CS6910: Programming Assignment 3 Report

Vimarsh Sathia
CS17B046

Indian Institute of Technology Madras — January 11, 2021

1 Part-A: Word2Vec embeddings

For the text8 dataset, embeddings were trained using the Continuous Bag of Words and the Skipgram training technique. 4 embeddings with the following details were trained, (also attached in the submission).

1. Continuous Bag of Words(CBOW)
   (a) 100 dimensions, context size 4, 5 epochs
   (b) 200 dimensions, context size 4, 5 epochs
2. Skipgram
   (a) 100 dimensions, context size 4, 5 epochs
   (b) 200 dimensions, context size 4, 5 epochs

In both training methods, the target word was in the middle of the context words.

2 Part-B: Sentiment Analysis

In this part, sentiment analysis on the Rotten Tomatoes dataset was carried out using pre-trained GloVe embeddings and the custom embeddings trained in section section 1 . The following GloVe embeddings were considered:

1. GloVe $6B$ tokens, 50 dimensions
2. GloVe $6B$ tokens, 100 dimensions
3. GloVe $6B$ tokens, 200 dimensions

table 1 summarizes the test accuracy achieved for each word embedding. Each LSTM model was trained for 10 epochs in this setup.

Table 1: Accuracy stats for every embedding after 10 epochs

Embedding	Dimensions	Train accuracy(%)	Test accuracy(%)
GloVe	50	63.461	63.108
	100	65.250	64.770
	200	**65.757**	**65.073**
CBOW	100	60.363	59.533
	200	64.166	62.822
Skipgram	100	60.275	59.016
	200	64.084	61.190

From table 1, we can see that maximum accuracy is achieved for the GloVe word embedding with 200 dimensions. However, since all accuracies are almost the same, we analyze classwise classification accuracy by viewing the confusion matrix for all different embeddings.

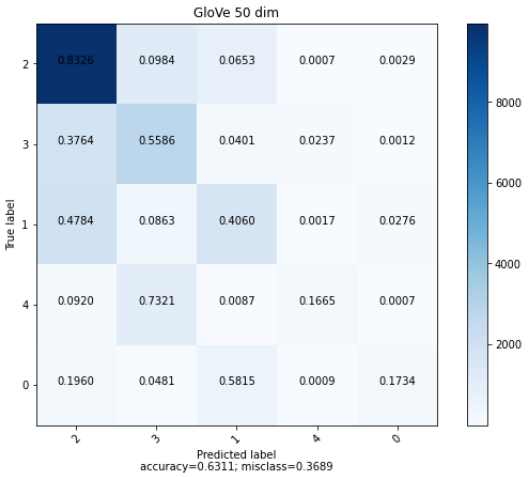

Figure 1: Confusion matrix for GloVe embedding with 50 dimensions

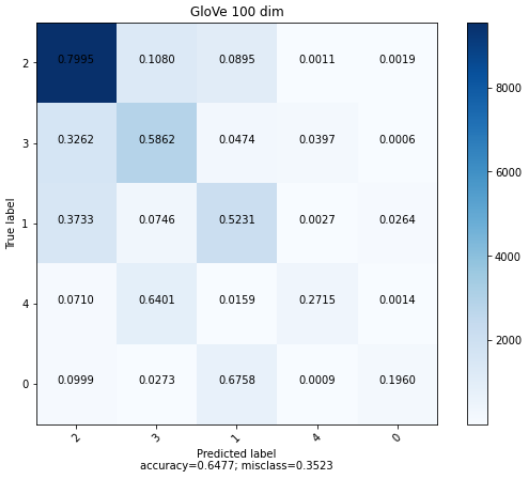

Figure 2: Confusion matrix for GloVe embedding with 100 dimensions

After analyzing the figures, we see that the Glove embeddings produce balanced output for all classes, compared to the CBOW and Skipgram embeddings. fig. 1 to fig. 7 show plots of the confusion matrices of every LSTM model with the appropriate embedding.

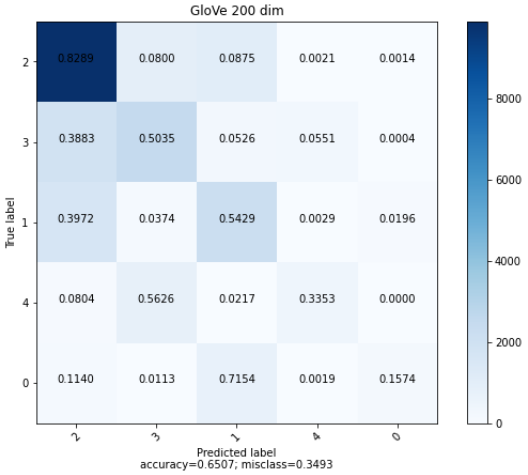

Figure 3: Confusion matrix for GloVe embedding with 200 dimensions

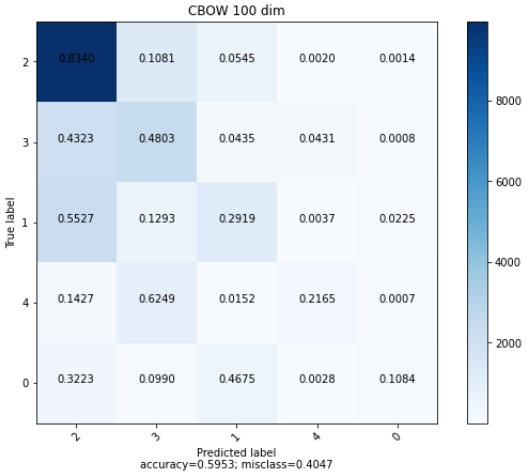

Figure 4: Confusion matrix for CBOW embedding with 100 dimensions

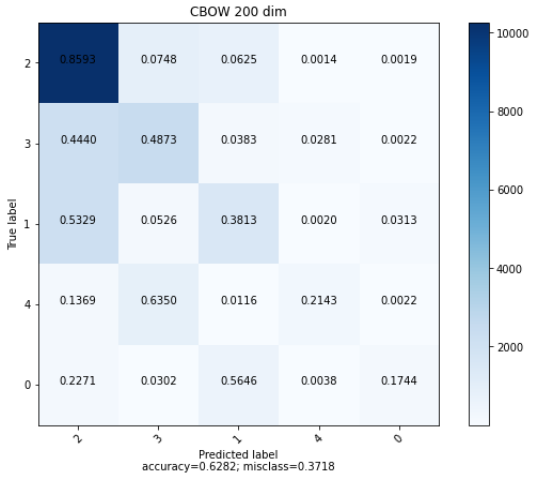

Figure 5: Confusion matrix for CBOW embedding with 200 dimensions

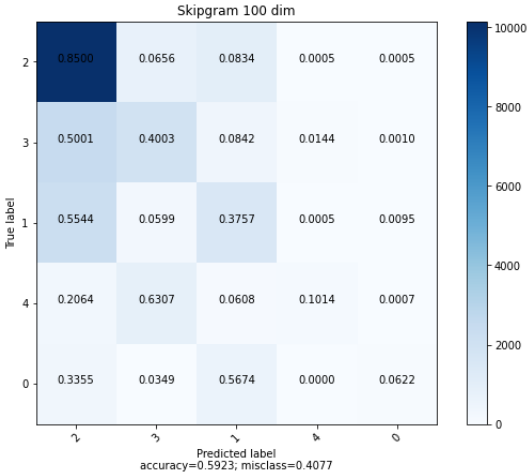

Figure 6: Confusion matrix for Skipgram embedding with 100 dimensions

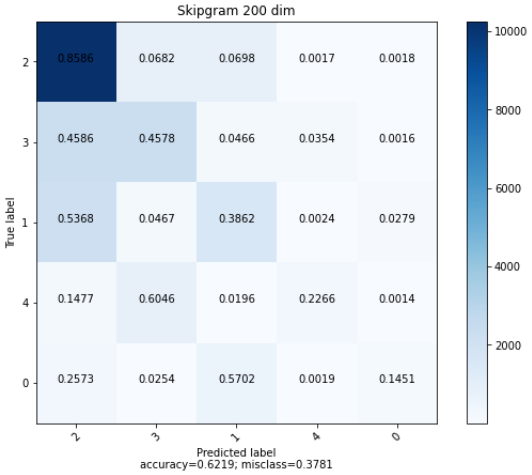

Figure 7: Confusion matrix for Skipgram embedding with 200 dimensions