ASSIGNMENT-4 (TEXTDATA)

The study focused on analyzing sentiment within the IMDB dataset, comprising 50,000 movie reviews, using a binary classification approach. It compared two methods for converting textual data into numerical representations: custom-trained embedding layers and pretrained word embeddings (GloVe). The evaluation also explored how varying training sample sizes influenced accuracy and test loss.

Data Preprocessing

- **Text Conversion**: Movie reviews were converted into numerical sequences by mapping each word to a corresponding integer index. To ensure consistency, sequences were padded to maintain a uniform length across all samples.
- **Embedding Techniques**:
    - **Custom-Trained Embedding Layer**: An embedding layer trained directly on the dataset.
    - **Pretrained Embedding Layer (GloVe)**: Leveraged word embeddings pre-trained on large text corpora.

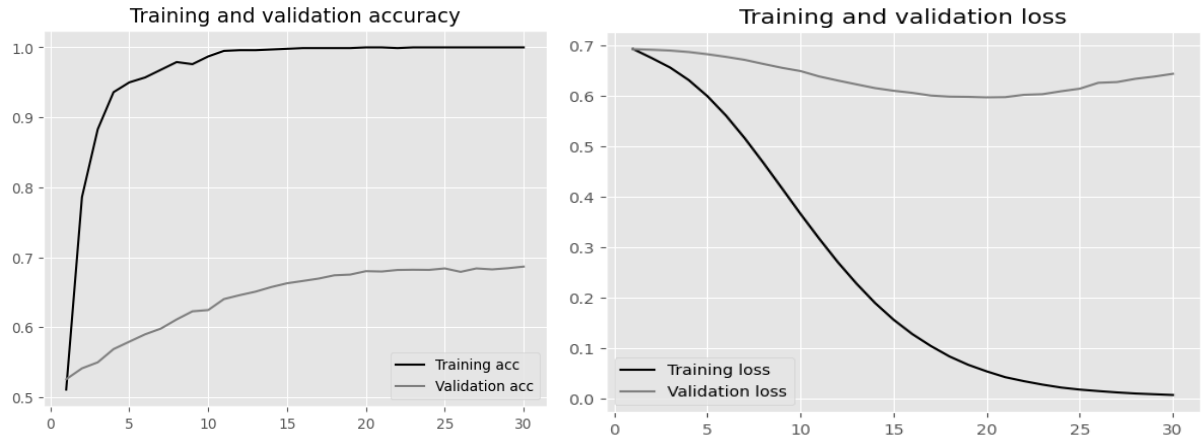
Procedure

- **Custom-Trained Embeddings**:
  Models were trained on various dataset subsets containing 100, 1,000, 5,000, and 10,000 samples. After training, performance was assessed by recording training accuracy and test loss using a fixed validation set.
- **Pretrained Embeddings (GloVe)**:
  The same subsets and evaluation process were applied, mirroring the custom-trained setup, to compare performance on the validation dataset.

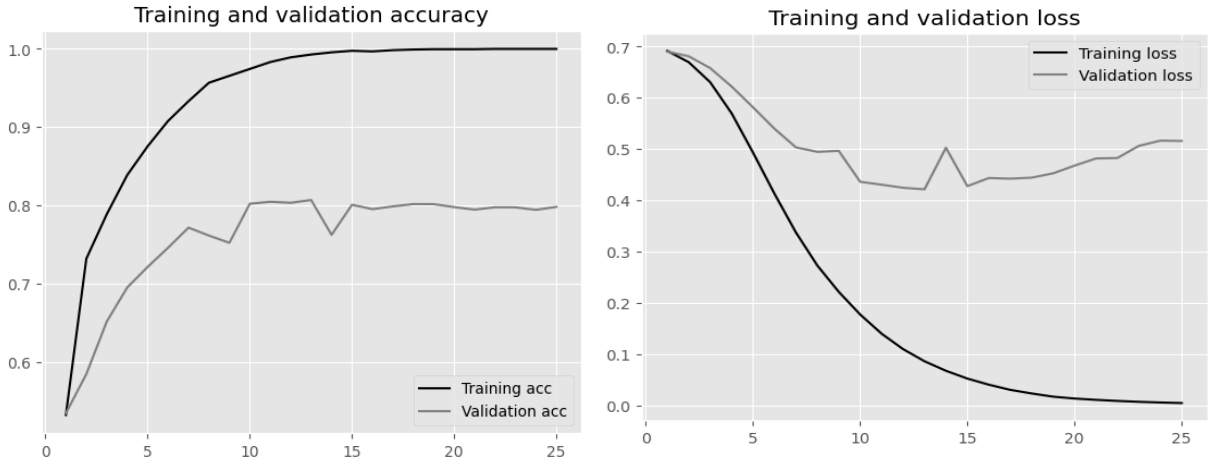
CUSTOM-TRAINED EMBEDDING LAYER

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

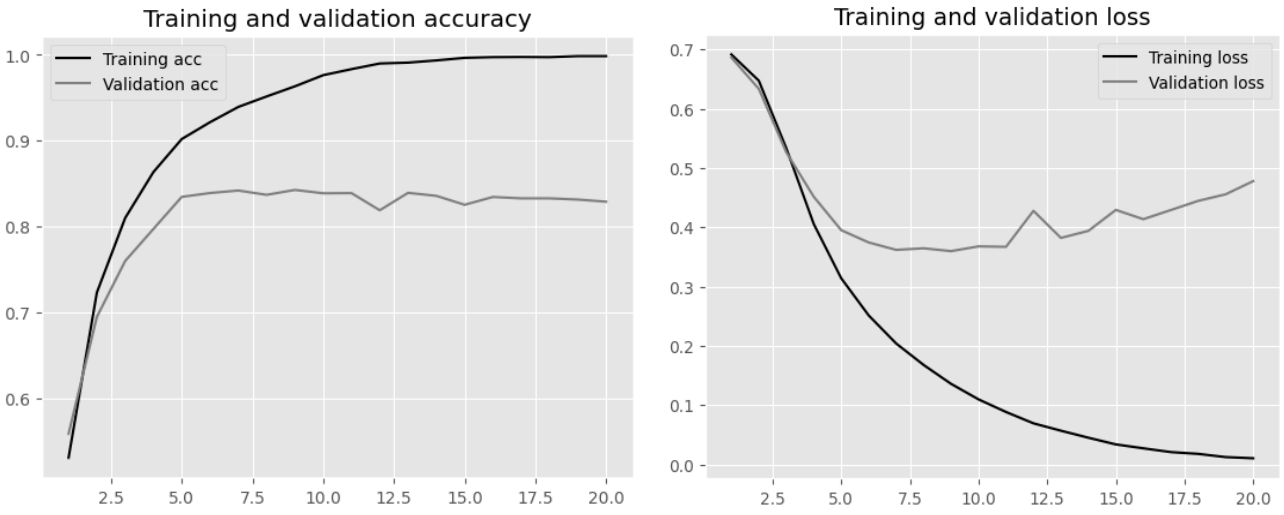
2. Custom-trained embedding layer with training sample size = 1000



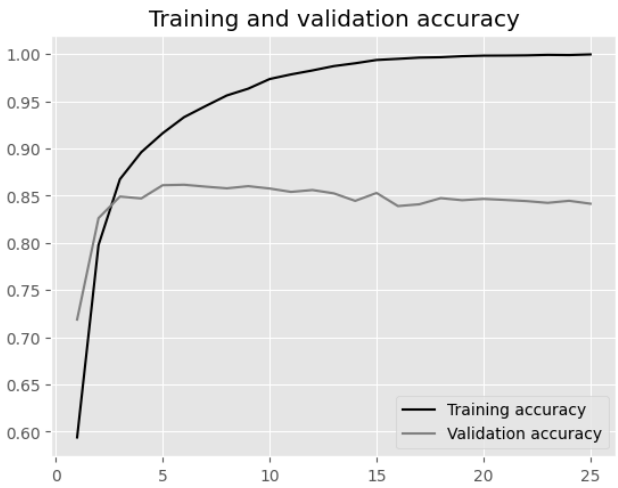
3. Custom-trained embedding layer with training sample size = 2500

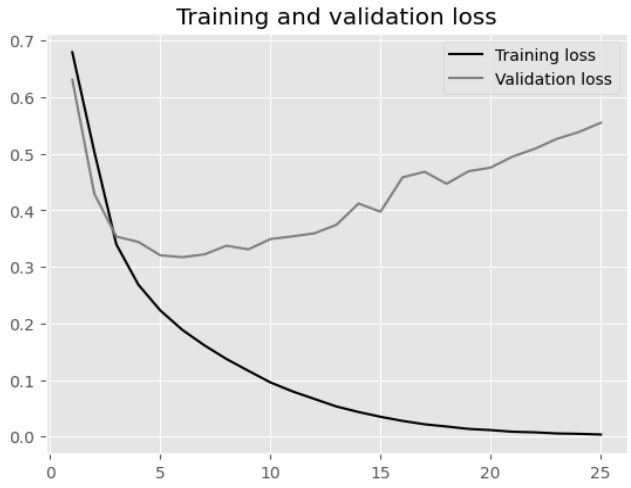


4. Custom-trained embedding layer with training sample size = 5000



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

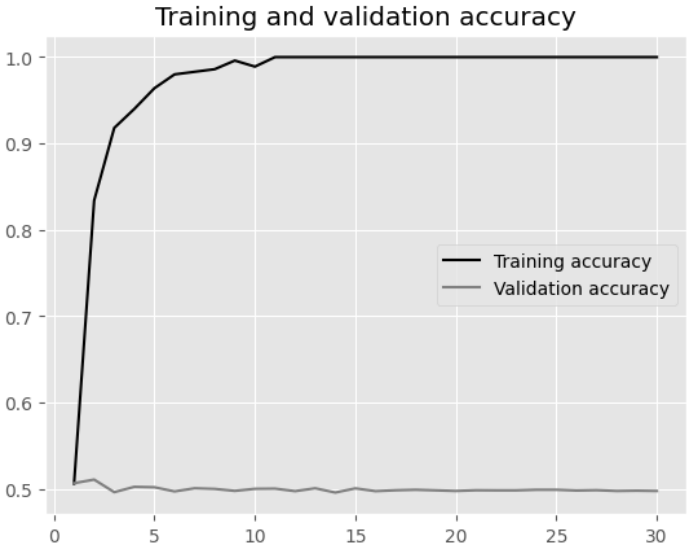


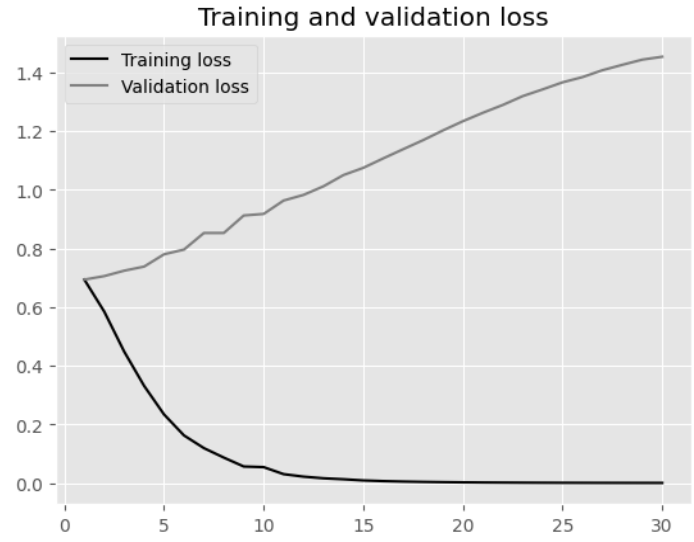


The accuracy of the custom-trained embedding layer ranged from 97.3% to 100%, depending on the training sample size, with the highest accuracy achieved using a sample size of 100.

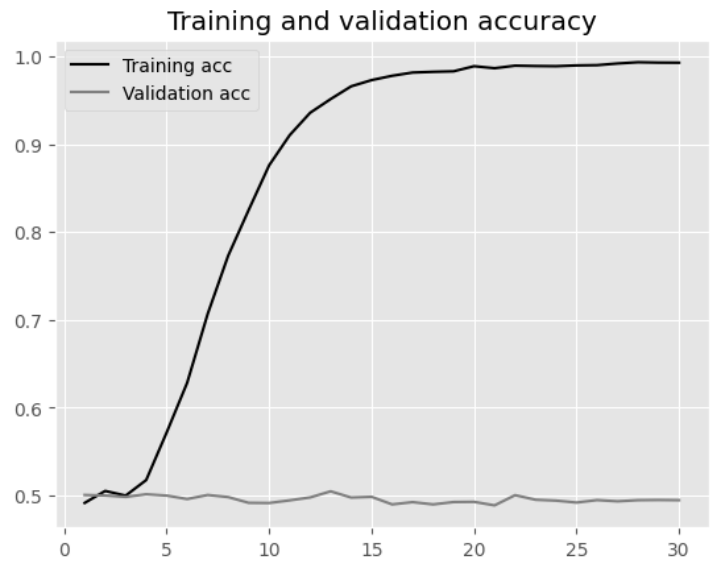
PRETRAINED WORD EMBEDDING LAYER :

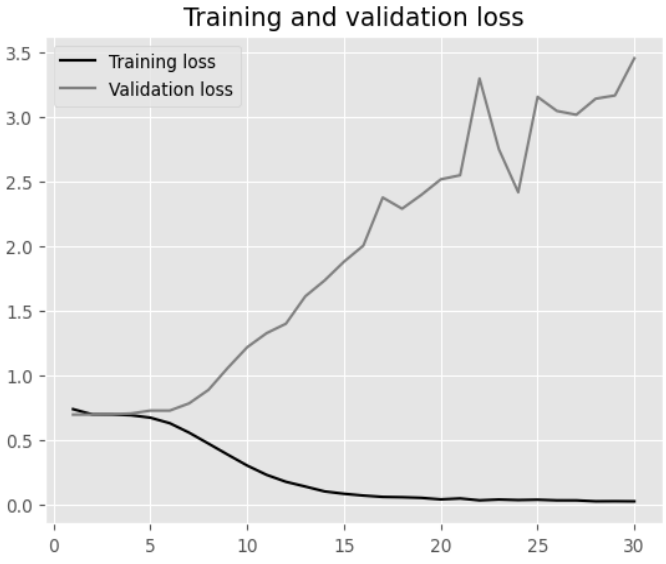
1. pretrained word embedding layer with training sample size = 1000



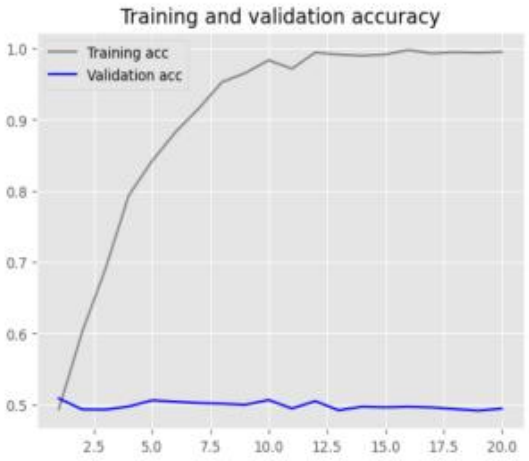


2. pretrained word embedding layer with training sample size = 10000

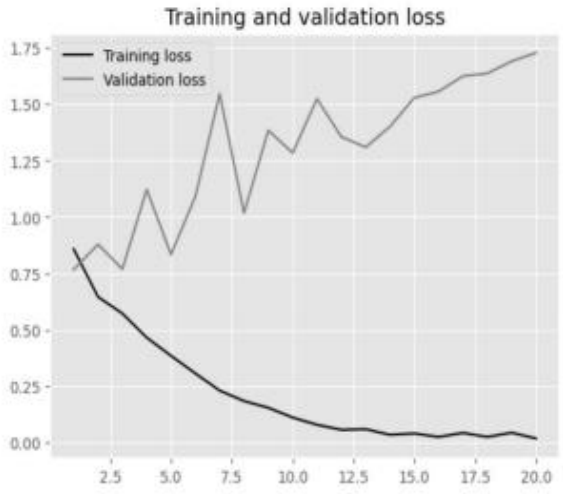




3. pretrained word embedding layer with training sample size = 5000



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



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

Table 1: Custom-Trained Embedding Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss	Test Loss
100	50.70	49.80	0.6952	0.69627
1000	68.56	68.63	0.6477	0.6487
2500	79.99	79.93	0.514	0.517
5000	82.31	82.49	0.486	0.480
10000	83.96	83.99	0.564	0.568

The table illustrates a gradual enhancement in model performance with larger training sample sizes:

- **Accuracy**: Both validation and test accuracy show steady improvement, increasing from around 50% with 100 samples to over 84% with 10,000 samples.
- **Loss**: Validation and test loss consistently decrease, indicating a better fit and improve generalizationas the training dataset grows.
  The custom embeddings effectively captured patterns in the data, leading to greater accuracy and stability as the training sample size expanded.

Table 2: Pretrained Embedding (GloVe) Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss	Test Loss
100	57.13	50.6	0.8023	0.850
1000	49.8	48.8	1.588	1.589
2500	49.6	50.36	1.7286	1.680
5000	48.8	50.38	1.7722	1.704
10000	48.4	49.9	3.804	3.705

The table highlights a clear disparity where increasing the training size had minimal impact on performance:

- **Accuracy**: Both validation and test accuracy remained stagnant around 50%, reflecting poor generalization and learning.
- **Loss**: Validation and test loss stayed high, indicating difficulties in adapting pretrained embeddings to the dataset.
  Pretrained embeddings failed to effectively utilize the data, resulting in consistently poor performance regardless of the training sample size.

Conclusion:
This study compared custom-trained embeddings and pretrained word embeddings (GloVe) for sentiment analysis on the IMDB dataset. Custom embeddings significantly outperformed GloVe across all metrics, achieving a test accuracy of 84.63% with 10,000 training samples, while GloVe plateaued at approximately 50%. Larger training sample sizes improved the performance of custom embeddings, showcasing their ability to adapt to the dataset. In contrast, GloVe struggled to generalize due to limited alignment with task-specific vocabulary. Overall, custom-trained embeddings proved to be more effective for this dataset, particularly with increased training samples.