Result Analysis

Details Regarding Experiment:

- Please run the Run.py file to execute the experiments
- OOV's are not included during training
- Used learning rate is 0.001
- I have calculated accuray scores on the development data in every epoch and stored the model with the best accuracy score on that data. Then during testing I used the stored model.

The sequence tagger model is developed using the given data set. Both macro and micro F1 scores of development data for every epoch are reported in table and in graph. The analysis is given below:

Number of Epochs	Macro F1 Score	Micro F1 score
1	0.37223	0.66071
2	0.40002	0.68354
3	0.41456	0.700504
4	0.41886	0.70328
5	0.42803	0.70988
6	0.42401	0.70271
7	0.42472	0.70233
8	0.436505	0.71666
9	0.435907	0.711914
10	0.43499	0.71024
11	0.43431	0.7093
12	0.44096	0.71507
13	0.44098	0.71263
14	0.44145	0.71272
15	0.441084	0.71231
16	0.44445	0.71734
17	0.446287	0.71744
18	0.4469	0.72185
19	0.44083	0.71145
20	0.440564	0.71634
For Test set	0.44767	0.72233

Table 1: F1 macro score and F1 micro score for all epochs

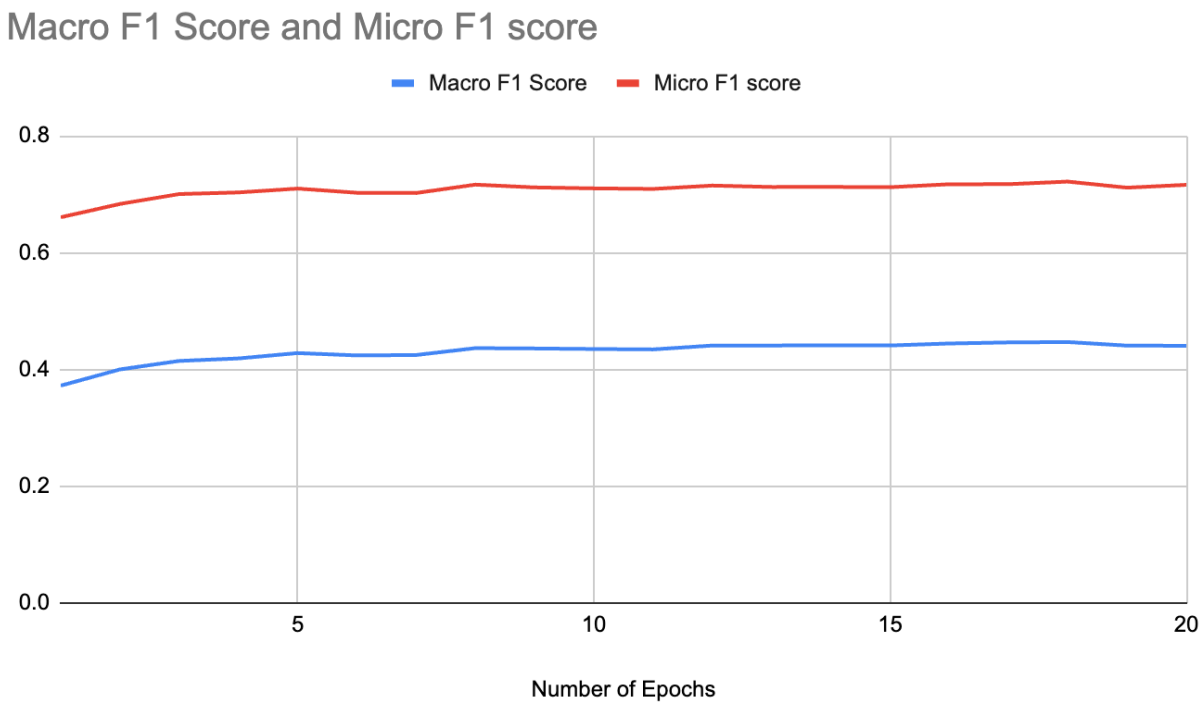

Figure 1: Graph of of the above table with macro and micro F1 scores per epoch

1) Here we have multiple labels or classes therefore we calculated macro-averaged and micro-averaged F1 scores for all classes. Here we can observe that the macro-averaged score per epoch is significantly lower than micro-averaged F1 scores. As we know, macro averaged F1 score is the unweighted mean of F1 score per class whereas micro averaged F1 score is calculated using observations of total number of classes. So, the micro F1 score gives more emphasis on each observation instead of each class. In case of imbalanced data, labels with higher occurrences have a larger effect on micro F1 score, see figure 1. On the other hand, macro F1 score gives equal importance to all classes meaning labels with low occurrences have the same effect as high occurrences

labels. Hence, micro F1 scores are higher than macro F1 scores.

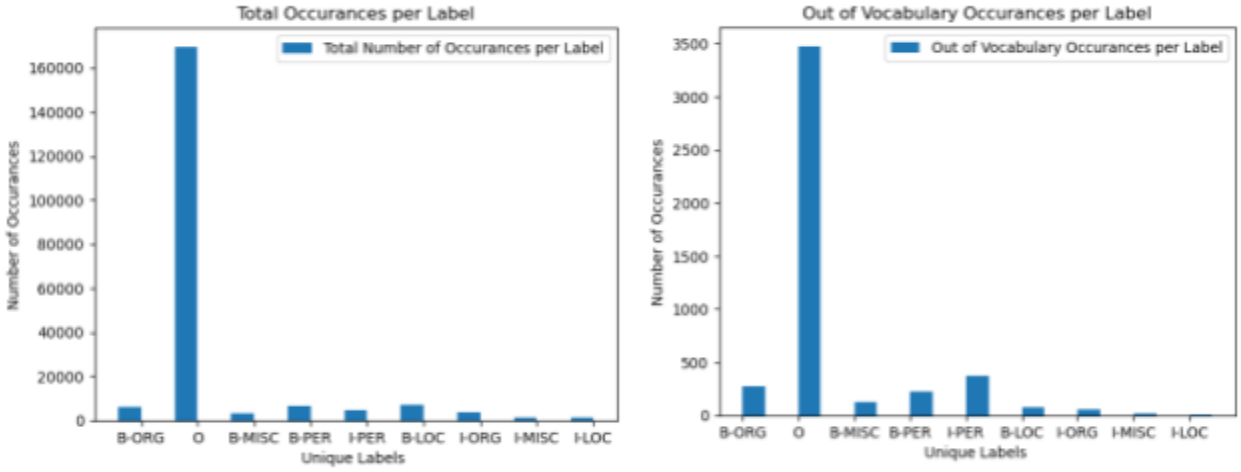

Figure 2: Graph for comparison between total words and out of vocabulary words

As we can see in figure 2 that out of 9 labels, the words labeled with “O” were found approximately 160000 times in train data compared to other labeled words. Moreover, out of vocabulary occurrences of words labeled “O” is also higher than other labeled words. As a result we can conclude that the given dataset is a highly imbalanced dataset with “O” labeled words majority.

I would prefer the macro F1 score for this multi labeled imbalanced dataset. Because macro F1 is calculated per label and all labels (higher or lower occurrences) are given equal importance.

2) The model fails to generate correct predictions mostly in case of “O” labeled words, see figure 3.

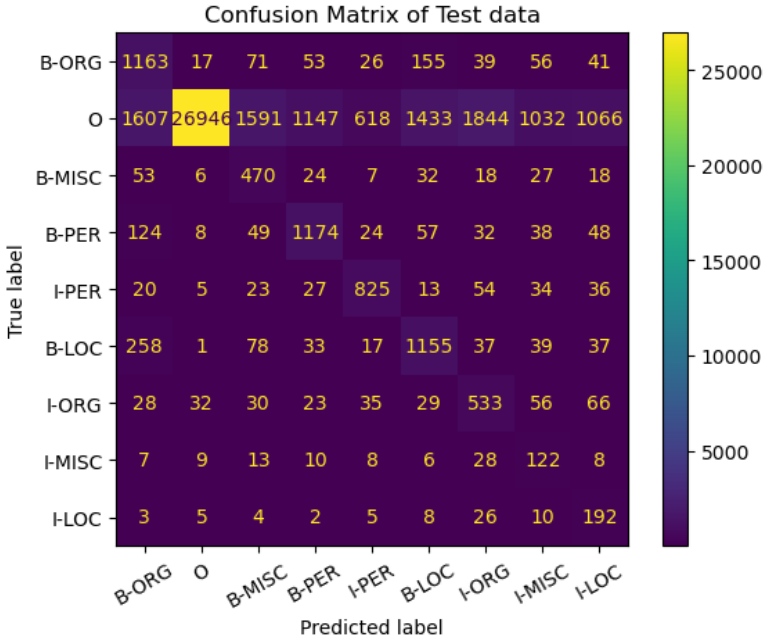

Figure 3 : Confusion Matrix

From the confusion matrix in Figure 3, it is evident that various wrong predictions are produced by the model mostly in the case of “O”,”B-ORG”, “B-LOC”, “I-ORG” etc. Even in case of low occurrences labels such as “I-MISC”, “I-LOC”, the model generated 89 and 63 wrong predictions respectively. However, the correct predictions (true positive values) can be seen diagonally.

3) There are a few problems that cause the error. As we can see from Figure 2, the dataset is highly imbalanced. Also we used batch size value 1 instead of mini batch. Setting the batch size to 1 might cause the learning curve of the model to oscillate.

   In case of imbalanced dataset, the suggestions to improve the model can be
   - ➔ Doing under-sampling of the data. In this way labels with higher occurrences can not effect low occurrences labels
   - ➔ Using weighted cross entropy to reduce the impact of higher occurrences of labels
   - ➔ Hyper-parameter tuning. Hyper-parameter such as batch size, learning rate, number of Bilstm layer and hidden size can be tuned to find the best results.