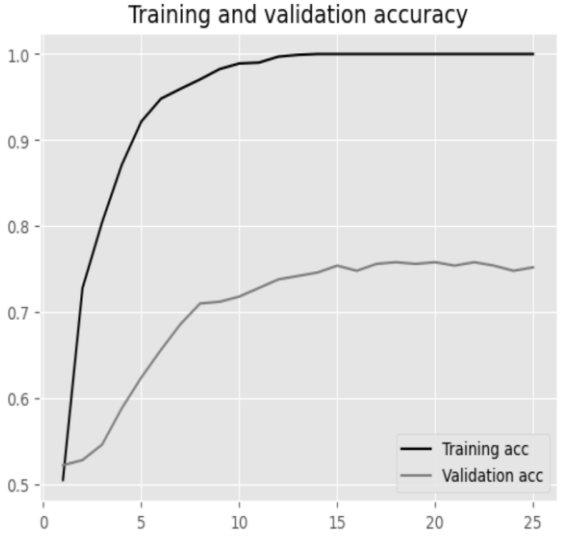




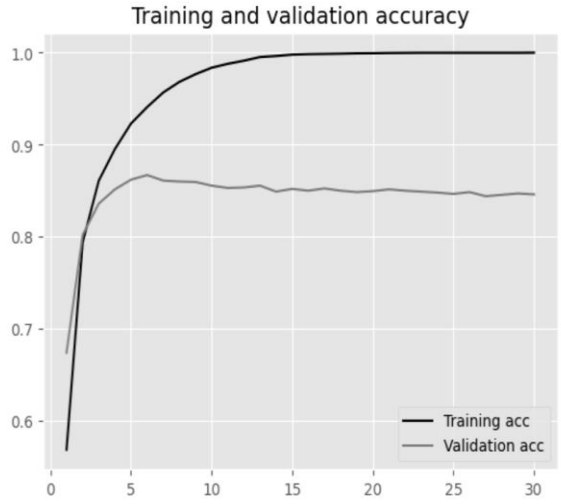
3. Custom-trained embedding layer with training sample size = 5000

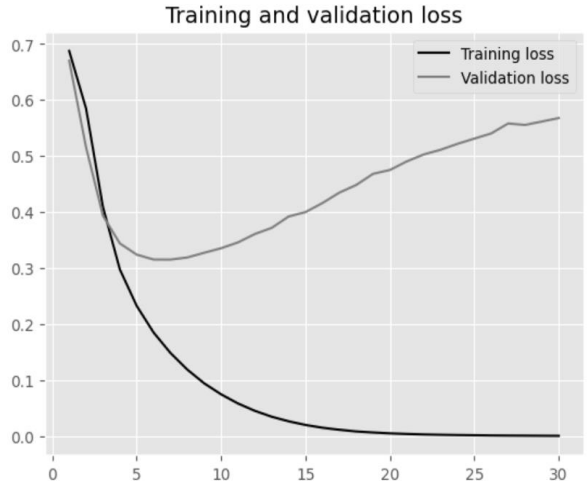


4. Custom-trained embedding layer with training sample size = 2500



5. Custom-trained embedding layer with training sample size = 10000

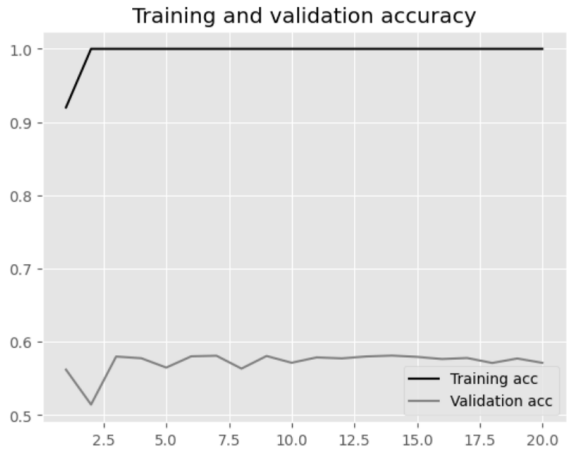


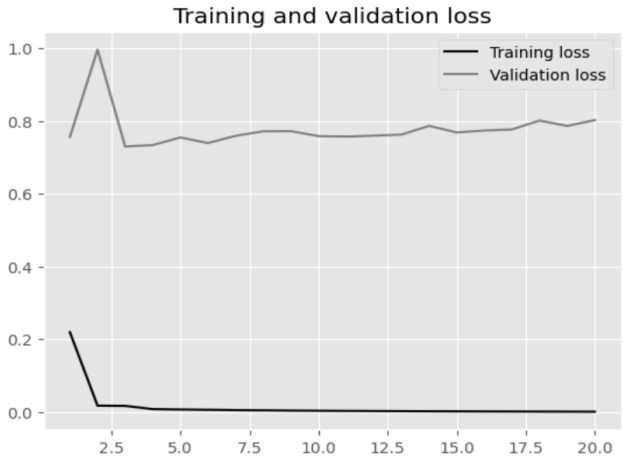


Depending on the size of the training sample, the accuracy of the custom-trained embedding layer varied from 97.3% to 100%. The training sample size of 100 produced the best accuracy.

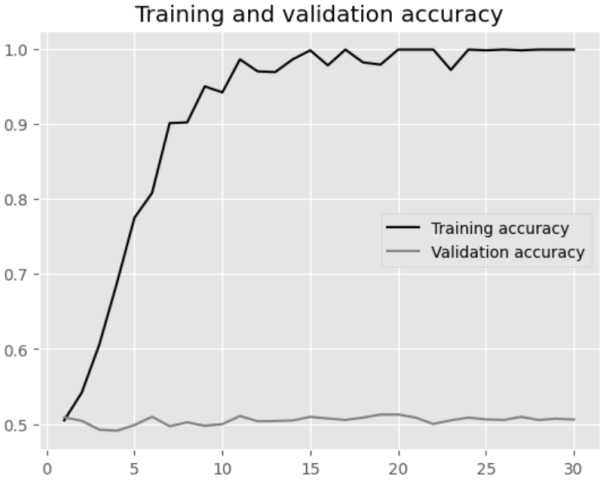
PRETRAINED WORD EMBEDDING LAYER

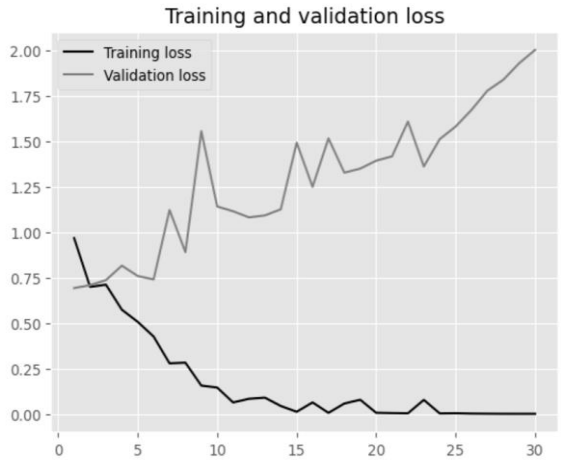
1. pretrained word embedding layer with training sample size = 100



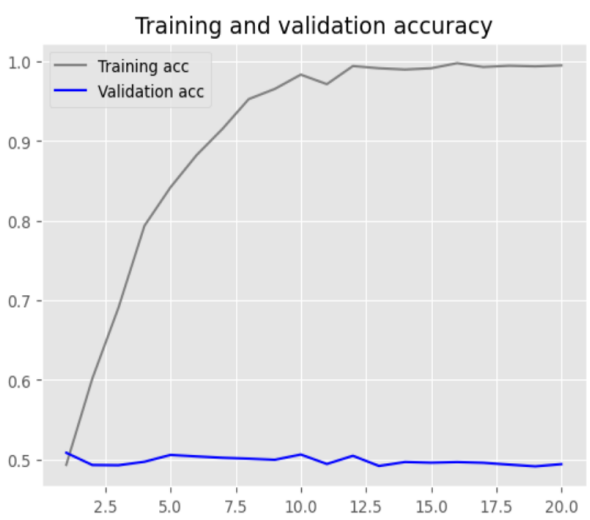


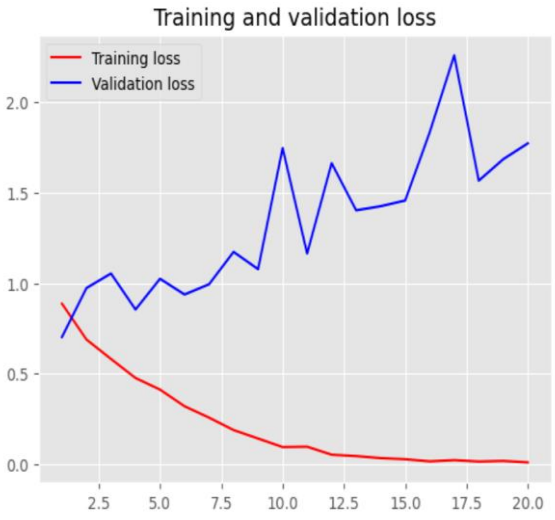
2. pretrained word embedding layer with training sample size = 1000



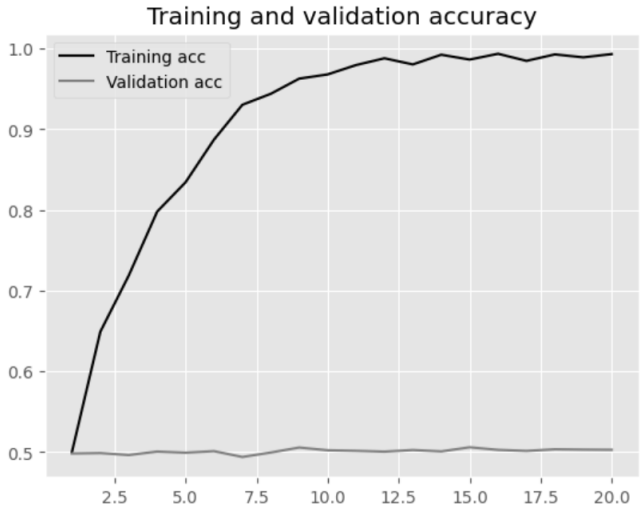


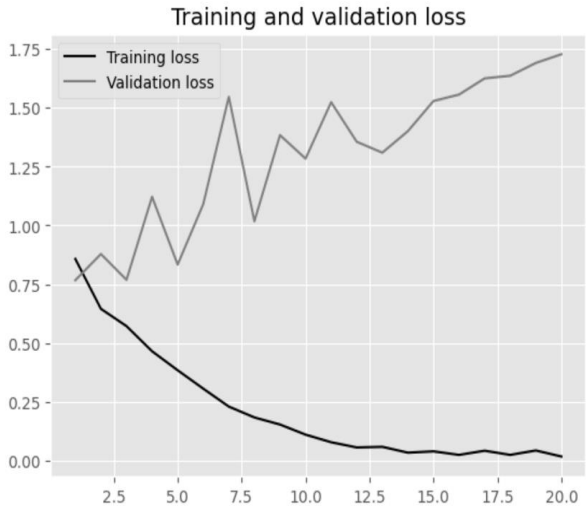
3.pretrained word embedding layer with training sample size = 5000



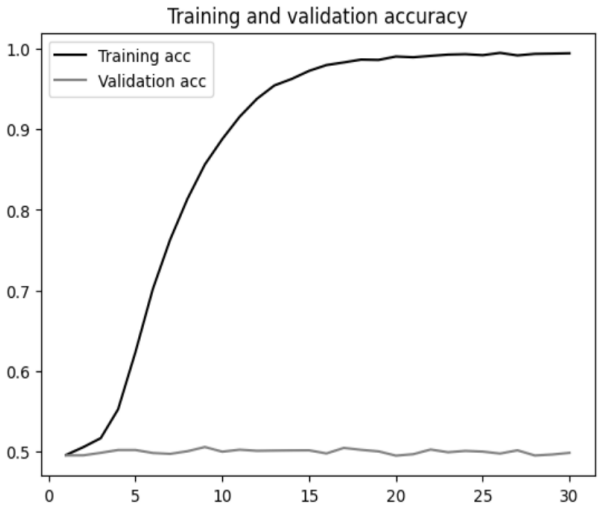


4.pretrained word embedding layer with training sample size = 2500





5. pretrained word embedding layer with training sample size = 10000



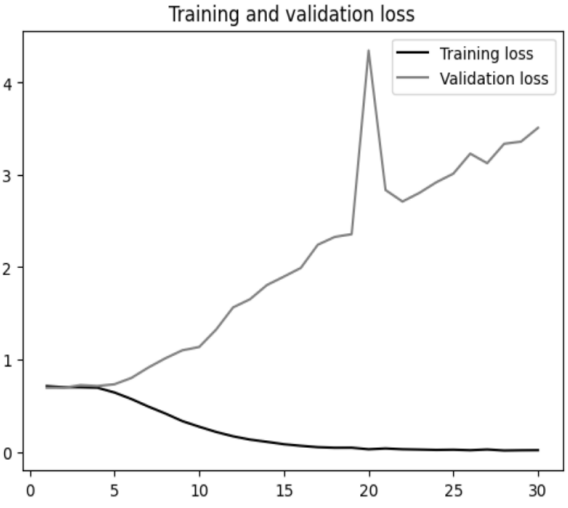


Table 1: Custom-Trained Embedding Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss	Test Loss
100	50.38	50.39	0.6937	0.6935
1,000	69.86	69.90	0.6272	0.626
2,500	80.89	80.80	0.4880	0.496
5,000	83.10	83.10	0.4670	0.466
10,000	84.34	84.63	0.5562	0.5554

Table 1: Custom-Trained Embedding Results

This table demonstrates the progressive improvement in model performance as the training sample size increases:

- Accuracy: Validation and test accuracy steadily improve, starting from approximately 50% with 100 samples to over 84% with 10,000 samples.
- Loss: Validation and test loss decrease consistently, showing better fit and generalization with larger training datasets.

The custom embeddings successfully learned patterns in the data, improving model accuracy and stability as the training size increased.

Table 2: Pretrained Embedding (GloVe) Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss	Test Loss
100	57.13	50.60	0.8023	0.850
1,000	50.61	49.30	2.0035	2.049
2,500	50.29	50.36	1.7286	1.685
5,000	49.40	50.38	1.7722	1.704
10,000	49.80	49.90	3.8040	3.705

Table 2: Pretrained Embedding (GloVe) Results

This table shows a stark contrast where increasing the training size did not significantly enhance performance:

- Accuracy: Validation and test accuracy stagnate around 50%, indicating poor generalization and learning.
- Loss: Validation and test loss values remain high, suggesting difficulty in adapting pretrained embeddings to the dataset.

The pretrained embeddings were unable to leverage the data effectively, resulting in consistently poor performance regardless of the training sample size.

Conclusion:

The study compared **custom-trained embeddings** and **pretrained word embeddings (GloVe)** for sentiment analysis on the IMDB dataset. Custom embeddings outperformed GloVe in all metrics, achieving **84.63% test accuracy** with 10,000 training samples, while GloVe stagnated at around 50%. Increasing the training sample size improved performance for custom embeddings, demonstrating their adaptability to the dataset, whereas GloVe struggled to generalize due to limited alignment with

the task-specific vocabulary. Overall, custom-trained embeddings are more effective for this dataset, especially with larger training samples.