iNLP Assignment-4 Report

Sk Abukhoyer

Datasets 1. Sentiment Analysis: Stanford Sentiment Treebank

We are asked the task, which involves sentiment analysis on the Stanford Sentiment Treebank dataset. It mentioned in the assignment pdf that I have to preprocess the data and convert the sentences to word embeddings using the ELMo model. However, when I tried to generate embeddings for the entire dataset, you encountered an error related to the shapes of the input arrays. I try padding the sentences with zeros to make them have the same length before generating their embeddings.
After generating the embeddings, I was asked about creating a classifier to train on the dataset. I suggested that you use PyTorch to create a classifier with a linear layer that takes in the batch embeddings of dimension 32x512 as input and outputs a probability distribution over the two classes.
I then trained the classifier, so I provided an example code snippet that uses cross-entropy loss and stochastic gradient descent to train the model. I used the trained model to predict the sentiment label for a new sentence by first converting the sentence to an embedding using the ELMo model and then passing it through the trained classifier.
Overall, our conversation covered the preprocessing of the Stanford Sentiment Treebank dataset, the use of the ELMo model to generate sentence embeddings, the creation and training of a PyTorch classifier on the embeddings, and the prediction of sentiment labels on new data.

Below is given the classification report for this above dataset:

	precision	recall	f1-score	support
0	0.5136	0.7244	0.6011	1143
1	0.4732	0.2652	0.3399	1067
accuracy			0.5027	2210
macro avg	0.4934	0.4948	0.4705	2210
weighted avg	0.4941	0.5027	0.4750	2210

And the ROC graph shown below: