ASSIGNMENT 4: TEXT AND SEQUENCE DATA

OBJECTIVE

The IMDB dataset aims to classify movie reviews into positive or negative categories through binary classification. It comprises 50,000 reviews, focusing on the top 10,000 words for evaluation. Training samples are limited to **100**, **5000**, **1000**, and **100,000**, with validation conducted on 10,000 samples. Data preparation is carried out before applying the embedding layer and pre-trained embedding model. Different approaches are tested to assess performance.

DATA PREPROCESSING

- As part of the data preparation process, each review transforms word embeddings, representing each word with a fixed-size vector. A limit of 10,000 samples constrains this procedure. Moreover, a numerical sequence is derived from the reviews, with each number corresponding to a unique word. However, the neural network's input cannot directly accommodate this list of numbers, necessitating the construction of tensors. These tensors are created using the integer list, forming a data structure with an integer data type, and arranged in a (samples, word indices) format. To achieve this, it's imperative to ensure uniformity in sample length by padding the reviews with dummy words or numbers.
- During dataset preparation, reviews are transformed into word embeddings, where each word is encoded into a fixed-size vector. This process is limited to 10,000 samples. Additionally, numerical representations are generated from the reviews, assigning a distinct number to each word. However, the neural network's input format doesn't directly support this numerical list. To address this, tensors need to be constructed. These tensors are formulated using the integer list, structured with an integer data type in a (samples, word indices) format. Ensuring consistency in sample length requires padding the reviews with dummy words or numbers.

PROCEDURE

We explored two distinct methods for generating word embeddings on the IMDB dataset:

- Developing a custom-trained embedding layer.
- Utilizing a pre-trained word embedding layer based on the GloVe model, a widely recognized pre-trained model trained on extensive textual data.

To evaluate the effectiveness of these embedding strategies, we employed two embedding layers on the IMDB review dataset and compared their accuracy across varying training sample sizes (**100**, **5000**, **1000**, and **10,000**).

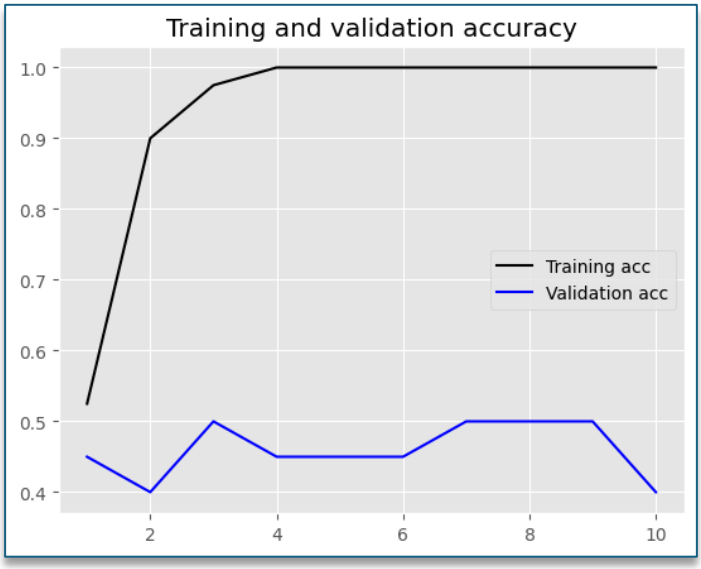
Our approach involved:

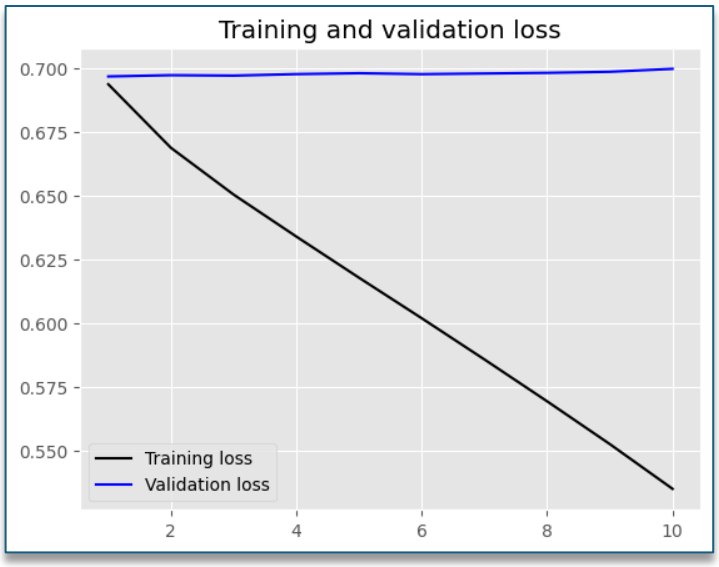
- Training a specially tailored embedding layer using the IMDB review dataset.
- Assessing the accuracy of each model on a testing set after training on different sample sizes.

- Contrasting these accuracy metrics with those obtained from a model incorporating a pre-trained word embedding layer.

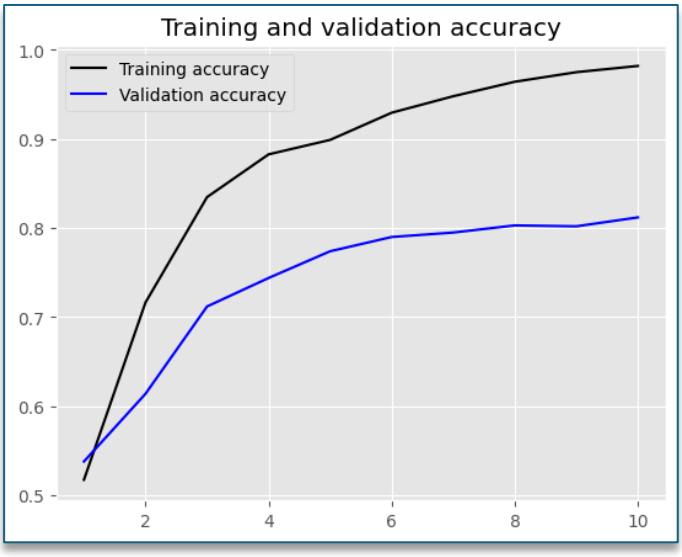
CUSTOM TRAINED EMBEDDING LAYER

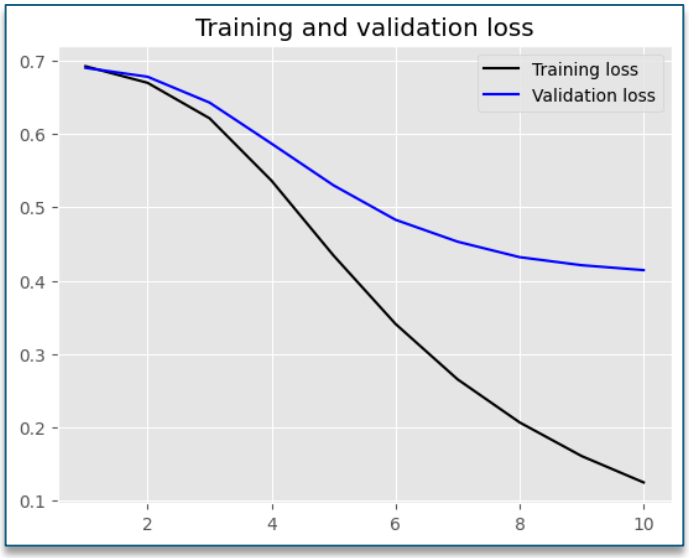
1. A custom-trained embedding layer was utilized with a training sample size of **100**.



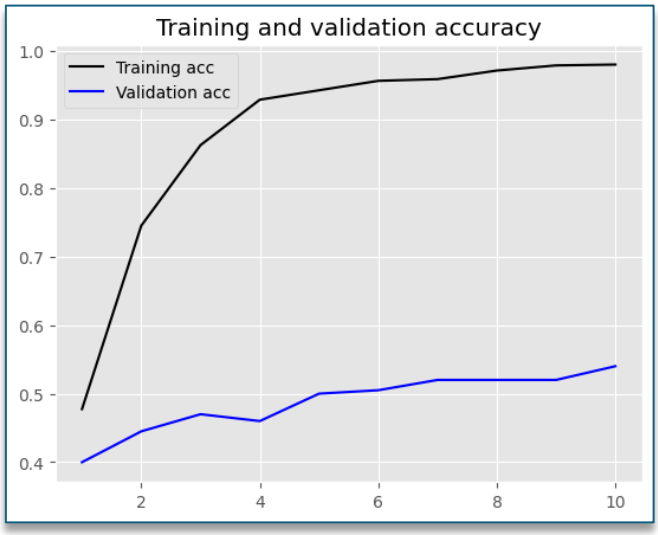


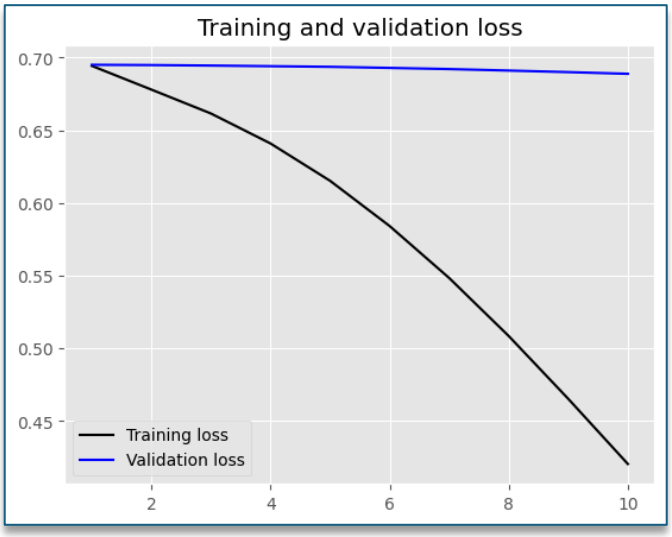
2. A custom-trained embedding layer was utilized with a training sample size of **5000**.



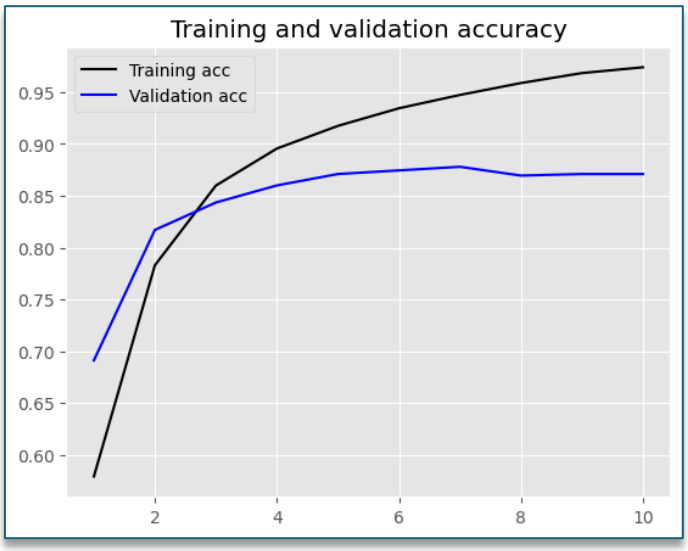


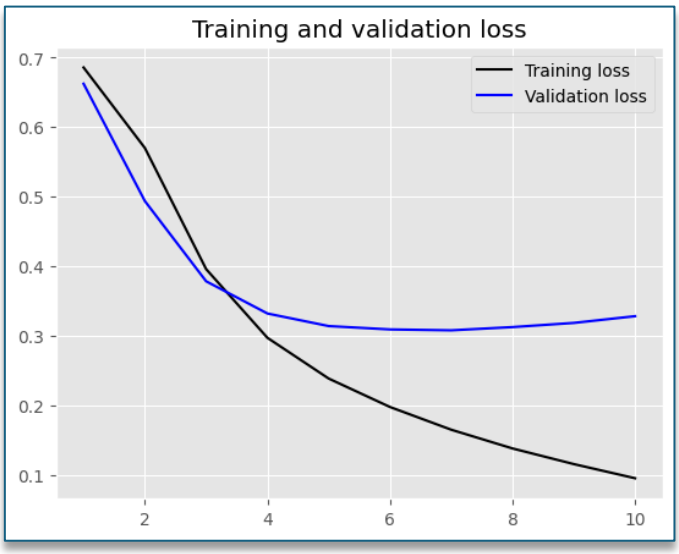
3. A custom-trained embedding layer was utilized with a training sample size of **1000.**





4. A custom-trained embedding layer was utilized with a training sample size of **10000.**

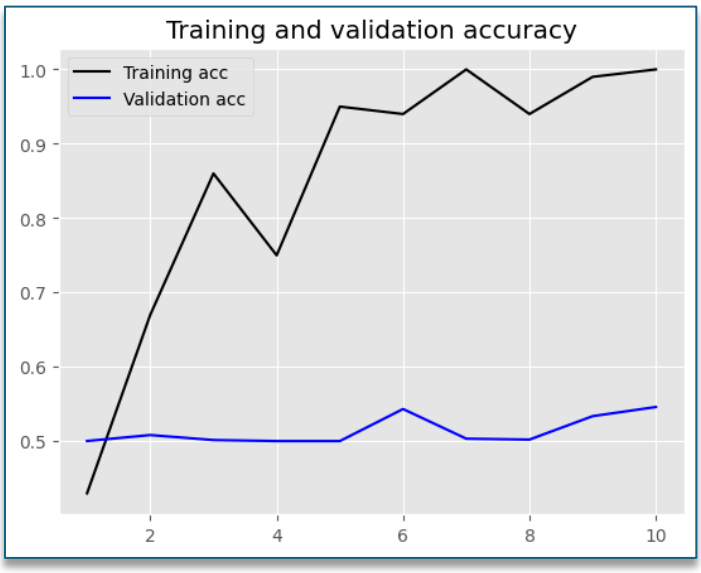


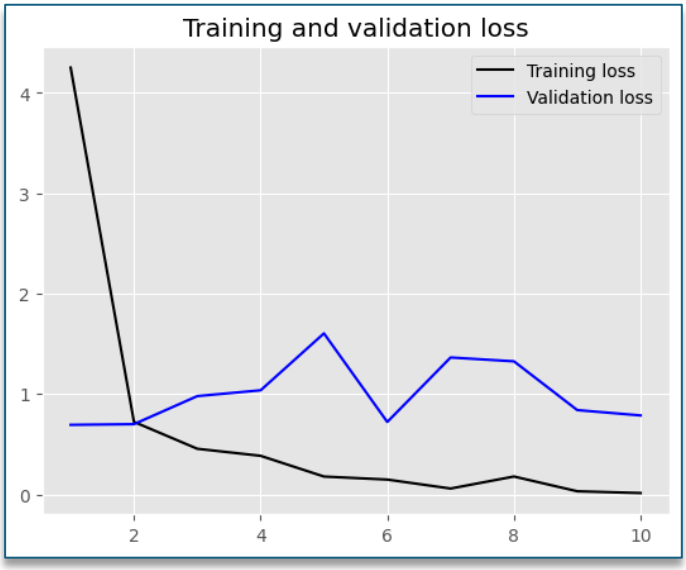


As the size of the training sample increases from 100 to 10,000, there is a consistent trend of improvement in both test accuracy and test loss. For instance, when the training sample size is 100, the test accuracy hovers around 50%. However, with a larger training sample size of 10,000, the test accuracy notably increases to approximately 85.77%. Concurrently, the test loss decreases as the sample size grows, reflecting enhanced model performance.

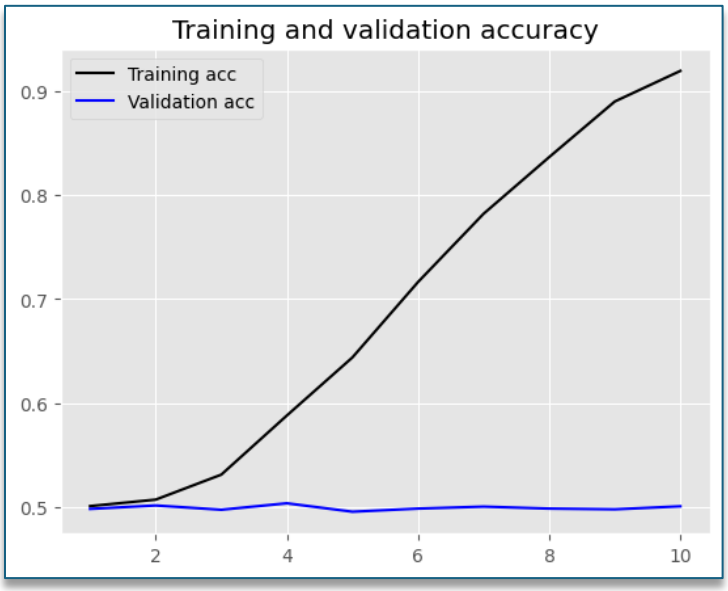
PRETRAINED WORD EMBEDDING LAYER

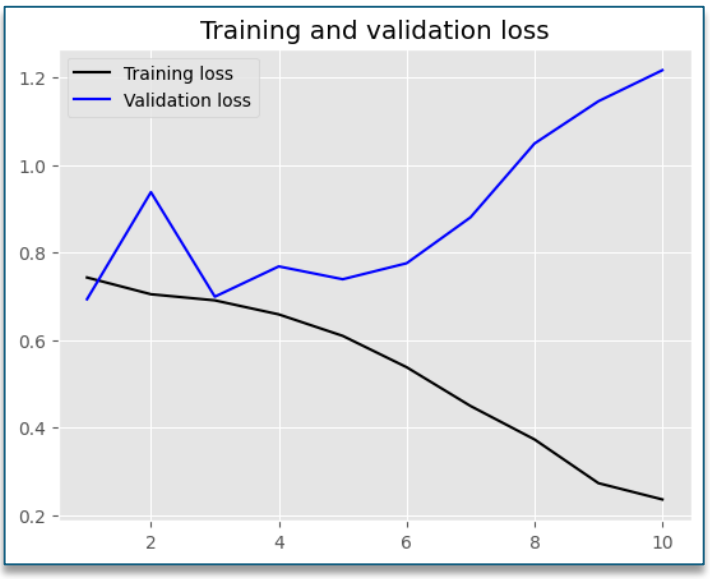
1. Using a pre-trained word embedding layer with a training sample size of **100**.



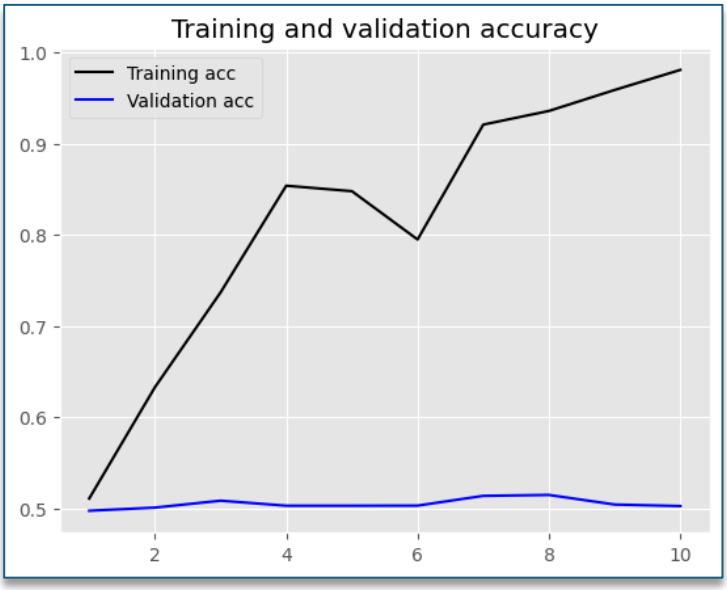


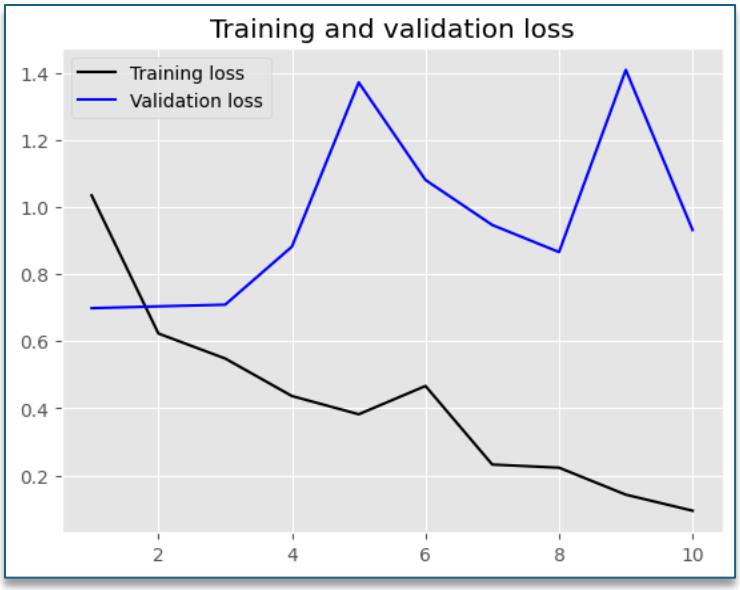
2. Using a pre-trained word embedding layer with a training sample size of **5000**.



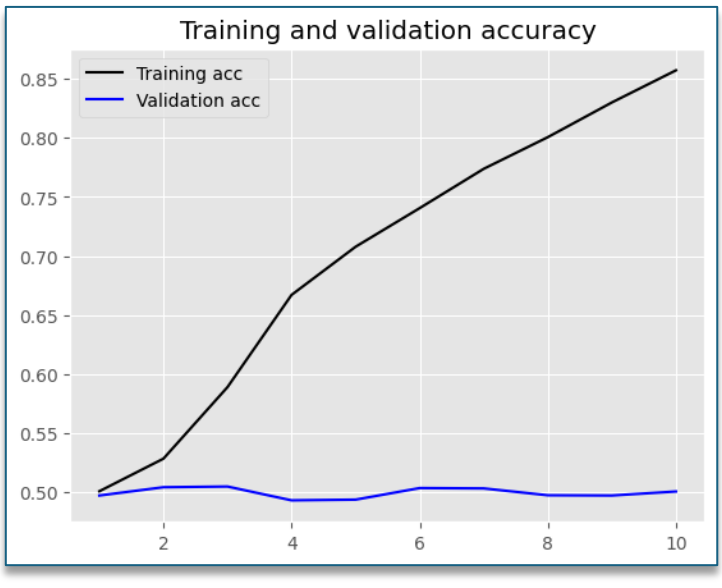


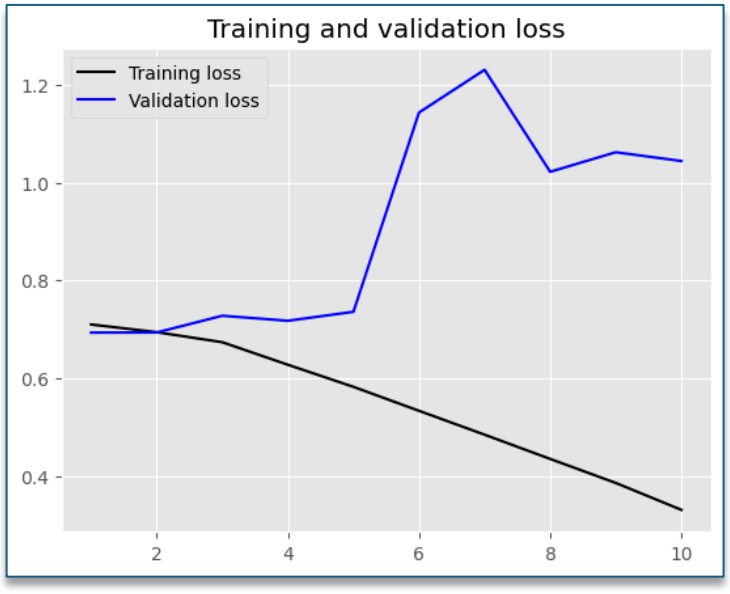
3. Using a pre-trained word embedding layer with a training sample size of **1000.**





4. Using a pre-trained word embedding layer with a training sample size of **10000.**





Global Vectors for Word Representation is widely utilized for pre-training word embeddings by leveraging global corpus statistics. Despite varying the training sample size from 100 to 10,000, there isn't a notable enhancement in test accuracy or test loss. Even with a substantial sample

size of 10,000, the test accuracy hovers around a relatively low 49.90%. Additionally, the test loss remains consistently high compared to custom-trained embeddings, indicating inferior performance.

RESULTS

Embedding Technique	Training Sample Size	Test Accuracy (%)	Test Loss
Custom-trained embedding layer	100	50.24	0.69
Custom-trained embedding layer	5000	81.93	0.36
Custom-trained embedding layer	1000	56.02	0.67
Custom-trained embedding layer	10000	85.54	0.33
Pre-trained word embedding (GloVe)	100	50.34	0.81
Pre-trained word embedding (GloVe)	5000	50.23	1.13
Pre-trained word embedding (GloVe)	1000	49.15	0.97
Pre-trained word embedding (GloVe)	10000	49.90	1.16

CONCLUSION

In summary, the results indicate that custom-trained embedding layers perform better than pre-trained GloVe embeddings regarding test accuracy and test loss, particularly with larger training sample sizes. This implies that for this particular tasks, training embeddings tailored to the specific task yields superior outcomes compared to employing pre-existing embeddings like GloVe.