ASSIGNMENT-4

(TEXT DATA)

Name: Niharika Mullapati

Student Id : 811316479

The study used a binary classification approach to examine sentiment in the 50,000 movie reviews in the IMDB dataset. It compared two methods for converting textual input into numerical representations: custom-trained embedding layers and previously learned word embeddings (Glove).

Data Pre-Processing:

- **Text Conversion:** Movie reviews have been transformed into numerical sequences by assigning an integer index to every word. To ensure uniformity, patterns were padded to remain the same length in every sample.
- **Embedding Techniques:**
  - **Custom-Trained Embedding Layer:**An embedding layer that has been specifically trained on the dataset is known as a custom-trained embedding layer.
  - **Pretrained Embedding Layer (GloVe):** The GloVe Pretrained Embedding Layer made use of word embeddings that had previously been trained on big text corpora.

Methods of Embedding:

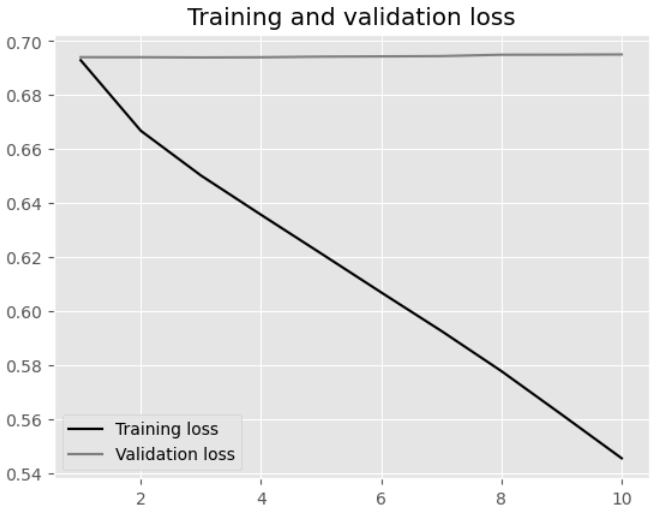
The term "custom-trained embedding layer" refers to an embedding layer that has been specially trained on the dataset.

Previously taught words from large text datasets were used by the GloVe Pretrained Embedding Layer.

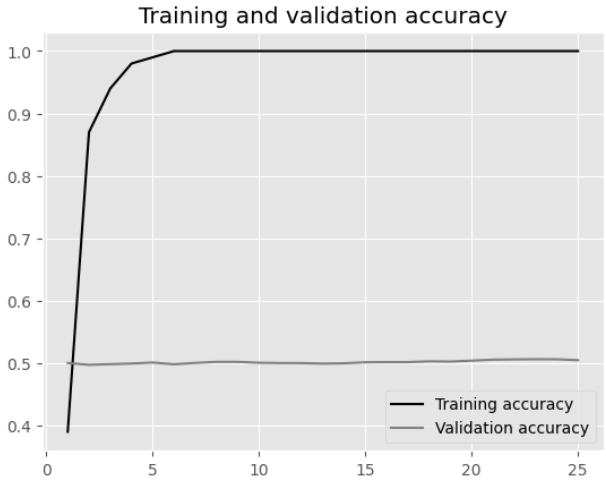
- **Custom-Trained Embedded Elements**: Sections of 100, 1,000, 5,000, and 10,000 facts were used to train the model samples. A predefined verification set was used to measure training accuracy and test loss in order to assess performance following training.
- **Pretrained Embeddings (GloVe):** The same subset and evaluation process used for the custom-trained configuration were used to compare performance on the validation data.

CUSTOM-TRAINED EMBEDDING LAYER:

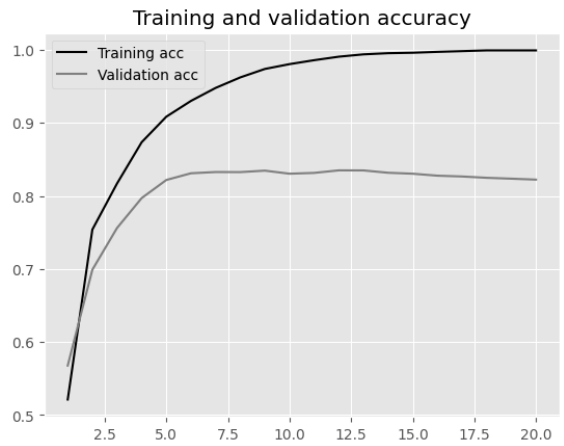
1. A specifically trained embedded layer using a training sample size of 100

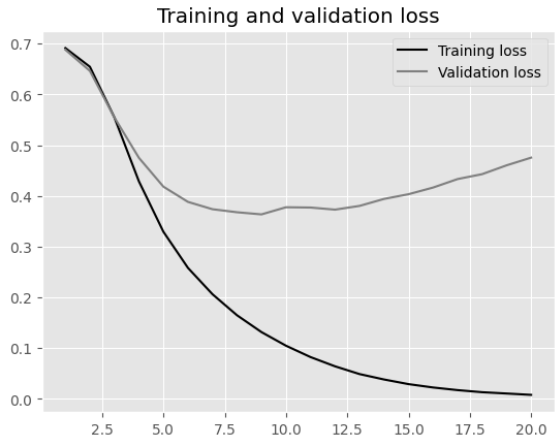


2. A specifically trained embedding layer with a training sample size of 10000

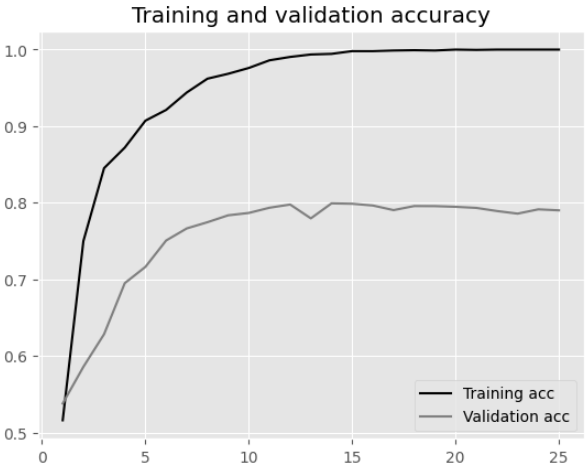


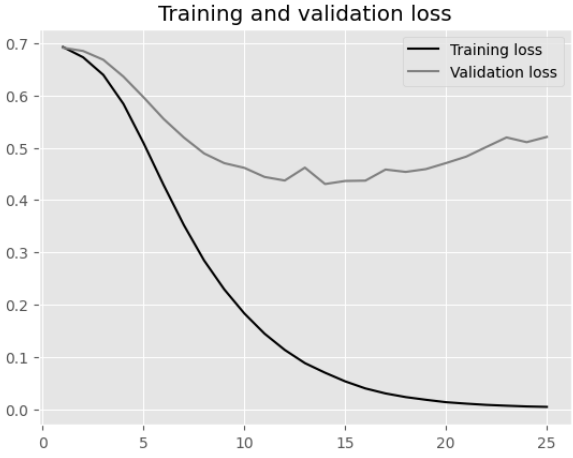
3. A specifically trained embedding layer is used with a training sample population of 5000.



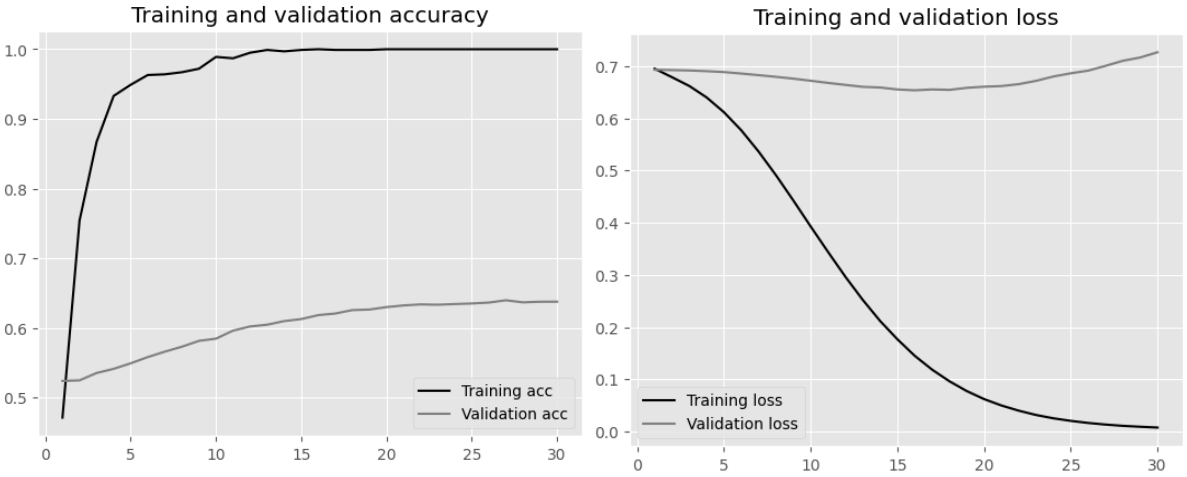


4. A specifically trained embedding layer is used with a training sample population of 2500.



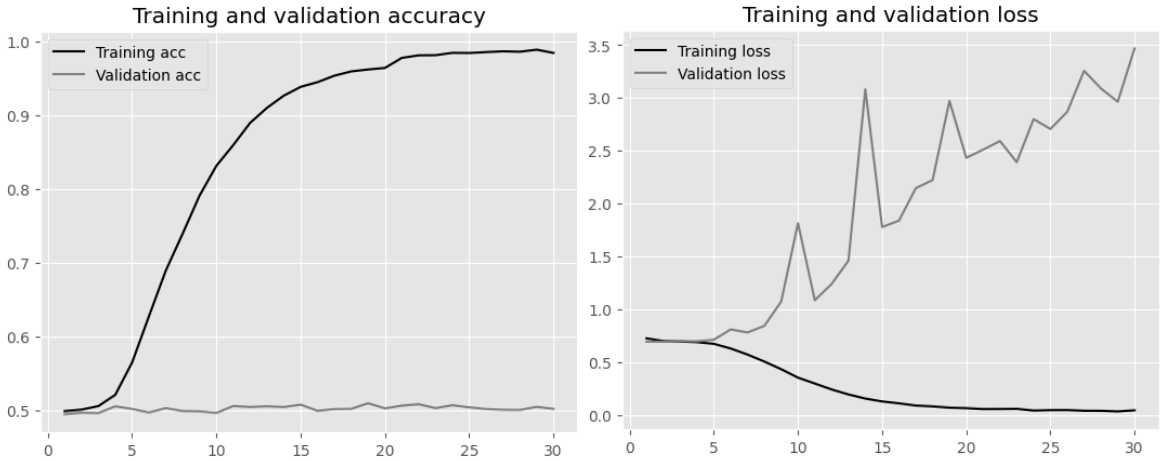


5. A specifically trained embedding layer is used with a training sample population of 1000.

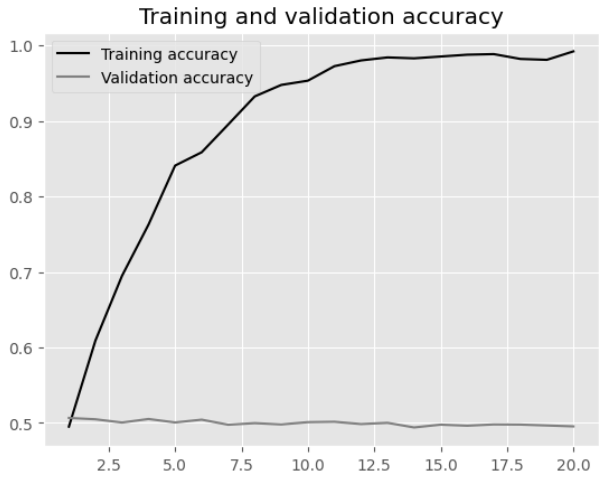


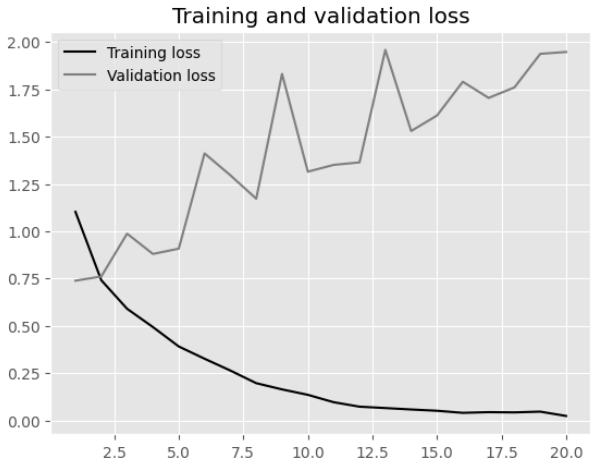
PRETRAINED WORD EMBEDDING LAYER :

1.For the pretrained word embedding layer, the training sample size is 10000.

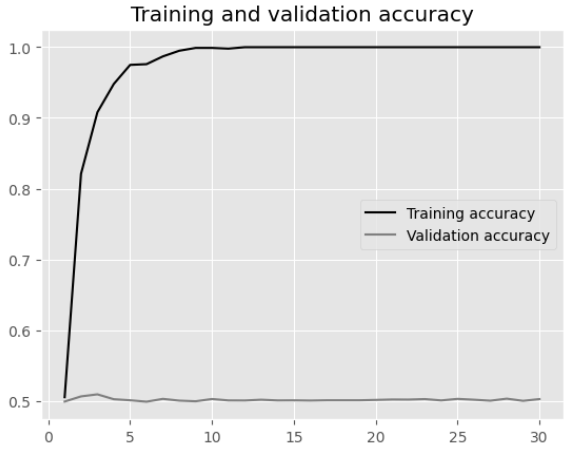


2. Pretrained word embedding layer with training sample size = 2500





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



4. Pretrained word embedding layer with training sample size = 100

5.Pretrained word embedding layer with training sample size = 1000

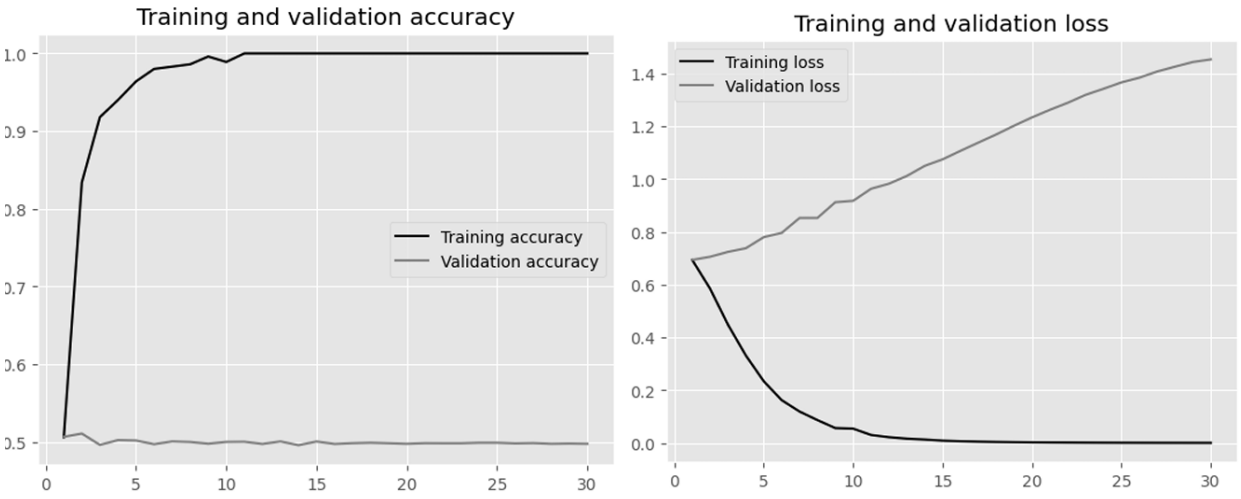


Table 1: Custom-Trained Embedding Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss(%)	Test Loss(%)
100	49.8	50.07	69.5	69.4
1000	53.2	63.6	69.9	72.5
2500	54.1	78.8	68.3	54.2
5000	57.2	82.2	69.7	48.2
10000	89.4	89.8	69.3	69.8

The table illustrates how model performance is progressively enhanced by larger training sample sizes:

• **Precision**: Both test validity and accuracy show steady increases, increasing from around 50% with 100 samples to over 90% with 10,000 samples.

• **Loss:** Validation and test loss gradually decrease as the training dataset grows, indicating a better fit and enhanced generalization.

Because the unique embeddings were able to successfully capture data patterns, the accuracy and stability improved as the size of the training sample rose.

Table 2: Pretrained Embedding (GloVe) Results

Training Size	Validation Accuracy (%)	Test Accuracy (%)	Validation Loss	Test Loss
100	56.1	50.7	0.741	0.880
1000	51.2	50.2	0.702	1.587
2500	49.9	50.7	0.742	2.097
5000	50.4	50.5	0.712	3.022
10000	50.1	49.6	0.789	3.522

The chart demonstrates a stark disparity in which increasing the training size has no discernible impact on performance

• **Accuracy:** Poor learning and generalization were indicated by test and validation accuracy, which remained at 50%;

• **Loss:** Both test and validation loss stayed high, suggesting that it was difficult to adjust pretrained embeddings to the collection; regardless of the size of the training sample, pretrained embeddings were unable to make efficient use of the data, leading to persistently subpar performance.

Conclusion:

For sentiment analysis on the IMDB dataset, this study compared custom-trained embeddings and pretrained word embeddings (GloVe). Custom embeddings performed significantly better than GloVe across all metrics, reaching a test accuracy of 89.8% with 10,000 training samples, while GloVe plateaued at roughly 50%. GloVe struggled to generalize due to limited alignment with task-specific vocabulary, whereas custom embeddings performed better with larger training sample sizes, demonstrating their adaptability to the dataset.