4. Analysis

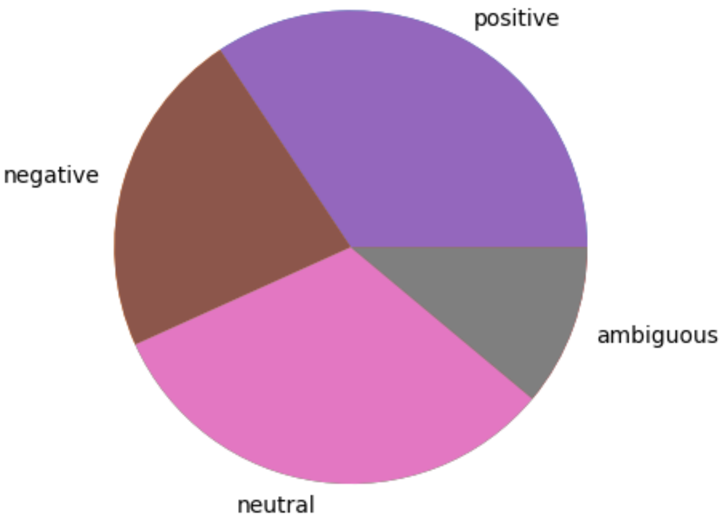


Figure 1: Pie Chart for the different Sentiments

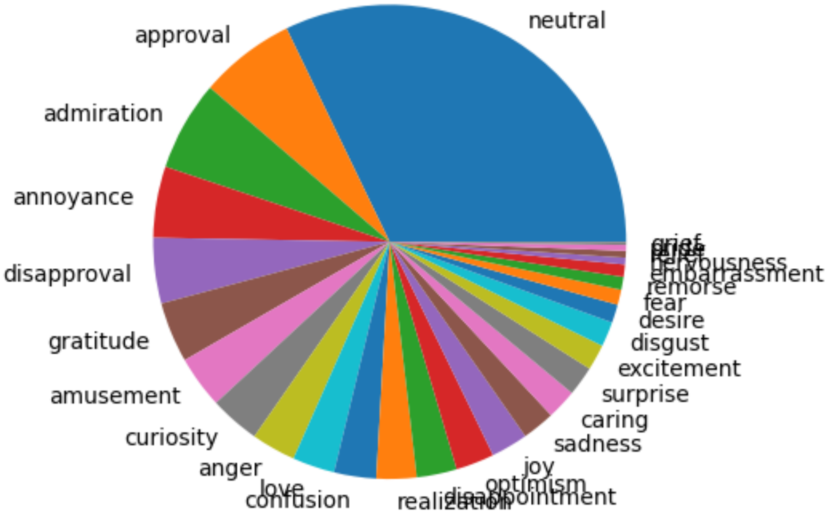
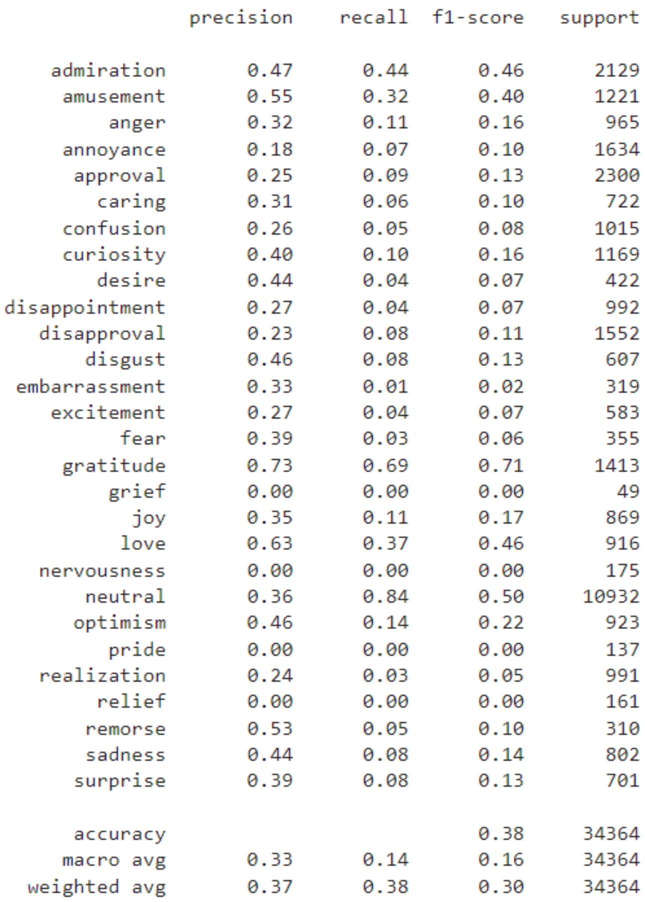


Figure 2: Pie Chart for the different emotions

4.1: By looking at the dataset given on Moodle and then visualizing it using Jupyter Notebook, we can have a better idea on the topic of classification. The plot function was used from the matplotlib library to print both the emotion classification and sentiment classifications. Furthermore, they were displayed in two different pie charts, one for the different type of emotions and one for the different type of sentiments. Figure 1 shows the sentiment pie chart with the 4 different types of sentiments. Those are the negative, positive, ambiguous and neutral classes. From the looks of it, the 4 types of sentiments seem to be more or less in equal amounts other than the ambiguous category. For this reason, an accuracy metric should be a good metric to use because the dataset is slightly balanced. On the other hand, Figure 2 shows the different category of emotions. We won't be naming them all of them because there are 28 different classes, but here are some of them: approval, admiration, curiosity… When it comes to the emotions pie chart, we can see that it is not a balanced dataset, and this is because the neutral class is present in a greater amount than the rest. This would mean that we can't use the accuracy metric to evaluate the performance of our classifiers. The reasoning behind why accuracy can't be used when the dataset is not balanced is a very logical explanation. As described in class, if one category is for example 99% of the dataset whereas the rest are just 1%, if our system just guesses that one category for all the options, then we would have 99% accuracy which is not a good way to view it because all the system was doing was just outputting that one value whereas it was not actually determining it by analysing anything. For this case, since accuracy won't be a good measure to evaluate the performance of the emotion's classifier, we would need to look at the other metrics. Moreover, there is the recall and precision metrics that can be looked at. However, as seen in class, depending on the different situations, precision or recall might be preferable. Hence, the weighted harmonic mean, also known as the F measure, can be calculated. When it comes to this project, the F1 measure was calculated for all the different models. The F1 measure means that our beta value is of 1 and thus the precision and recall have the same importance as we learned in class. In addition, since we have many classes that we are interested in and not just one, we let Jupyter Notebook also calculate the macro average and weighted average of F1. This means we combine the F1 measure values into one number to then be able to compare the different models together. Furthermore, when it comes to the different models to use for the emotion or sentiment categories, the general idea we have is that the Multi-Layered Perceptron will take the longest time, and this is due to the fact that it has to go through many epochs since we've determined through Jupyter Notebook that the vocabular is very large with a size of 30449 words. Thus, it must forward propagate and back propagate many times and adjust the weights accordingly. In comparison to the Naïve Bayes which is just computing probabilities which will be quicker. Lastly, for the decision dree, it is computing information gain and then creating a tree, which should make it faster than MLP but slower than Naïve Bayes.

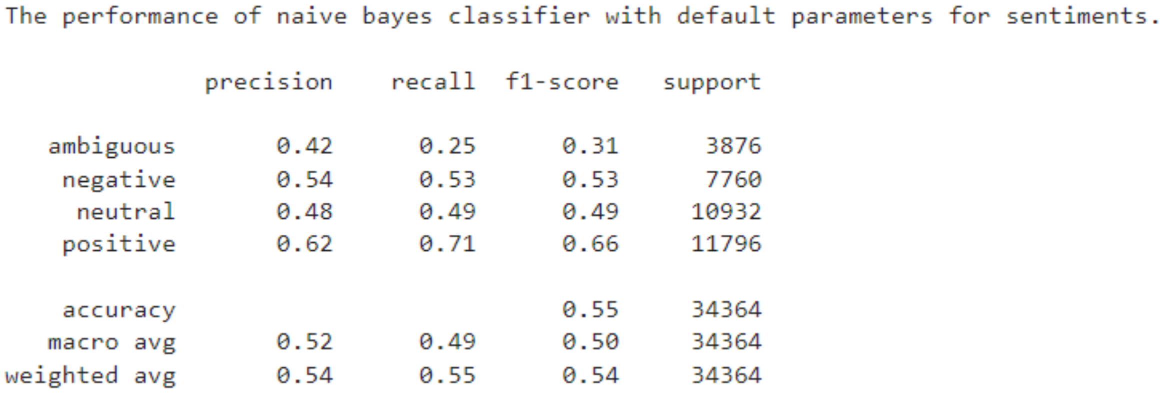
4.2: Part 2 Figures

Since the figures will be used for all of Part 2, we decided to put them at the beginning with the corresponding Figure numbers and throughout the Part 2 Analysis they will be referenced.



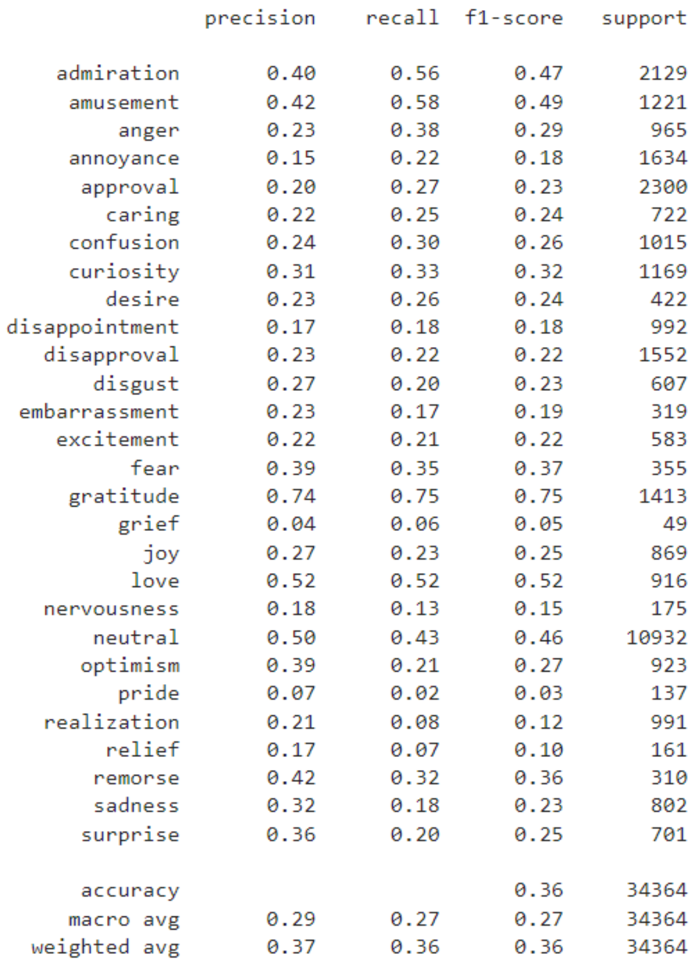
```
                precision    recall  f1-score   support

      admiration       0.47      0.44      0.46      2129
       amusement       0.55      0.32      0.40      1221
           anger       0.32      0.11      0.16       965
       annoyance       0.18      0.07      0.10      1634
        approval       0.25      0.09      0.13      2300
          caring       0.31      0.06      0.10       722
       confusion       0.26      0.05      0.08      1015
       curiosity       0.40      0.10      0.16      1169
          desire       0.44      0.04      0.07       422
  disappointment       0.27      0.04      0.07       992
     disapproval       0.23      0.08      0.11      1552
         disgust       0.46      0.08      0.13       607
   embarrassment       0.33      0.01      0.02       319
      excitement       0.27      0.04      0.07       583
            fear       0.39      0.03      0.06       355
       gratitude       0.73      0.69      0.71      1413
           grief       0.00      0.00      0.00        49
             joy       0.35      0.11      0.17       869
            love       0.63      0.37      0.46       916
     nervousness       0.00      0.00      0.00       175
         neutral       0.36      0.84      0.50     10932
        optimism       0.46      0.14      0.22       923
           pride       0.00      0.00      0.00       137
     realization       0.24      0.03      0.05       991
          relief       0.00      0.00      0.00       161
         remorse       0.53      0.05      0.10       310
         sadness       0.44      0.08      0.14       802
        surprise       0.39      0.08      0.13       701

        accuracy                           0.38     34364
       macro avg       0.33      0.14      0.16     34364
    weighted avg       0.37      0.38      0.30     34364
```

Figure 3: Base-MNB Classification Report (Emotion)



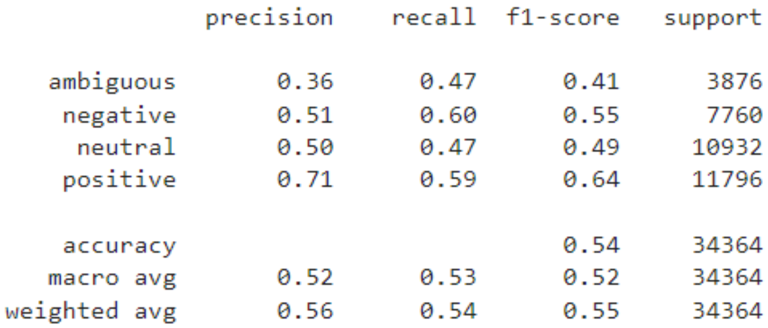
```
The performance of naive bayes classifier with default parameters for sentiments.

              precision    recall  f1-score   support

   ambiguous       0.42      0.25      0.31      3876
    negative       0.54      0.53      0.53      7760
     neutral       0.48      0.49      0.49     10932
    positive       0.62      0.71      0.66     11796

    accuracy                           0.55     34364
   macro avg       0.52      0.49      0.50     34364
weighted avg       0.54      0.55      0.54     34364
```

Figure 4: Base-MNB Classification Report (Sentiment)



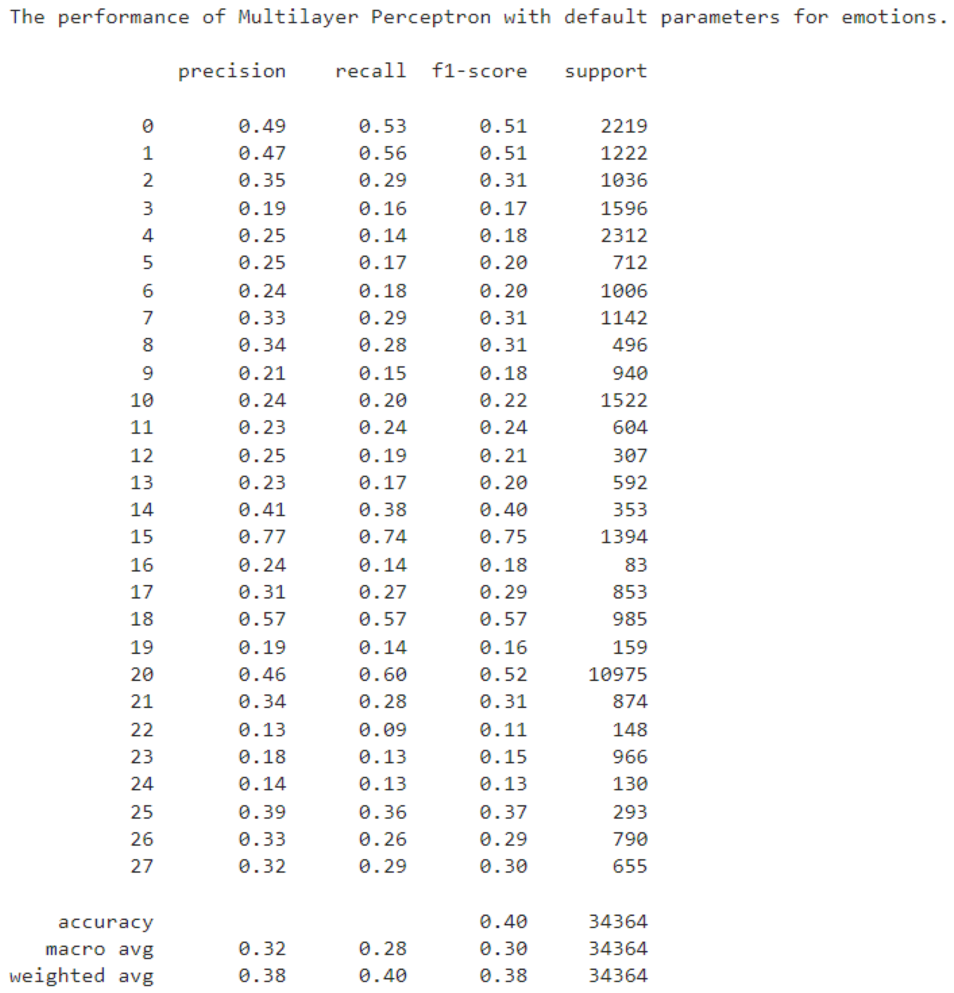
```
                precision    recall  f1-score   support

    admiration       0.40      0.56      0.47      2129
     amusement       0.42      0.58      0.49      1221
         anger       0.23      0.38      0.29       965
     annoyance       0.15      0.22      0.18      1634
      approval       0.20      0.27      0.23      2300
        caring       0.22      0.25      0.24       722
     confusion       0.24      0.30      0.26      1015
     curiosity       0.31      0.33      0.32      1169
        desire       0.23      0.26      0.24       422
disappointment       0.17      0.18      0.18       992
   disapproval       0.23      0.22      0.22      1552
       disgust       0.27      0.20      0.23       607
 embarrassment       0.23      0.17      0.19       319
    excitement       0.22      0.21      0.22       583
          fear       0.39      0.35      0.37       355
     gratitude       0.74      0.75      0.75      1413
         grief       0.04      0.06      0.05        49
           joy       0.27      0.23      0.25       869
          love       0.52      0.52      0.52       916
   nervousness       0.18      0.13      0.15       175
       neutral       0.50      0.43      0.46     10932
      optimism       0.39      0.21      0.27       923
         pride       0.07      0.02      0.03       137
   realization       0.21      0.08      0.12       991
        relief       0.17      0.07      0.10       161
       remorse       0.42      0.32      0.36       310
       sadness       0.32      0.18      0.23       802
      surprise       0.36      0.20      0.25       701

      accuracy                           0.36     34364
     macro avg       0.29      0.27      0.27     34364
  weighted avg       0.37      0.36      0.36     34364
```

Figure 5: Base-DT Classification Report (Emotion)



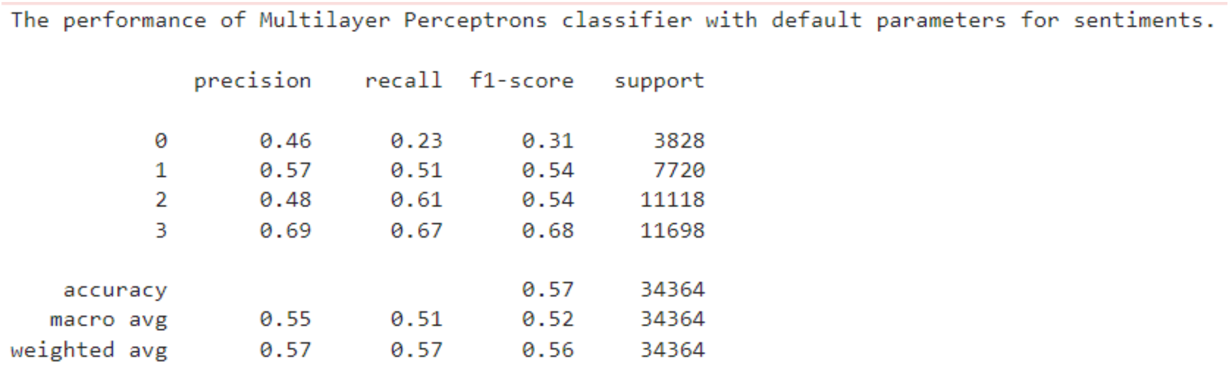
```
              precision    recall  f1-score   support

   ambiguous       0.36      0.47      0.41      3876
    negative       0.51      0.60      0.55      7760
     neutral       0.50      0.47      0.49     10932
    positive       0.71      0.59      0.64     11796

    accuracy                           0.54     34364
   macro avg       0.52      0.53      0.52     34364
weighted avg       0.56      0.54      0.55     34364
```

Figure 6: Base-DT Classification Report (Sentiment)



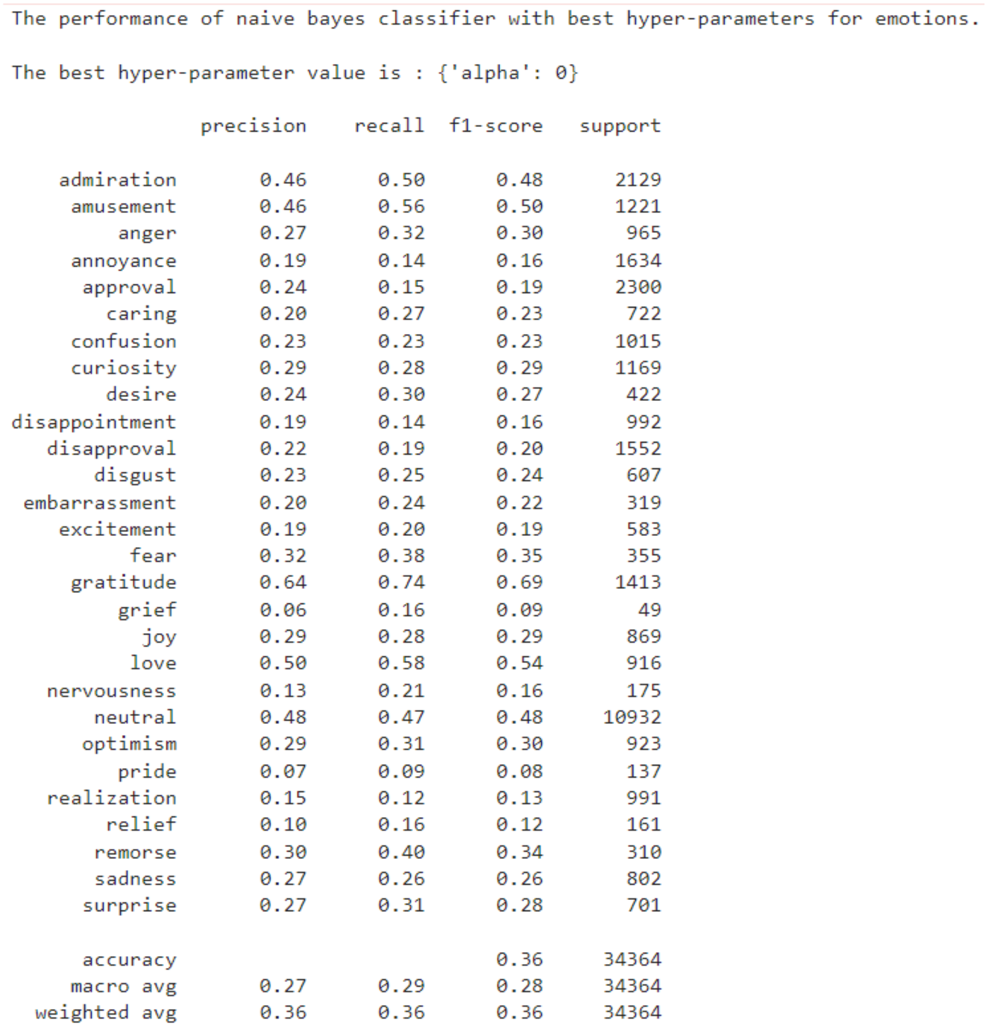
```
The performance of Multilayer Perceptron with default parameters for emotions.

              precision    recall  f1-score   support

           0       0.49      0.53      0.51      2219
           1       0.47      0.56      0.51      1222
           2       0.35      0.29      0.31      1036
           3       0.19      0.16      0.17      1596
           4       0.25      0.14      0.18      2312
           5       0.25      0.17      0.20       712
           6       0.24      0.18      0.20      1006
           7       0.33      0.29      0.31      1142
           8       0.34      0.28      0.31       496
           9       0.21      0.15      0.18       940
          10       0.24      0.20      0.22      1522
          11       0.23      0.24      0.24       604
          12       0.25      0.19      0.21       307
          13       0.23      0.17      0.20       592
          14       0.41      0.38      0.40       353
          15       0.77      0.74      0.75      1394
          16       0.24      0.14      0.18        83
          17       0.31      0.27      0.29       853
          18       0.57      0.57      0.57       985
          19       0.19      0.14      0.16       159
          20       0.46      0.60      0.52     10975
          21       0.34      0.28      0.31       874
          22       0.13      0.09      0.11       148
          23       0.18      0.13      0.15       966
          24       0.14      0.13      0.13       130
          25       0.39      0.36      0.37       293
          26       0.33      0.26      0.29       790
          27       0.32      0.29      0.30       655

    accuracy                           0.40     34364
   macro avg       0.32      0.28      0.30     34364
weighted avg       0.38      0.40      0.38     34364
```

Figure 7: Base-MLP Classification Report (Emotion)



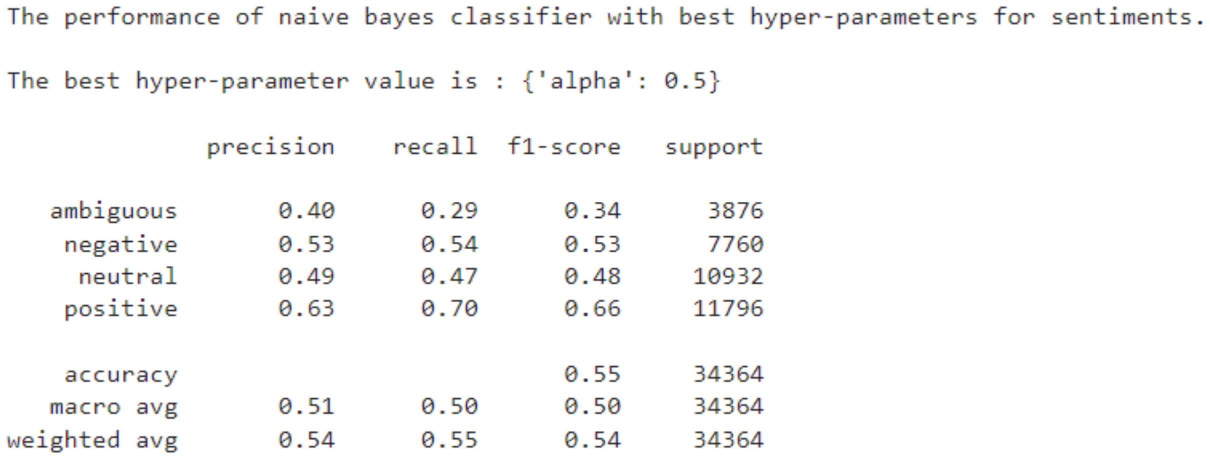
```
The performance of Multilayer Perceptrons classifier with default parameters for sentiments.

              precision    recall  f1-score   support

           0       0.46      0.23      0.31      3828
           1       0.57      0.51      0.54      7720
           2       0.48      0.61      0.54     11118
           3       0.69      0.67      0.68     11698

    accuracy                           0.57     34364
   macro avg       0.55      0.51      0.52     34364
weighted avg       0.57      0.57      0.56     34364
```

Figure 8 : Base-MLP Classification Report (Sentiment)



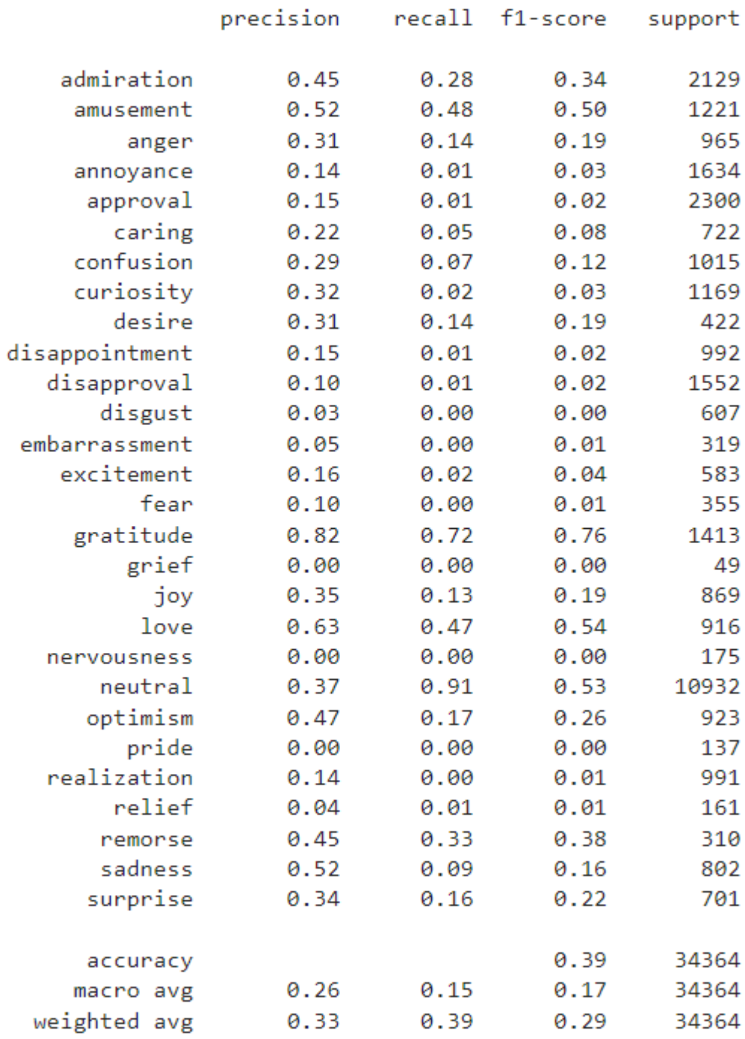
```
The performance of naive bayes classifier with best hyper-parameters for emotions.

The best hyper-parameter value is : {'alpha': 0}

                precision    recall  f1-score   support

    admiration       0.46      0.50      0.48      2129
     amusement       0.46      0.56      0.50      1221
         anger       0.27      0.32      0.30       965
     annoyance       0.19      0.14      0.16      1634
      approval       0.24      0.15      0.19      2300
        caring       0.20      0.27      0.23       722
     confusion       0.23      0.23      0.23      1015
     curiosity       0.29      0.28      0.29      1169
        desire       0.24      0.30      0.27       422
disappointment       0.19      0.14      0.16       992
   disapproval       0.22      0.19      0.20      1552
       disgust       0.23      0.25      0.24       607
 embarrassment       0.20      0.24      0.22       319
    excitement       0.19      0.20      0.19       583
          fear       0.32      0.38      0.35       355
     gratitude       0.64      0.74      0.69      1413
         grief       0.06      0.16      0.09        49
           joy       0.29      0.28      0.29       869
          love       0.50      0.58      0.54       916
   nervousness       0.13      0.21      0.16       175
       neutral       0.48      0.47      0.48     10932
      optimism       0.29      0.31      0.30       923
         pride       0.07      0.09      0.08       137
   realization       0.15      0.12      0.13       991
        relief       0.10      0.16      0.12       161
       remorse       0.30      0.40      0.34       310
       sadness       0.27      0.26      0.26       802
      surprise       0.27      0.31      0.28       701

      accuracy                           0.36     34364
     macro avg       0.27      0.29      0.28     34364
  weighted avg       0.36      0.36      0.36     34364
```

Figure 9: Top-MNB Classification Report (Emotion)



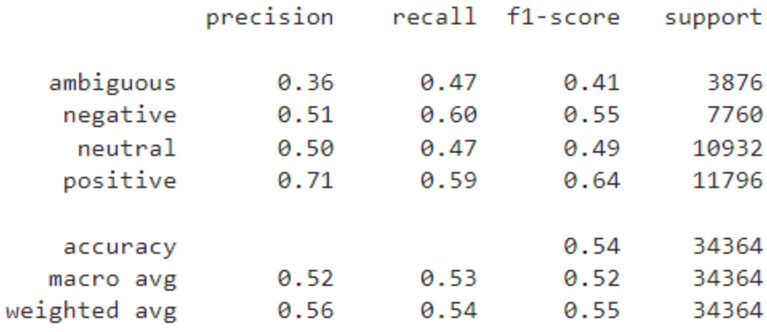
```
The performance of naive bayes classifier with best hyper-parameters for sentiments.

The best hyper-parameter value is : {'alpha': 0.5}

              precision    recall  f1-score   support

   ambiguous       0.40      0.29      0.34      3876
    negative       0.53      0.54      0.53      7760
     neutral       0.49      0.47      0.48     10932
    positive       0.63      0.70      0.66     11796

    accuracy                           0.55     34364
   macro avg       0.51      0.50      0.50     34364
weighted avg       0.54      0.55      0.54     34364
```

Figure 10: Top-MNB Classification Report (Sentiment)



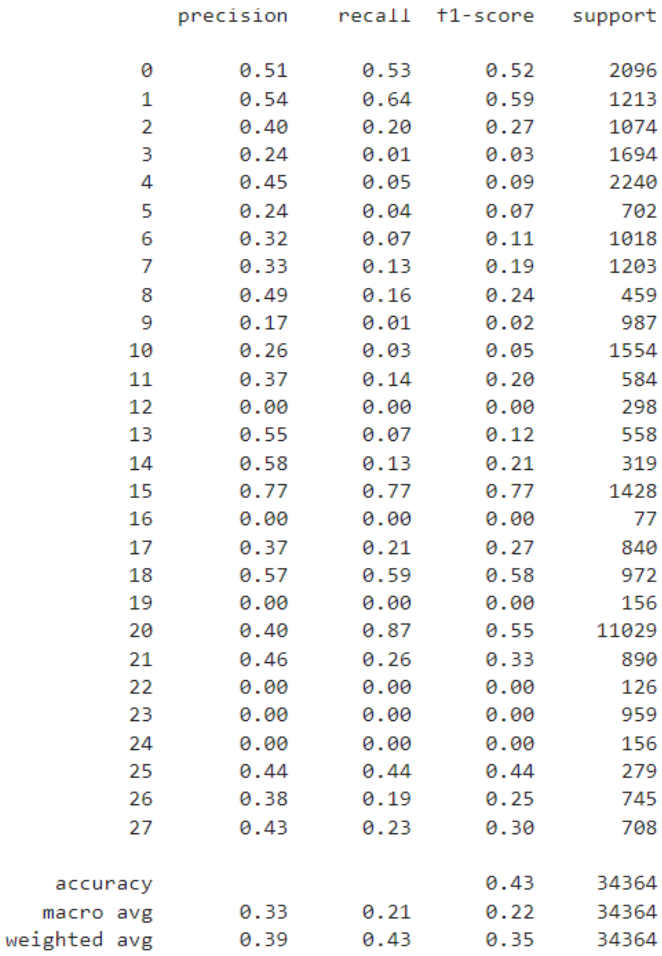
```
                precision    recall  f1-score   support

    admiration       0.45      0.28      0.34      2129
     amusement       0.52      0.48      0.50      1221
         anger       0.31      0.14      0.19       965
     annoyance       0.14      0.01      0.03      1634
      approval       0.15      0.01      0.02      2300
        caring       0.22      0.05      0.08       722
     confusion       0.29      0.07      0.12      1015
     curiosity       0.32      0.02      0.03      1169
        desire       0.31      0.14      0.19       422
disappointment       0.15      0.01      0.02       992
   disapproval       0.10      0.01      0.02      1552
       disgust       0.03      0.00      0.00       607
 embarrassment       0.05      0.00      0.01       319
    excitement       0.16      0.02      0.04       583
          fear       0.10      0.00      0.01       355
     gratitude       0.82      0.72      0.76      1413
         grief       0.00      0.00      0.00        49
           joy       0.35      0.13      0.19       869
          love       0.63      0.47      0.54       916
   nervousness       0.00      0.00      0.00       175
       neutral       0.37      0.91      0.53     10932
      optimism       0.47      0.17      0.26       923
         pride       0.00      0.00      0.00       137
   realization       0.14      0.00      0.01       991
        relief       0.04      0.01      0.01       161
       remorse       0.45      0.33      0.38       310
       sadness       0.52      0.09      0.16       802
      surprise       0.34      0.16      0.22       701

      accuracy                           0.39     34364
     macro avg       0.26      0.15      0.17     34364
  weighted avg       0.33      0.39      0.29     34364
```

Figure 11: Top-DT Classification Report (Emotion)



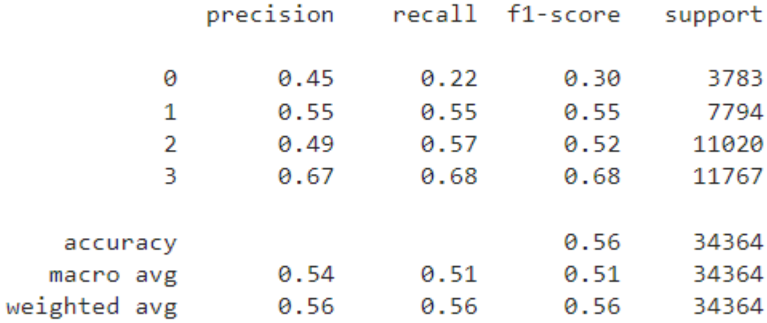
```
              precision    recall  f1-score   support

   ambiguous       0.36      0.47      0.41      3876
    negative       0.51      0.60      0.55      7760
     neutral       0.50      0.47      0.49     10932
    positive       0.71      0.59      0.64     11796

    accuracy                           0.54     34364
   macro avg       0.52      0.53      0.52     34364
weighted avg       0.56      0.54      0.55     34364
```

Figure 12: Top-DT Classification Report (Sentiment)



```
              precision    recall  f1-score   support

           0       0.51      0.53      0.52      2096
           1       0.54      0.64      0.59      1213
           2       0.40      0.20      0.27      1074
           3       0.24      0.01      0.03      1694
           4       0.45      0.05      0.09      2240
           5       0.24      0.04      0.07       702
           6       0.32      0.07      0.11      1018
           7       0.33      0.13      0.19      1203
           8       0.49      0.16      0.24       459
           9       0.17      0.01      0.02       987
          10       0.26      0.03      0.05      1554
          11       0.37      0.14      0.20       584
          12       0.00      0.00      0.00       298
          13       0.55      0.07      0.12       558
          14       0.58      0.13      0.21       319
          15       0.77      0.77      0.77      1428
          16       0.00      0.00      0.00        77
          17       0.37      0.21      0.27       840
          18       0.57      0.59      0.58       972
          19       0.00      0.00      0.00       156
          20       0.40      0.87      0.55     11029
          21       0.46      0.26      0.33       890
          22       0.00      0.00      0.00       126
          23       0.00      0.00      0.00       959
          24       0.00      0.00      0.00       156
          25       0.44      0.44      0.44       279
          26       0.38      0.19      0.25       745
          27       0.43      0.23      0.30       708

    accuracy                           0.43     34364
   macro avg       0.33      0.21      0.22     34364
weighted avg       0.39      0.43      0.35     34364
```

Figure 13: Top-MLP Classification Report (Emotion)



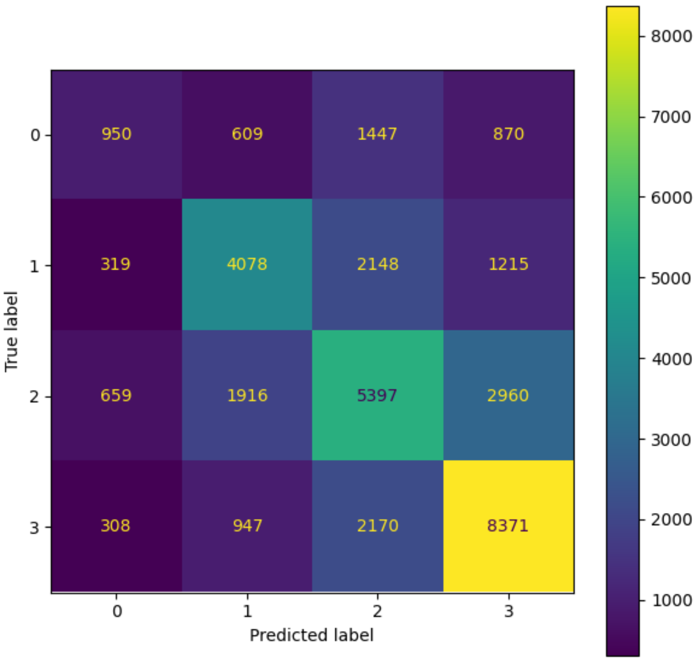
```
              precision    recall  f1-score   support

           0       0.45      0.22      0.30      3783
           1       0.55      0.55      0.55      7794
           2       0.49      0.57      0.52     11020
           3       0.67      0.68      0.68     11767

    accuracy                           0.56     34364
   macro avg       0.54      0.51      0.51     34364
weighted avg       0.56      0.56      0.56     34364
```

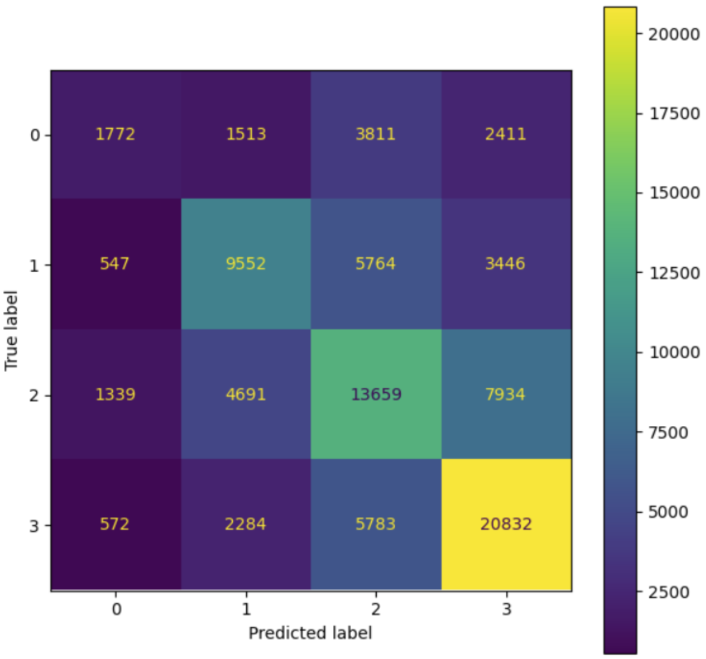
Figure 14: Top-MLP Classification Report (Sentiment)

4.2: Part 2 Analysis

Comparing the base versions of the different models together:

When comparing the base versions of the different models for the emotion classification, we do see some consistency in the values across the different models. For example, the accuracy for the 3 models varies from 0.36 to 0.40 as seen in Figures 3,5,7. Furthermore, we can see that as predicted, the sentiment accuracy should be higher since the sentiment classes were more balanced as seen in 1.3. As seen from figures 4,6,8 the accuracy for the sentiment varies from 0.51 to 0.55 which are higher than the ones for emotion. Moreover, the macro averages in the sentiment seem to be higher than the ones in the emotion category. This also makes sense because since there are less options for the sentiment, the recall and precisions are also better since the odds of getting a right option is higher. In turn, the macro averages and weighted averages are also better in the sentiment classifications. Lastly, when comparing the three models together when looking at the base version, we notice that the Naïve Bayes has lower recall and macro average than the other two models and this makes sense since Naïve Bayes is using probability as its basis to determine everything. Furthermore, when it comes to the change of the dataset size for the training and testing, we changed the training set size from 0.8 to 0.5 and the testing size from 0.2 to 0.5. This seems to have led to all the values of accuracy, precision, recall, f1 measure, macro average and weighted average to drop. This makes sense because of the theory learned called underfitting. This is happening because we are reducing the training set and therefore it could be that not enough cases were encountered and therefore the testing section doesn't have enough information to go from. As can be observed from the following confusion matrices of the two different sizes of training and testing sets.





This is for the Sentiment category and as can be seen from the confusion matrices, since we have more data for the testing part, we therefore have higher wrong predictions as well.

Comparing the top versions of the different models together:

When comparing the different top versions of the different models, we see that the accuracies are in general higher in the sentiment classification than for the emotions part. This is also the case for the base versions comparison. This makes sense as we had predicted because even though not perfectly balanced, the dataset for the sentiments is more or less balanced. Furthermore, the trends observed between the different top versions are very similar to the differences observed in the base versions.

Comparing the base with the top of the same models together:

When comparing the base versions with the top versions of the models, although not big, we see a difference. This is the fact that the top versions of the models seem to be doing better and this is because we put specific hyperparameters to improve them. For example, if we look at figure 3 which is the base version of MNB and compare it to figure 9 which is the top version of MNB, we can see that the 0.16 to 0.28. This is quite a significant change and shows how adjusting the hyperparameters can improve the results of our tests.

4.2: Part 3 Analysis

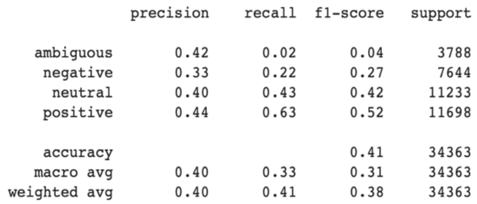
Base MLP:

We started running the Base MLP, with maximum iterations to its default value (200) and couldn’t finish the running as it took several hours and was still running. So, we interrupted it and again ran it with 20 epochs, just so we can get the results. But the results are not good enough because of very few iterations. For, emotions the F1 measure is almost zero for all the classes, that is the reason we got 0.06 and 0.20 as macro and weighted average respectively.

```
                precision    recall  f1-score   support

    admiration       0.42      0.03      0.06      2044
     amusement       0.40      0.04      0.07      1225
         anger       0.55      0.03      0.05      1038
     annoyance       0.00      0.00      0.00      1618
      approval       0.50      0.00      0.00      2213
        caring       0.00      0.00      0.00       748
     confusion       0.00      0.00      0.00       988
     curiosity       0.23      0.01      0.01      1190
        desire       0.00      0.00      0.00       436
disappointment       0.00      0.00      0.00       954
   disapproval       0.50      0.00      0.01      1541
       disgust       0.00      0.00      0.00       566
 embarrassment       0.00      0.00      0.00       308
    excitement       0.00      0.00      0.00       598
          fear       0.00      0.00      0.00       356
     gratitude       0.69      0.39      0.50      1379
         grief       0.00      0.00      0.00        77
           joy       0.29      0.02      0.04       901
          love       0.38      0.24      0.30       977
   nervousness       0.00      0.00      0.00       155
       neutral       0.34      0.98      0.50     11233
      optimism       0.00      0.00      0.00       881
         pride       0.00      0.00      0.00       142
   realization       0.00      0.00      0.00       931
        relief       0.00      0.00      0.00       154
       remorse       0.61      0.06      0.11       294
       sadness       0.33      0.00      0.00       737
      surprise       0.00      0.00      0.00       679

      accuracy                           0.35     34363
     macro avg       0.19      0.06      0.06     34363
  weighted avg       0.29      0.35      0.20     34363
```

Since, the dataset is skewed we will focus on weighted average.

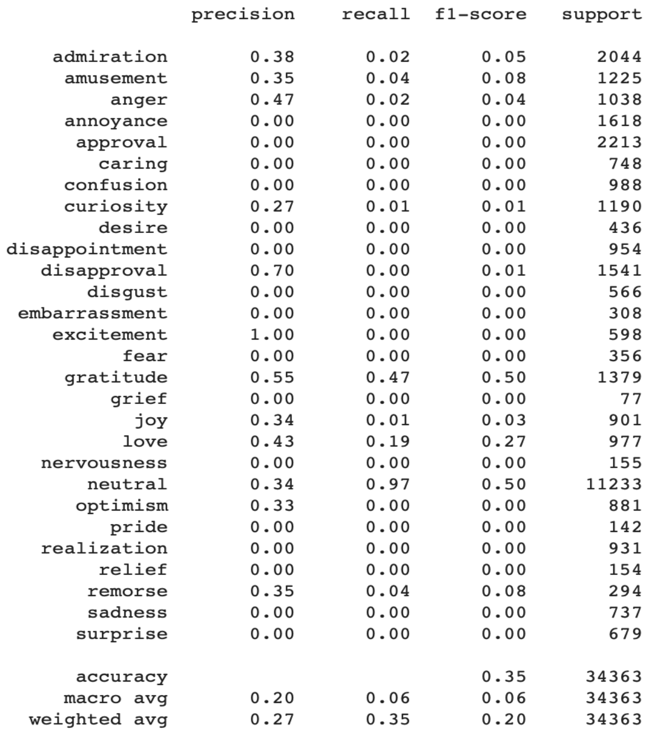
For the sentiments, we got the results we expected as there were less classes so, the training went effortlessly. In sentiments, we were focusing on accuracy as all classes were almost equally likely and we got the accuracy of 0.41 in 20 epochs, which is not bad, considering very few iterations.



```
              precision    recall  f1-score   support

   ambiguous       0.42      0.02      0.04      3788
    negative       0.33      0.22      0.27      7644
     neutral       0.40      0.43      0.42     11233
    positive       0.44      0.63      0.52     11698

    accuracy                           0.41     34363
   macro avg       0.40      0.33      0.31     34363
weighted avg       0.40      0.41      0.38     34363
```

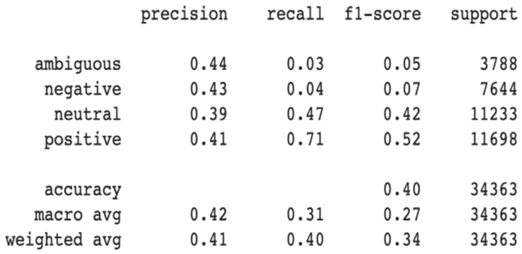
Top MLP:

Now, if we look at top MLP, we tried to maximize the weighted average by giving a *scoring* parameter in grid search. But we couldn't get the good results due to same (very few) iterations.



```
                precision    recall  f1-score   support

      admiration       0.38      0.02      0.05      2044
       amusement       0.35      0.04      0.08      1225
           anger       0.47      0.02      0.04      1038
       annoyance       0.00      0.00      0.00      1618
        approval       0.00      0.00      0.00      2213
          caring       0.00      0.00      0.00       748
       confusion       0.00      0.00      0.00       988
       curiosity       0.27      0.01      0.01      1190
          desire       0.00      0.00      0.00       436
  disappointment       0.00      0.00      0.00       954
     disapproval       0.70      0.00      0.01      1541
         disgust       0.00      0.00      0.00       566
   embarrassment       0.00      0.00      0.00       308
      excitement       1.00      0.00      0.00       598
            fear       0.00      0.00      0.00       356
       gratitude       0.55      0.47      0.50      1379
           grief       0.00      0.00      0.00        77
             joy       0.34      0.01      0.03       901
            love       0.43      0.19      0.27       977
     nervousness       0.00      0.00      0.00       155
         neutral       0.34      0.97      0.50     11233
        optimism       0.33      0.00      0.00       881
           pride       0.00      0.00      0.00       142
     realization       0.00      0.00      0.00       931
          relief       0.00      0.00      0.00       154
         remorse       0.35      0.04      0.08       294
         sadness       0.00      0.00      0.00       737
        surprise       0.00      0.00      0.00       679

        accuracy                           0.35     34363
       macro avg       0.20      0.06      0.06     34363
    weighted avg       0.27      0.35      0.20     34363
```

For the sentiments, we got 0.34 as weighted average and 0.40 as the accuracy. There is not much difference between these values because of equally likely classes.



```
              precision    recall  f1-score   support

   ambiguous       0.44      0.03      0.05      3788
    negative       0.43      0.04      0.07      7644
     neutral       0.39      0.47      0.42     11233
    positive       0.41      0.71      0.52     11698

    accuracy                           0.40     34363
   macro avg       0.42      0.31      0.27     34363
weighted avg       0.41      0.40      0.34     34363
```

Comparison:

There was no difference at all in the performance of Base and Top MLP for emotions, this was due to a smaller number of epochs that the classifier was not trained good enough to show better results. But there was difference in F1 score of individual classes and the score for Top MLP is slightly higher than that of Base MLP.

On the contrary, the scores for sentiments in Top MLP were lower than that of the base version. But it is not a significant difference. This shows that the parameters used for grid search were not good enough to get in the better results and looking at the overall performance, Base MLP performed better for the embeddings.

4.3: The tasks were split equally among all the teammates. For part 1, we decided to have 1 person take care of it so that we would be consistent with the work. Then for part 2, the tasks were split by the different models. Then for part 3, the people in charge of the models from part 2 took care of it. Lastly for part 4 a team effort was done, and everyone tried giving their input.