Text Classification using Word2Vec and RNN

N Surya Prakash Reddy(20CS10038)

1 Introduction

The objective of this project is to classify text data using Word2Vec and neural networks.

2 Data Cleaning

For cleaning the data, we have some observations:

- There are some unicode decimal characters like #8217, #36, #147, etc.
- Several html tags were in the form of &lt;Ahyperlink&gt;, etc.

These garbage values were removed along with not alphanumeric, non whitespace characters. Stop words have also been removed.

3 Tokenization and data loading

The sentences remaining after the cleaning have been tokenized using the word_tokenize tool. The vocabulary is computed from these tokenized forms and the embeddings for each word is computed through the Word2Vec tool.

Additionally, each sentence is converted to the average embedding of all its words for the neural network part.

4 Simple Neural Network

In this simple neural network, we add the layers as a Linear layer followed by a ReLU layer.

Out of three different models chosen with sizes (64), (128) and (64, 32), The (64) model gave highest accuracy of 0.57 on the validation set.

The accuracy and f1-score for the test set were 0.576, 0.581.

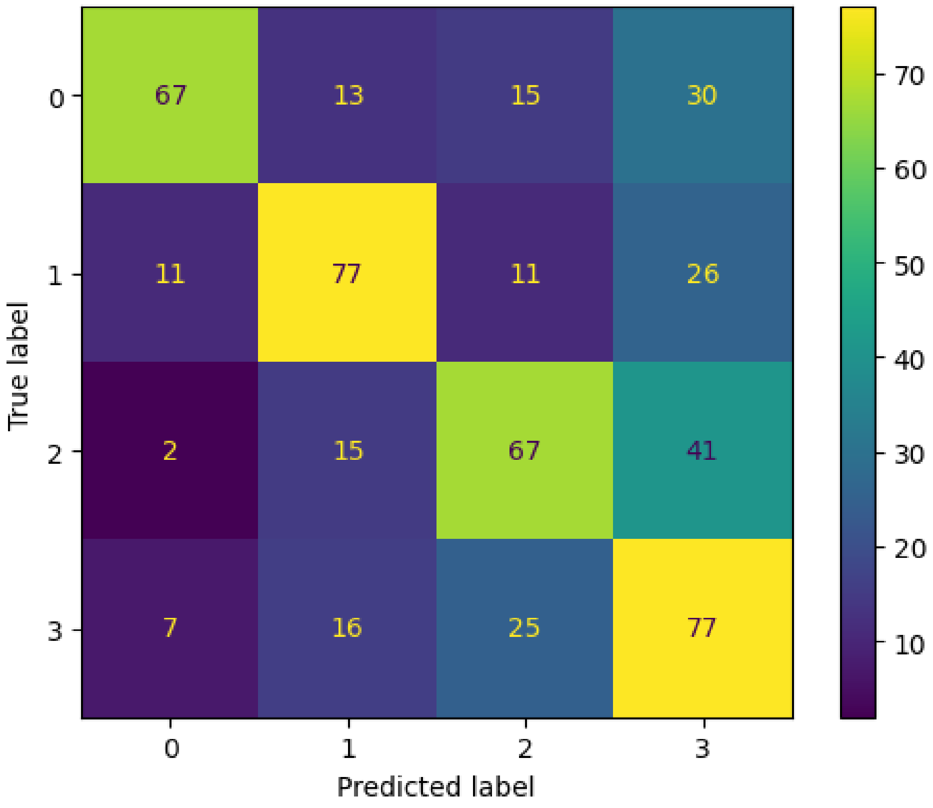

Figure 1: Confusion Matrix for word2vec

5 Recurrent Neural Network

For this we first pad and truncate the sentences to a max sequence length. Here we chose the sequence length of 50 since the maximum number of tokens in a line were 108 and not many lines have that many tokens, so half of it was chosen. Then we add an Embeddings layer to the nn.Module so that we can give indices and get the embeddings for the words in a padded sentence.

5.1 Simple RNN

Out of two different models chosen with sizes (32), (64), The (32) model gave highest accuracy of 0.77 on the validation set.

The accuracy and f1 score for the test set were 0.796, 0.797.

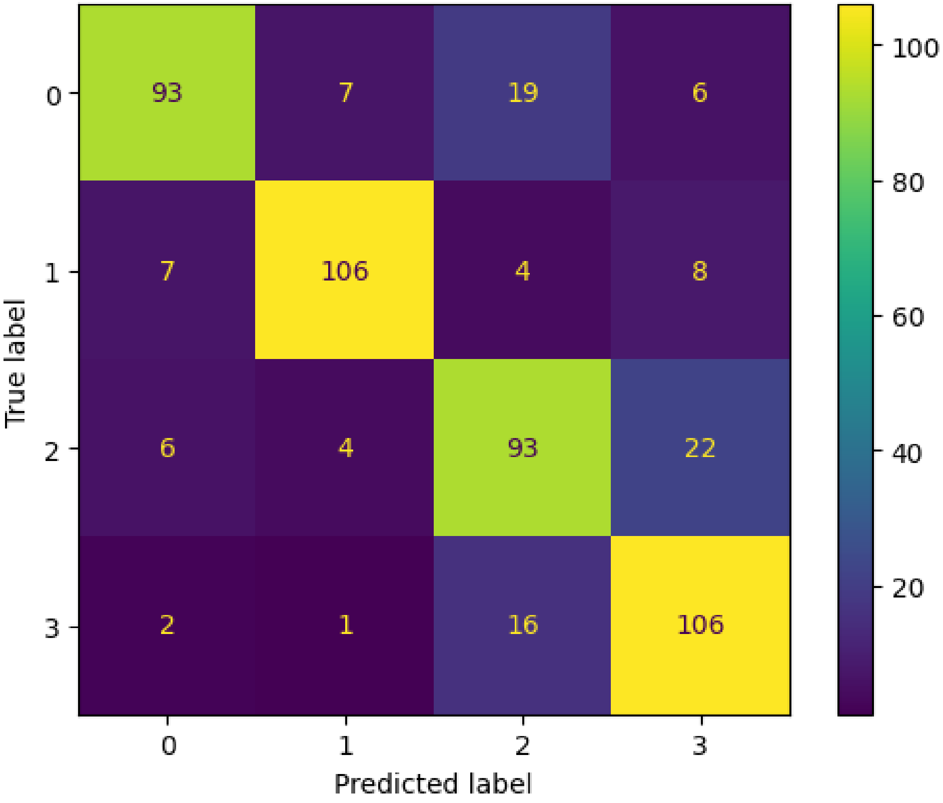

Figure 2: Confusion Matrix for RNN

5.2 LSTM Model

Out of two different models chosen with sizes (32), (64), The (32) model gave highest accuracy of 0.754 on the validation set.

The accuracy for the test set was 0.788, 0.786.

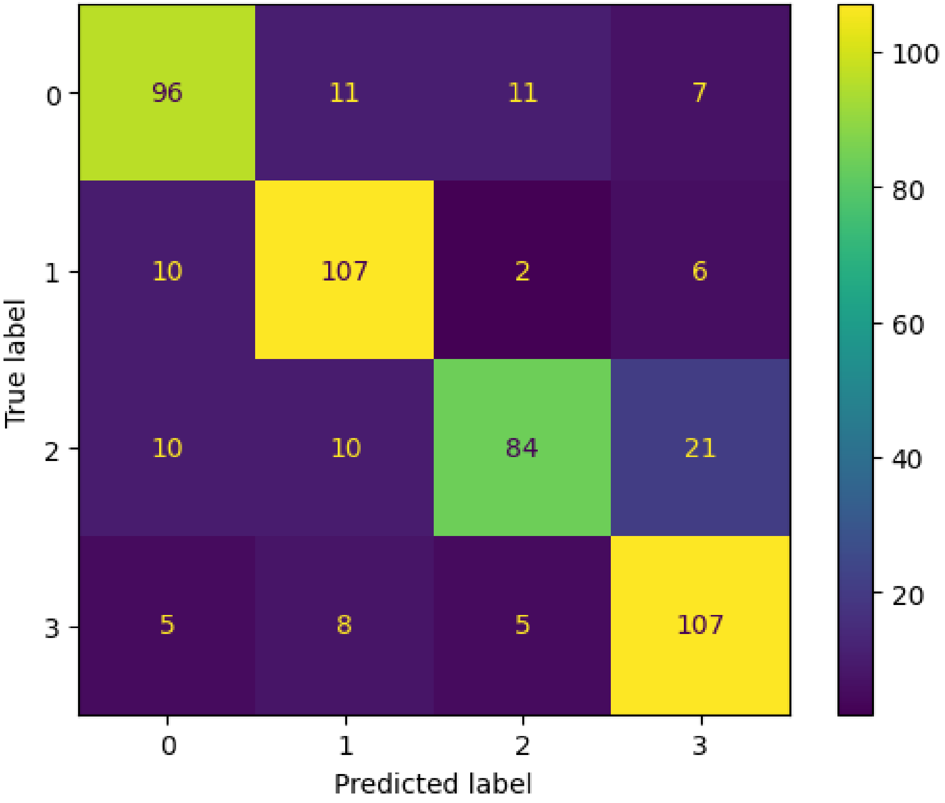

Figure 3: Confusion Matrix for LSTM