AML Assignment-4

Text and Sequence Data

Sushma Chiluveru

Aim:

The binary classification task for the IMDB dataset is to categorize movie reviews as positive or negative. The dataset contains 50,000 reviews, evaluating the top 10,000 words. Training samples are limited to sizes of 100, 5,000, 10,000, and 100,000, while validation is conducted on 10,000 samples. After preparing the data, it is input into a pretrained embedding model and embedding layer, and various techniques are tested to assess performance.

Preparing the Data:

- The dataset preparation process converts each review into word embeddings, with each word represented by a fixed-size vector.
- This process limits the number of samples to 10,000. Instead of using a string of words, the reviews are transformed into sets of numbers representing individual words. However, these numeric lists are not directly suitable for the neural network's input.
- Tensors must be created from these numbers. One method is to generate a tensor with samples and word indices in an integer data type and format.
- To achieve this, all samples must be of the same length, which necessitates using dummy words or numbers to ensure uniform length for every review.

Methods Used:

For the IMDB dataset, I identified two methods for creating word embeddings:

1. Custom-trained embedding layer.

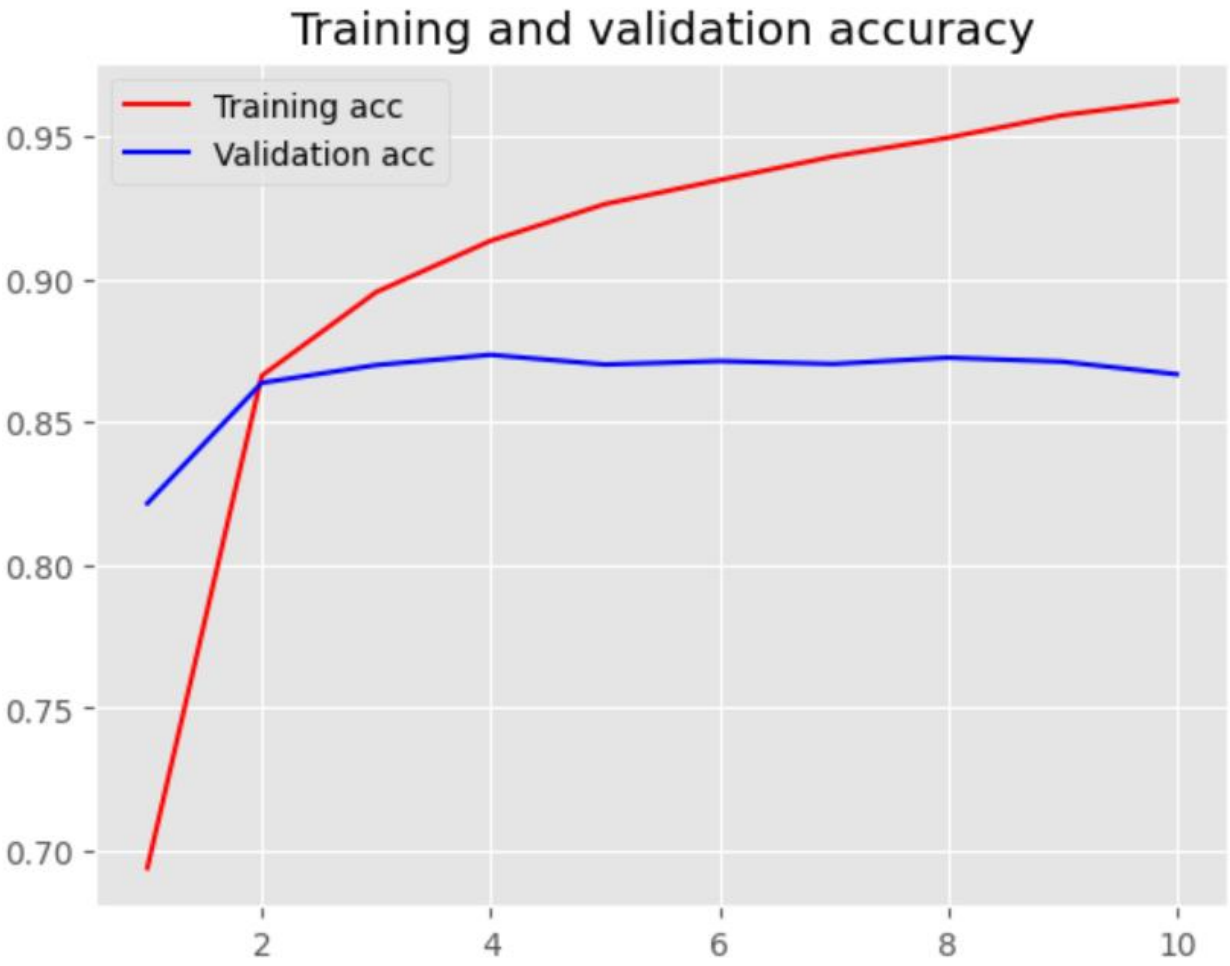
2. Pretrained word embedding layer using the GloVe model.

In this work, we used the widely known pretrained GloVe model, which is trained on extensive textual data. We evaluated accuracy across sample sizes of 100, 5,000, 1,000, and 10,000 by comparing custom-trained and pretrained embedding layers on the IMDB dataset. We tested

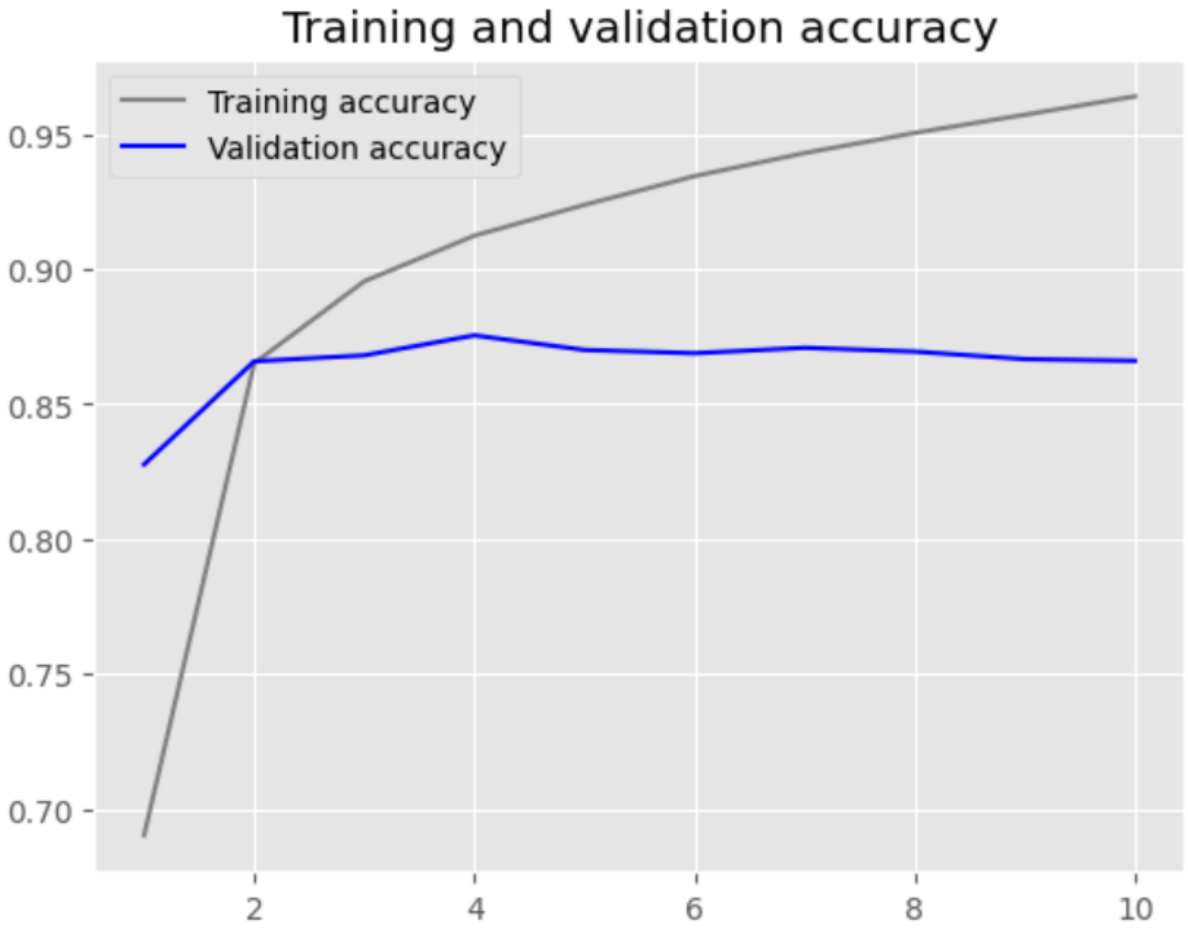
models using both pretrained and custom-trained embeddings on IMDB reviews with different sample sizes, assessing their accuracy on the test sets.

CUSTOM-TRAINED EMBEDDING LAYER:

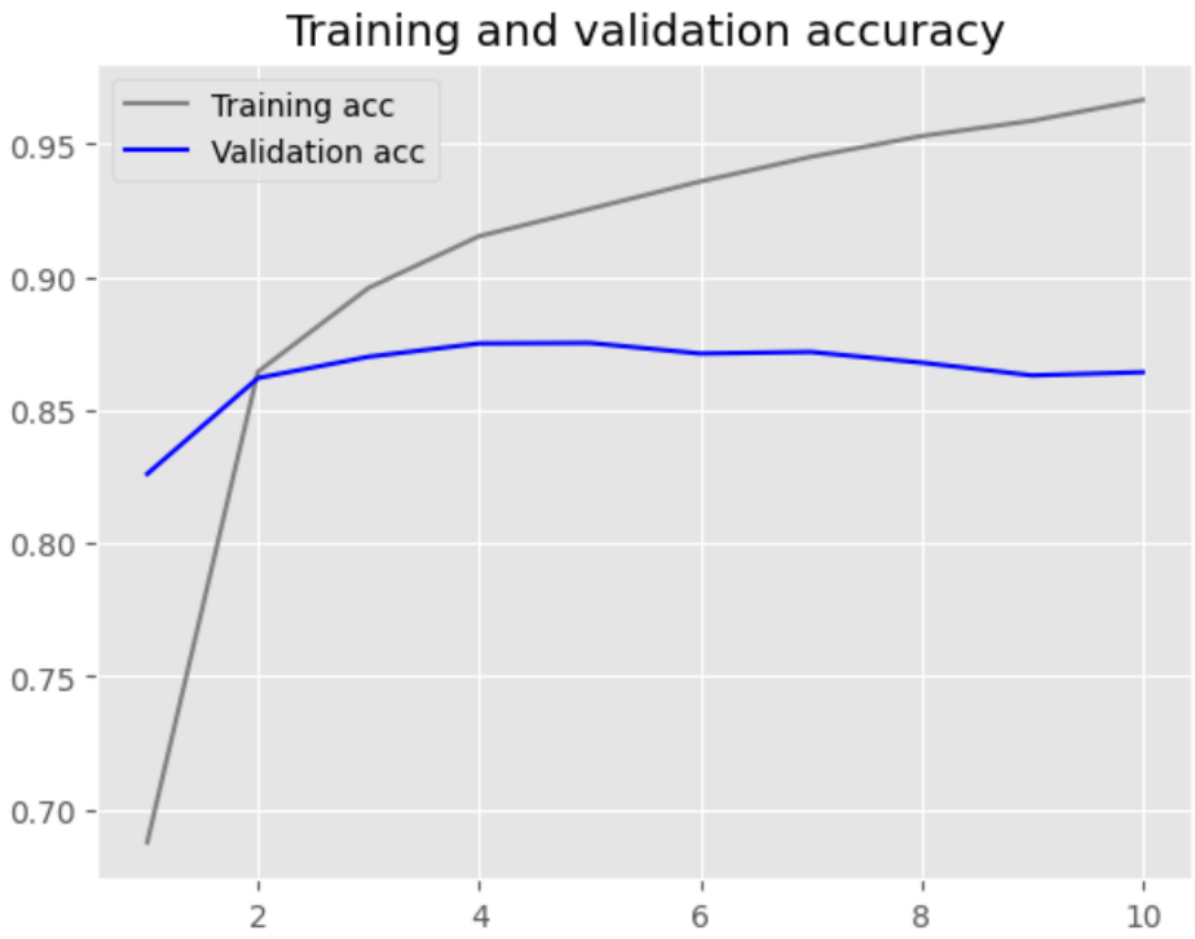
1. Custom-trained embedding layer with training sample size = 100

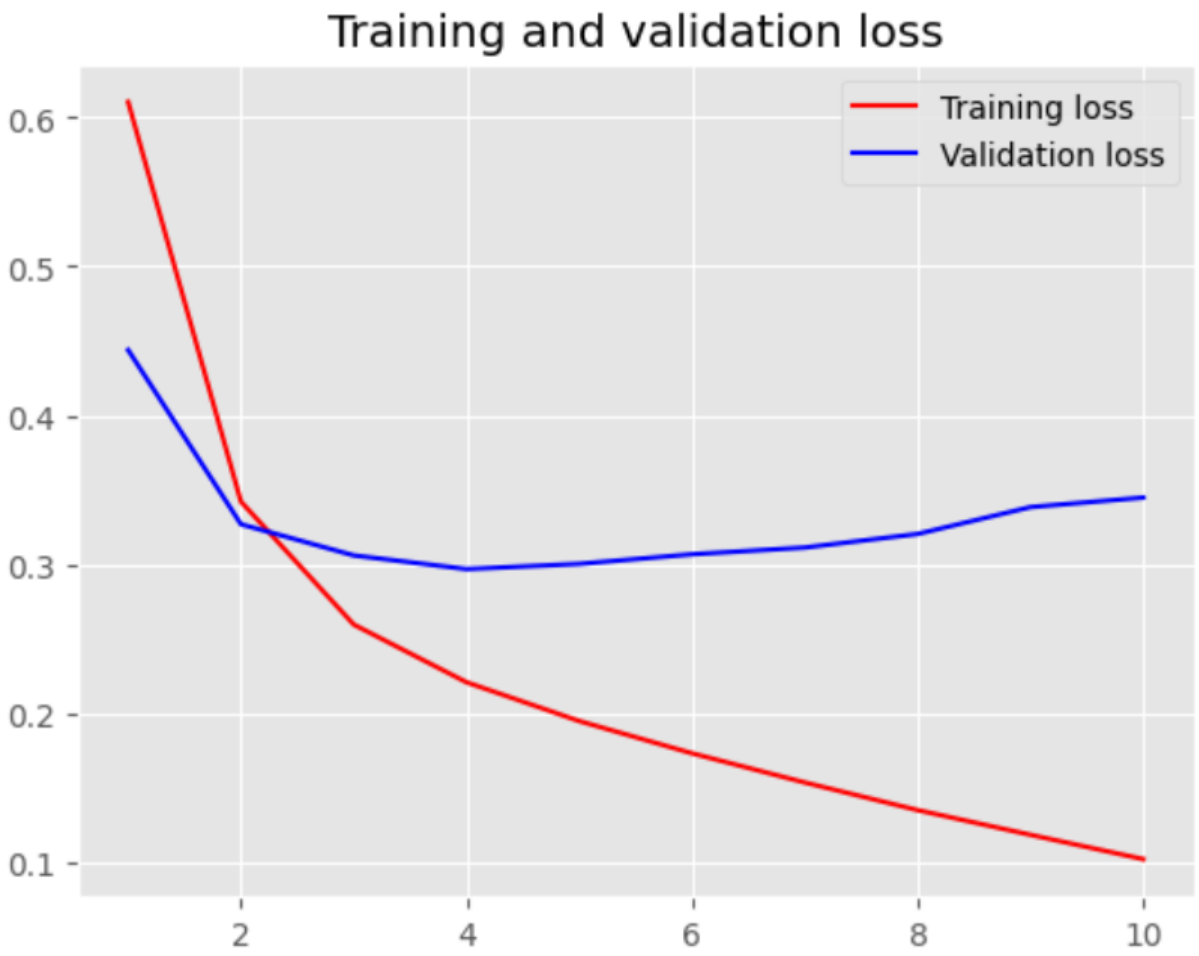


2. Custom-trained embedding layer with training sample size = 5000

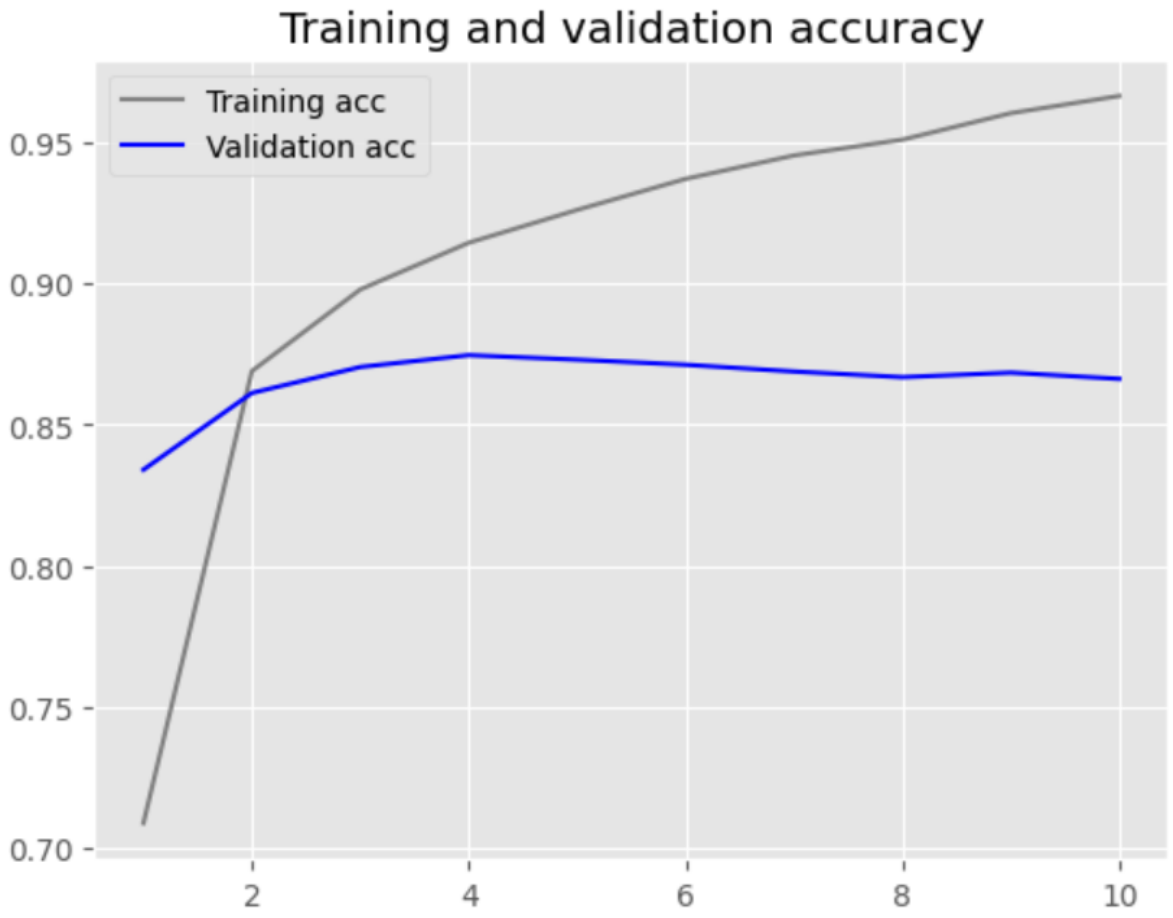

Training and validation accuracy
Training accuracy
Validation accuracy
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10

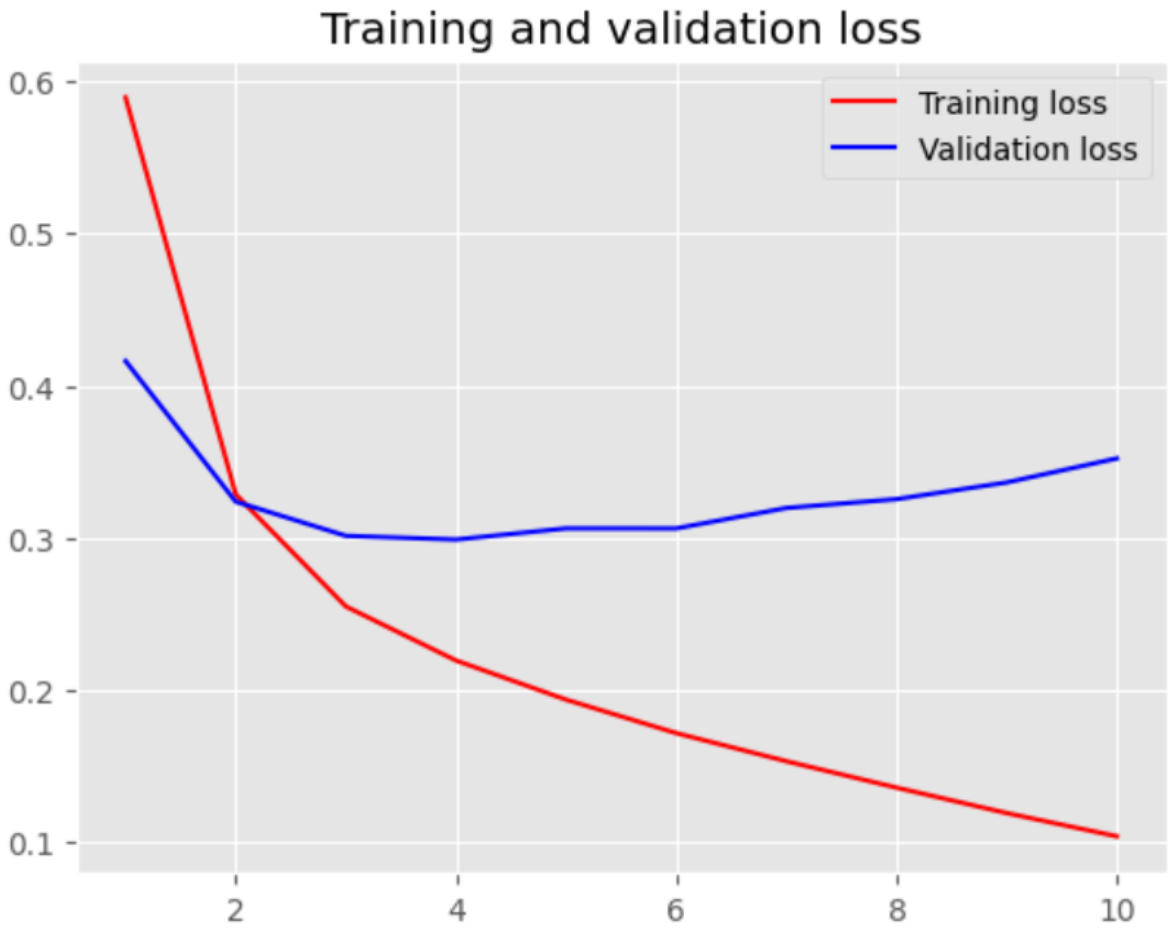
3. Custom-trained embedding layer with training sample size = 1000


Training and validation accuracy
Training acc
Validation acc
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10



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

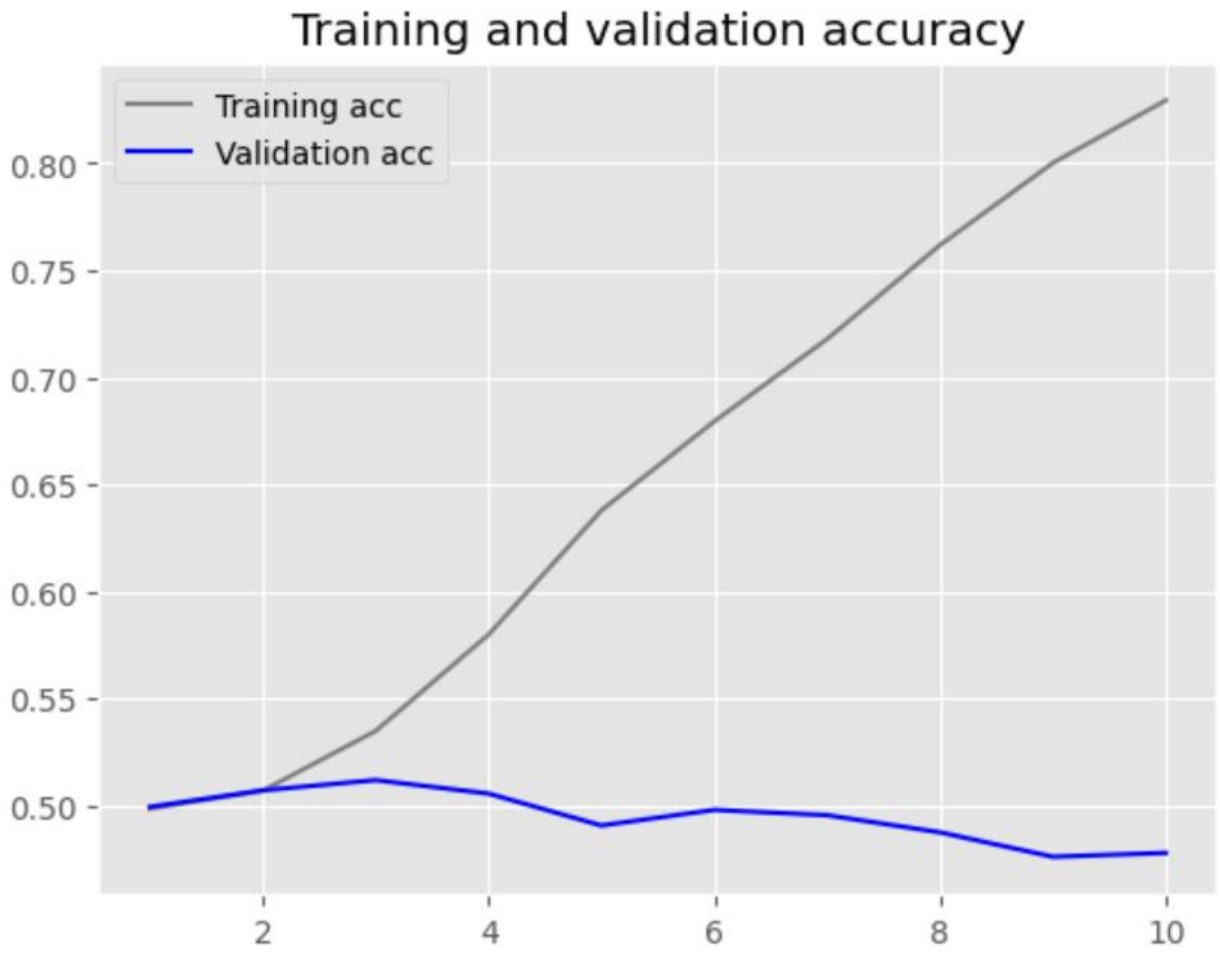

Training and validation accuracy
Training acc
Validation acc
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10

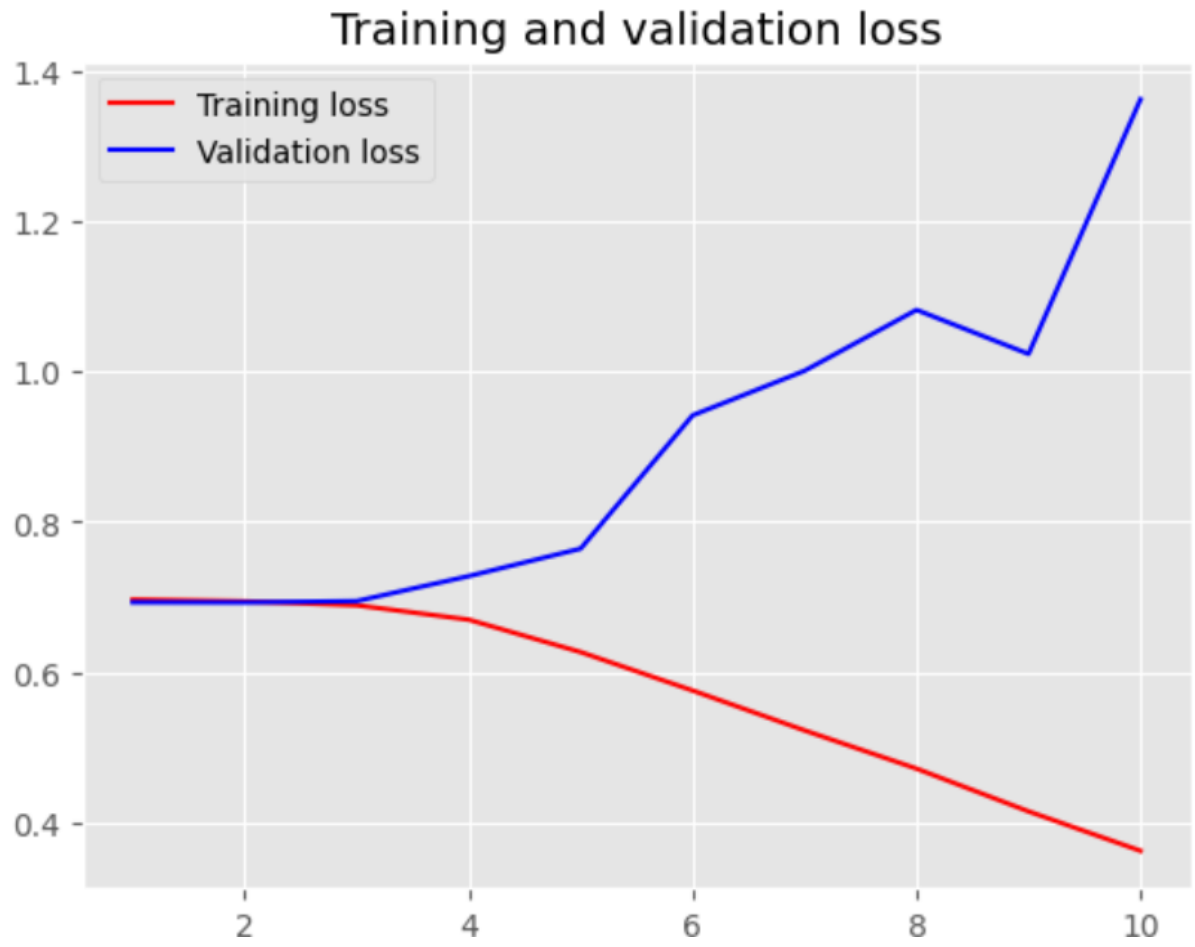


The accuracy of the custom-trained embedding layer ranged from 96.26% to 96.65%, depending on the training sample size, with the highest accuracy achieved with a 1000 and 10000 sample size which is same 96.65

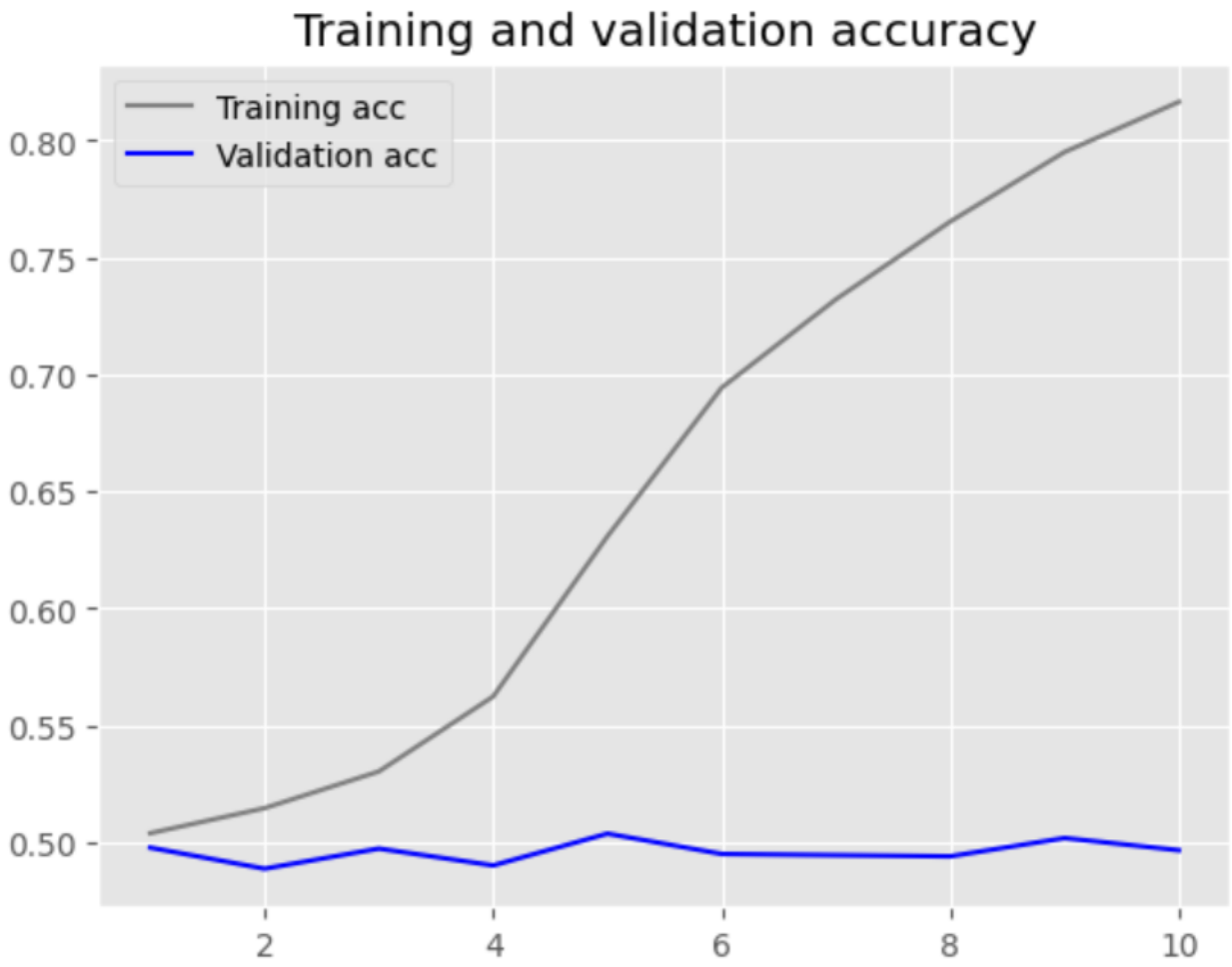
PRETRAINED WORD EMBEDDING LAYER

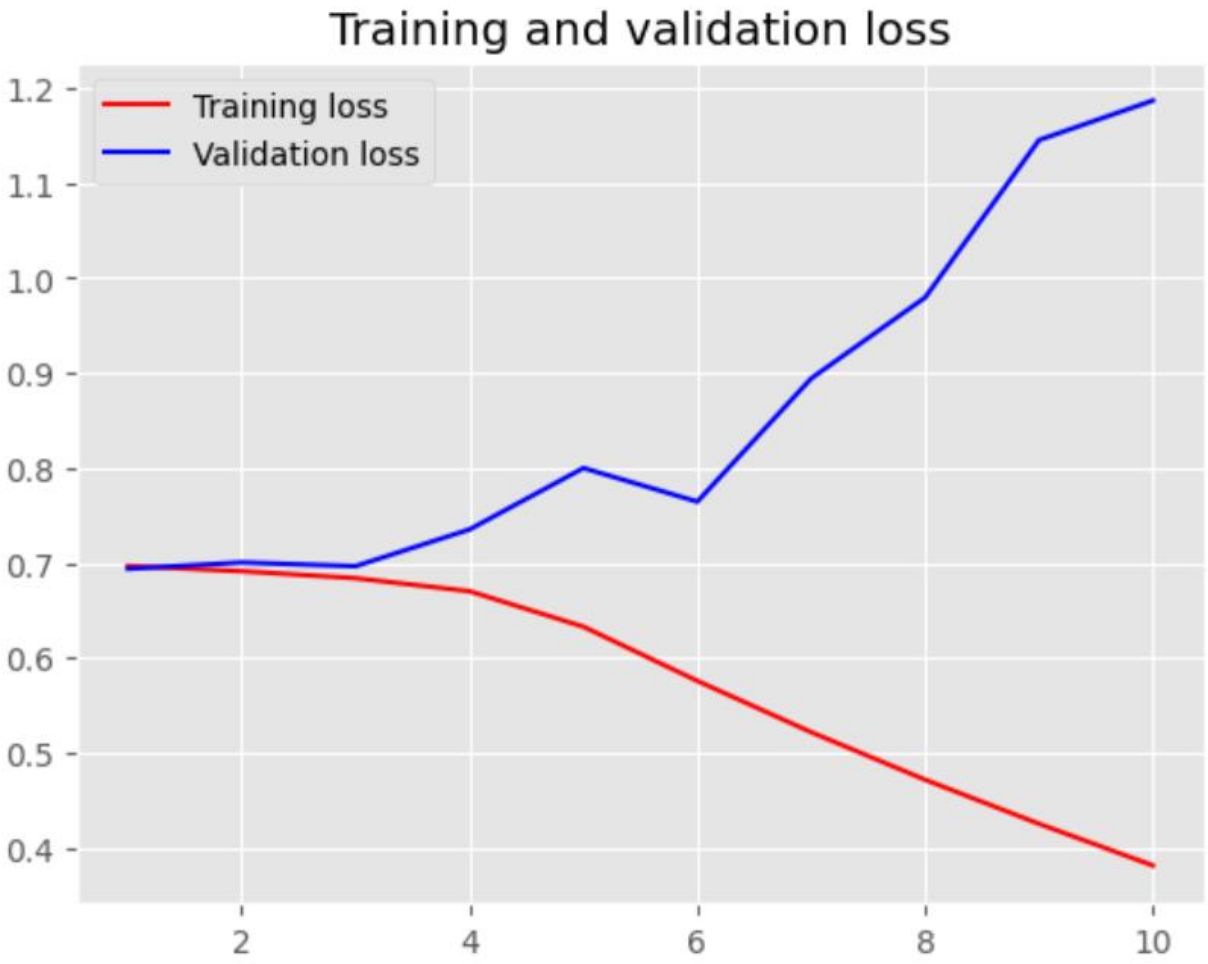
1. Pretrained word embedding layer with training sample size = 100


Training and validation accuracy
Training acc
Validation acc
0.80
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10

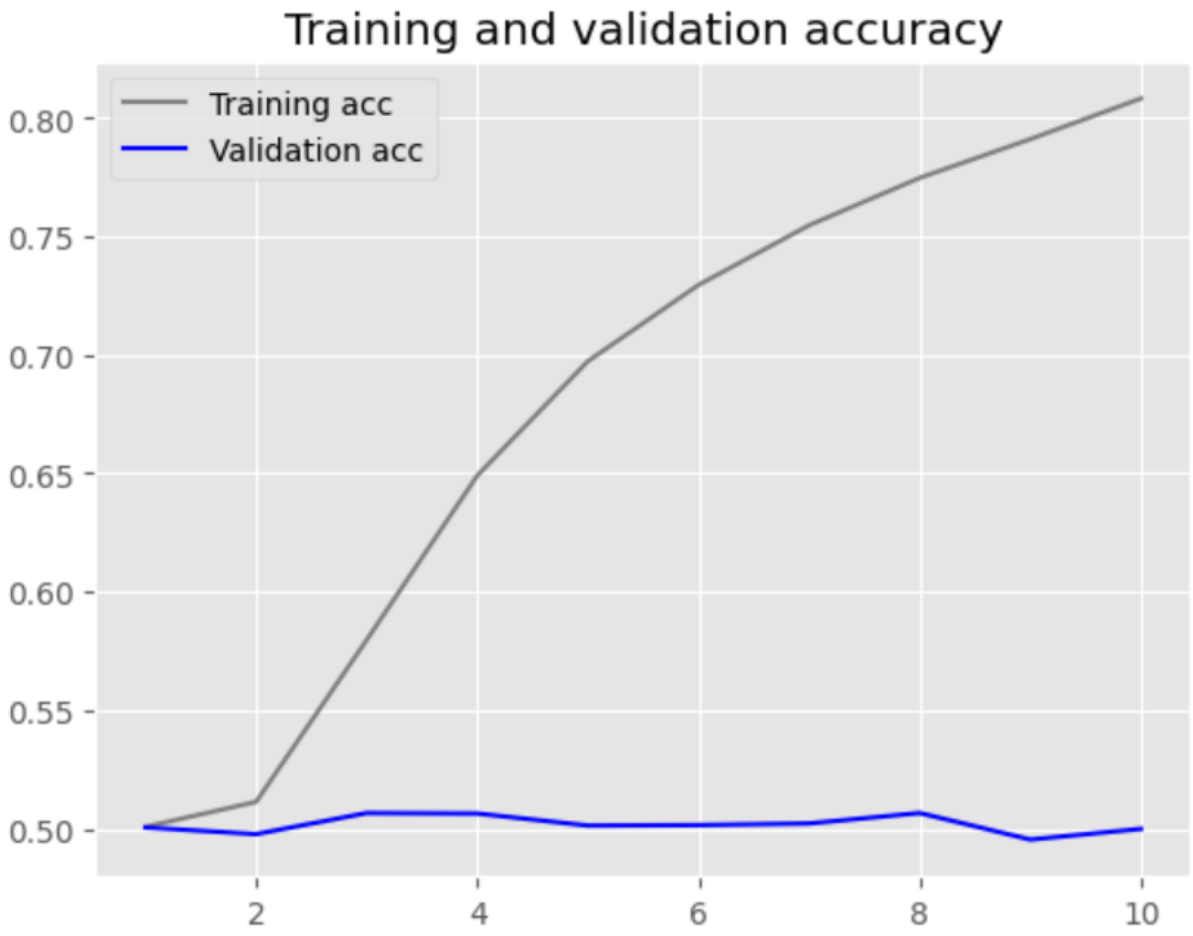


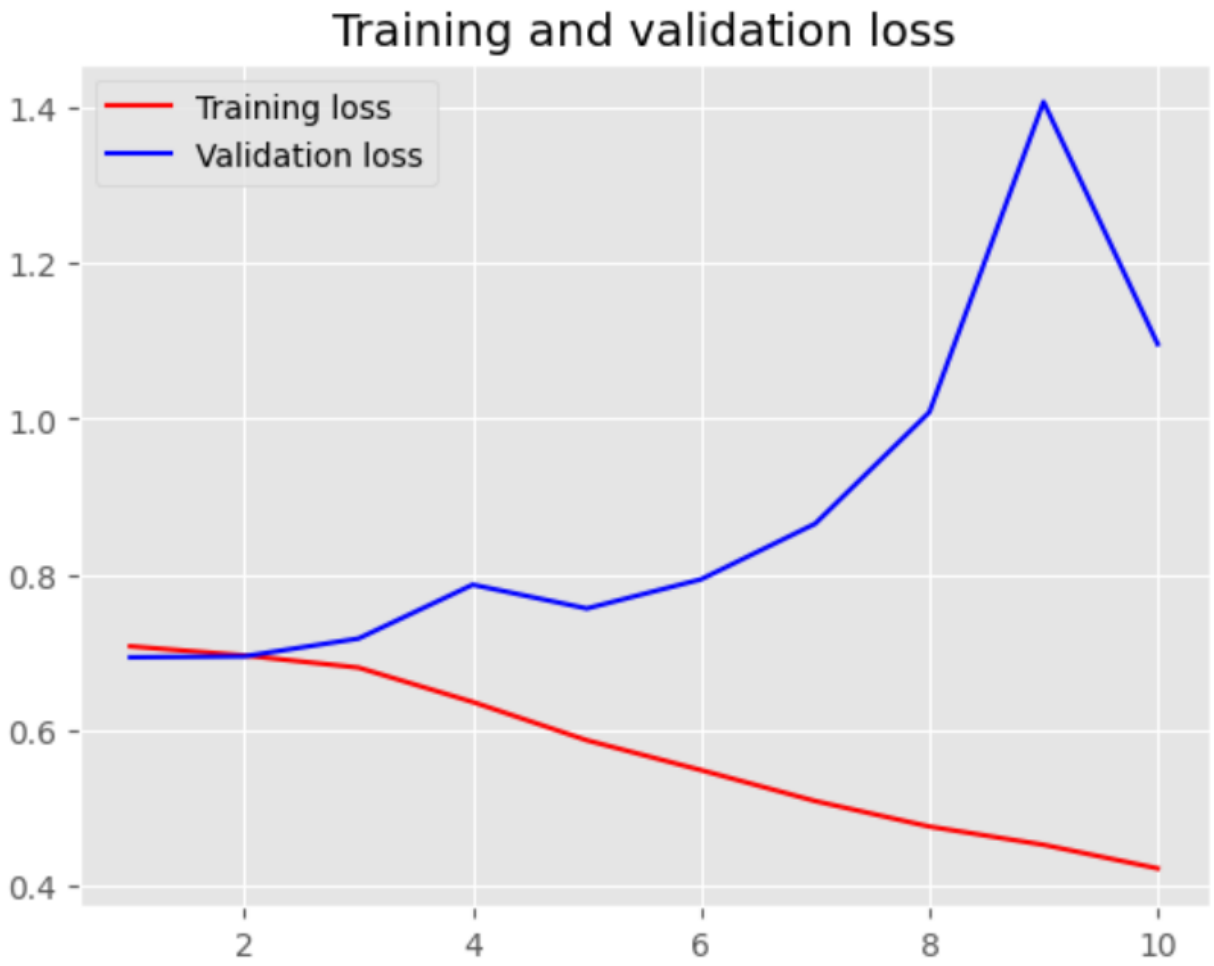
2. Pretrained word embedding layer with training sample size = 5000


Training and validation accuracy
Training acc
Validation acc
0.80
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10

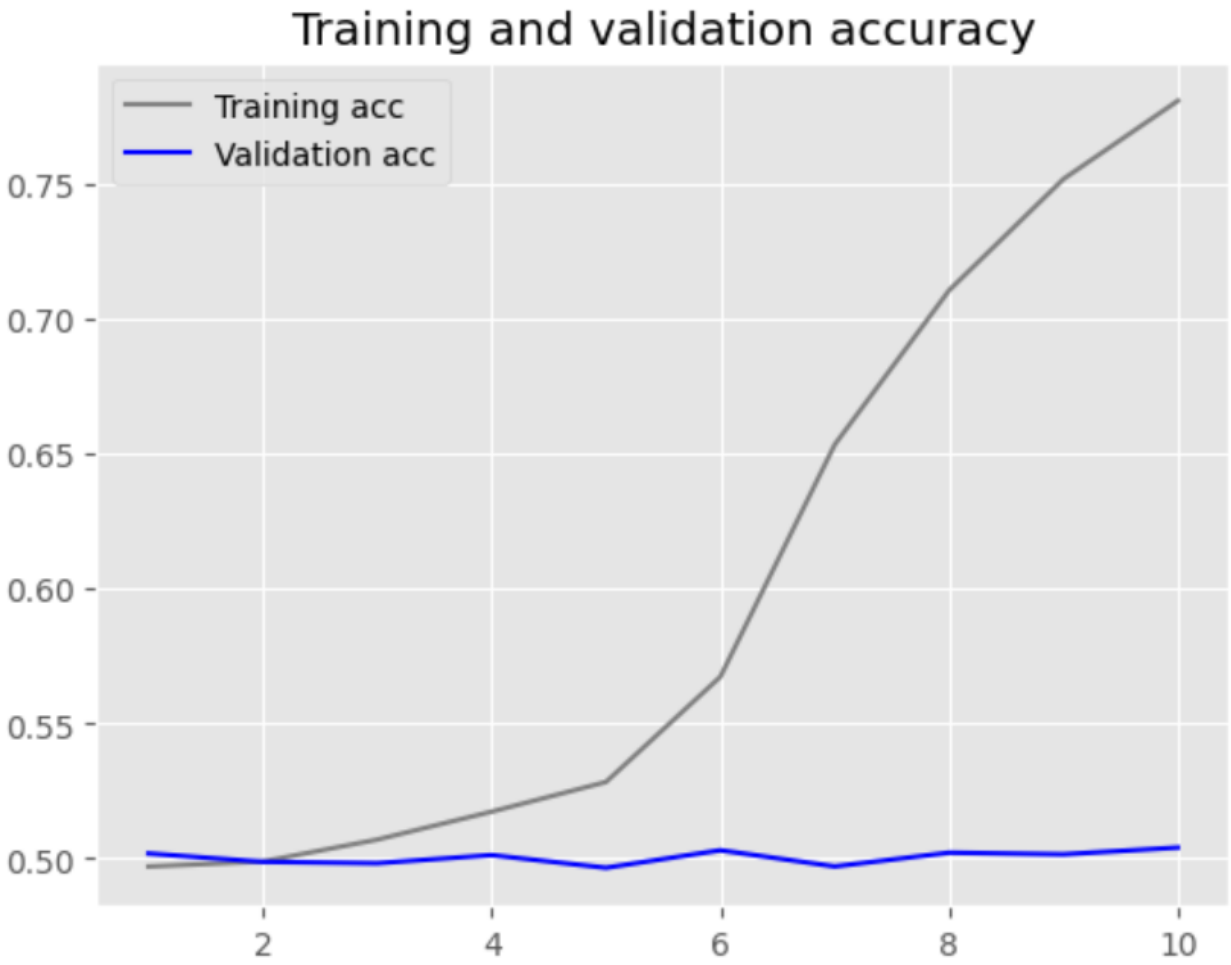


3. Pretrained word embedding layer with training sample size = 1000


Training and validation accuracy
Training acc
Validation acc
0.80
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10



4. Pretrained word embedding layer with training sample size = 10000


Training and validation accuracy
Training acc
Validation acc
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10

GloVe, a pretrained word embedding technique, achieves an accuracy range of 78.1% to 82.94AML%. Its accuracy peaks with 100 samples, but as the sample size increases, it tends to overfit and loses accuracy. The optimal strategy varies depending on task constraints, leading to some uncertainty.

RESULTS:

Embedding Technique	Training Sample Size	Training Accuracy (%)	Test loss
Custom trained embedding layer	100	96.26	0.34
Custom trained embedding layer	5000	96.39	0.34
Custom trained embedding layer	1000	96.65	0.34
Custom trained embedding layer	10000	96.65	0.35
Pretrained word embedding (GloVe)	100	82.94	1.13
Pretrained word embedding (GloVe)	5000	81.6	1.03
Pretrained word embedding (GloVe)	1000	80.83	0.92
Pretrained word embedding (GloVe)	10000	78.1	0.91

CONCLUSION:

The report aims to categorize movie reviews as positive or negative using the IMDB dataset. The dataset contains 50,000 reviews, and the task evaluates the top 10,000 words. Training samples were limited to sizes of 100, 5,000, 10,000, and 100,000, with validation on 10,000 samples. Two methods for creating word embeddings were tested: a custom-trained embedding layer and a pretrained GloVe model. The custom-trained embedding layer achieved higher accuracy,

ranging from 96.26% to 96.65%, peaking with 1000 and 10,000 samples. The GloVe model, however, ranged from 78.1% to 82.94%, with peak accuracy at 100 samples but exhibited overfitting with larger sample sizes. Overall, the custom-trained embeddings performed better, particularly with larger training samples, indicating their effectiveness over pretrained embeddings for this specific task.

In this experiment, the custom-trained embedding layer outperforms the pretrained word embedding layer when trained with more training samples.