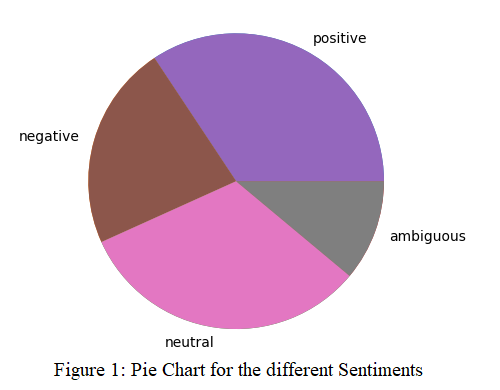
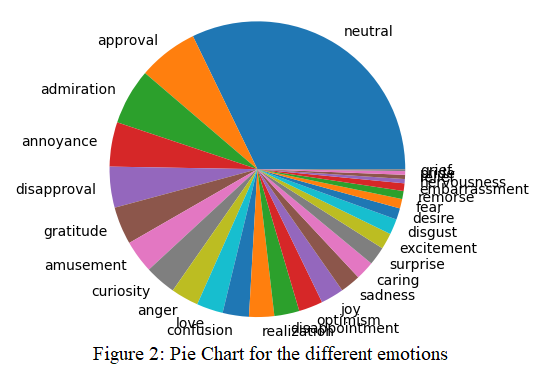
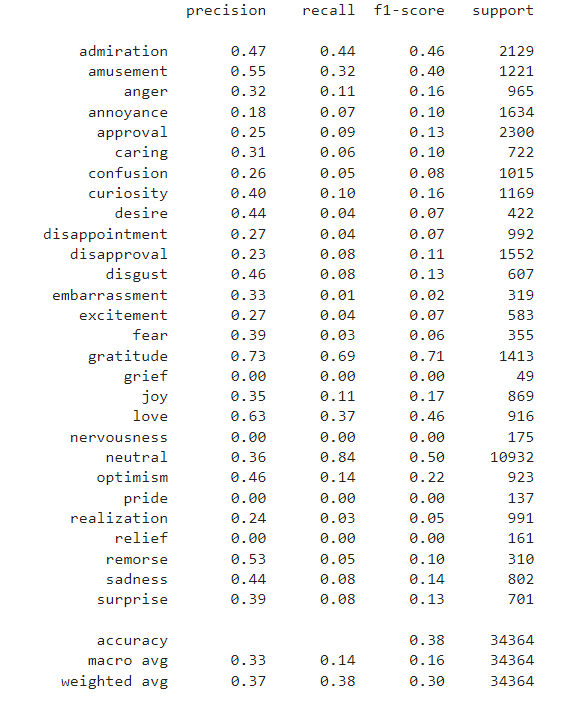
**4. Analysis**

**4.1**: By looking at the dataset given on Moodle and then visualizing it through the use of Jupyter Notebook, we are able to have a better idea on the topic of classification. The plot function was used from the matplotlib library in order 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 an accuracy metric is good to use when the dataset is 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 as a way 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 epoch since we’ve determined through Jupyter Notebook that the vocabular is very large with a size of 30449 words. Thus, it has to 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.

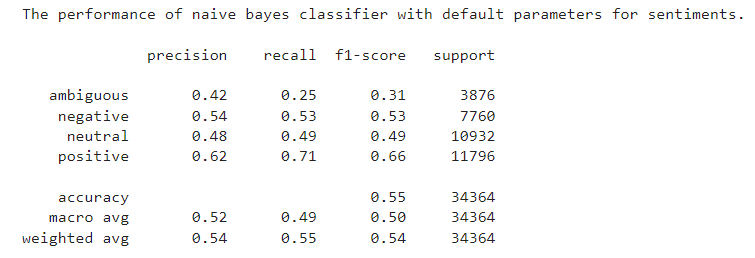
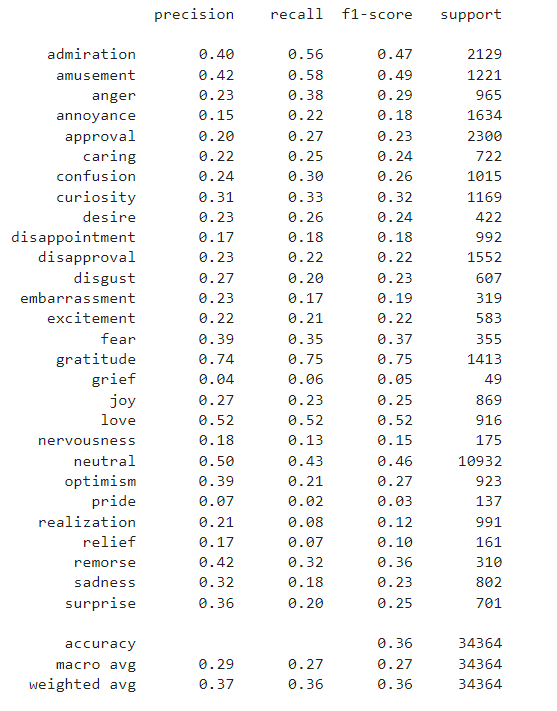
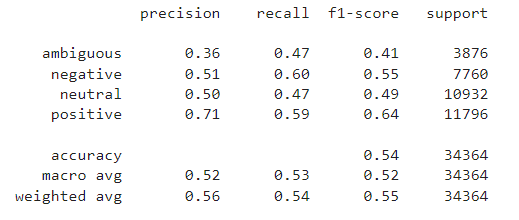
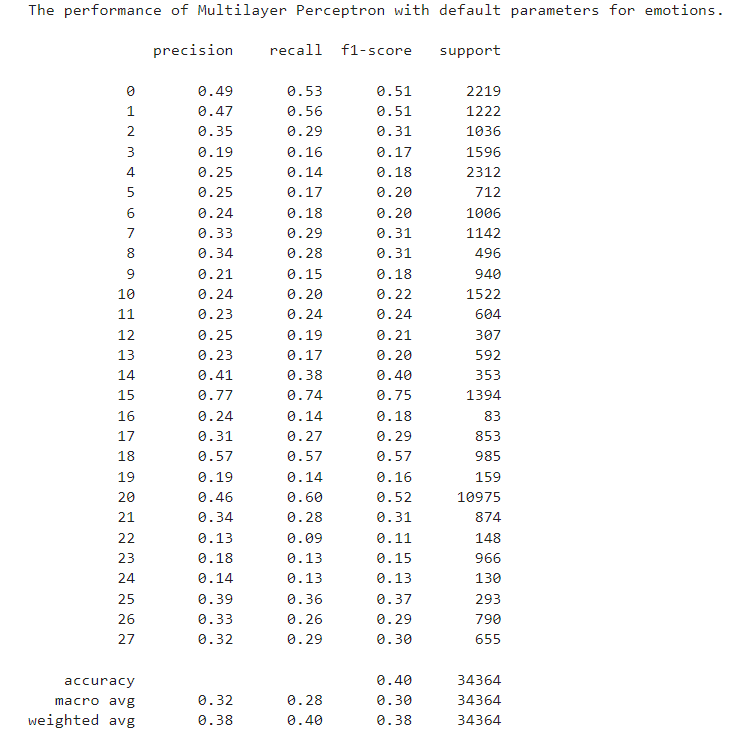
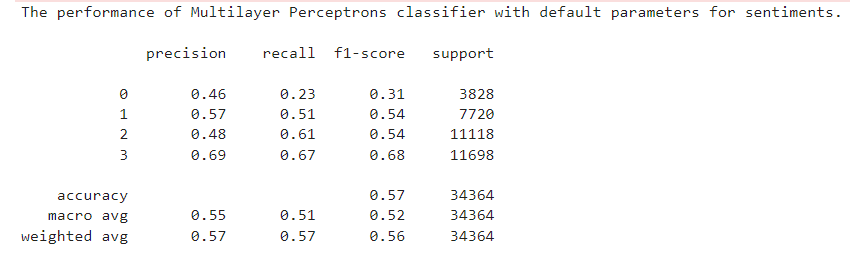
Figure 3: Base-MNB Classification Report (Emotion)

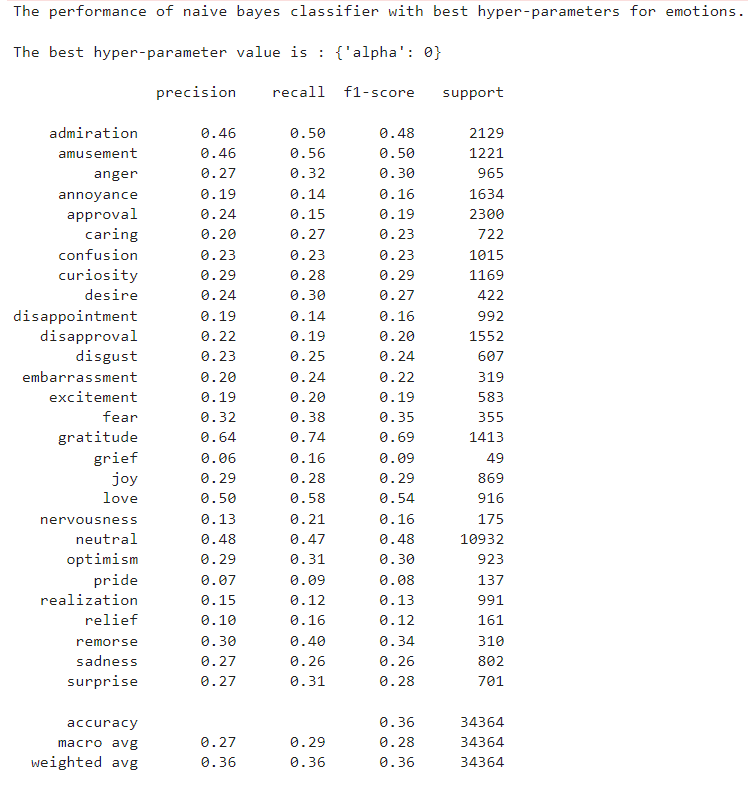
Figure 4: Base-MNB Classification Report (Sentiment)

****Figure 5: Base-DT Classification Report (Emotion)

****Figure 6: Base-DT Classification Report (Sentiment)

****Figure 7: Base-MLP Classification Report (Emotion)

****Figure 8 : Base-MLP Classification Report (Sentiment)

Figure 9: Top-MNB Classification Report (Emotion)

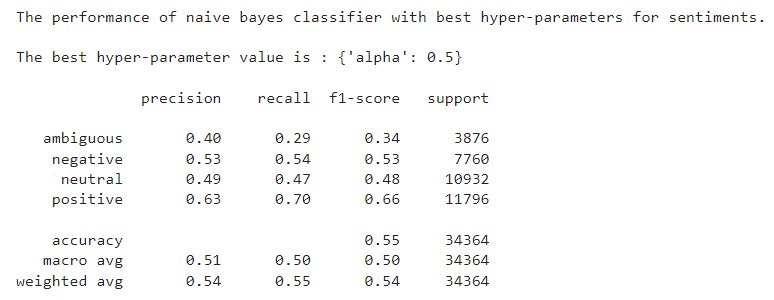
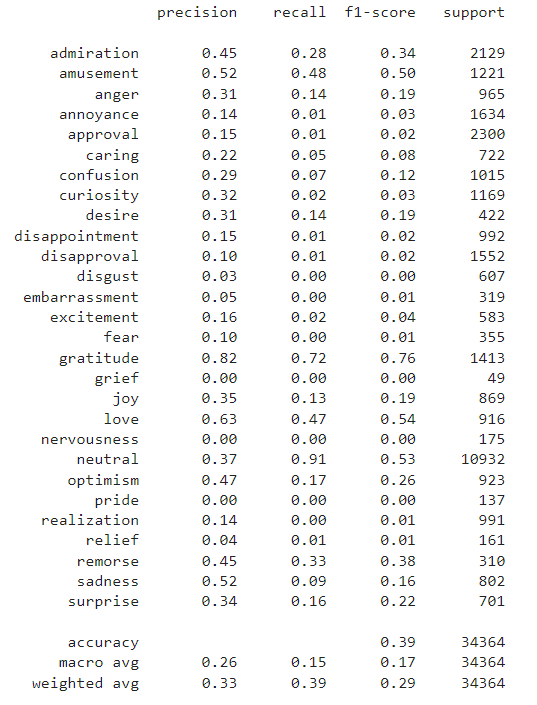


Figure 10: Top-MNB Classification Report (Sentiment)



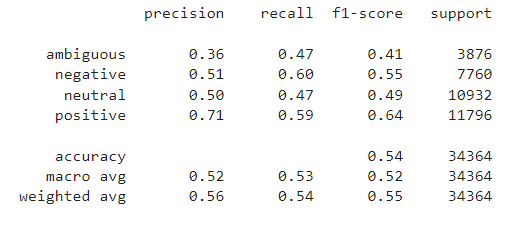
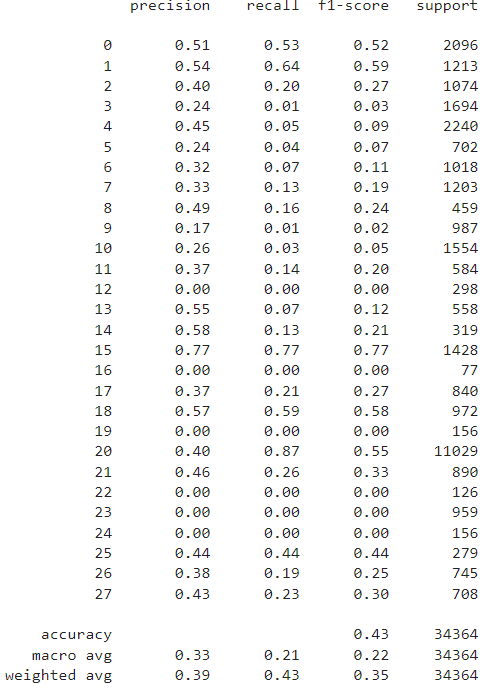
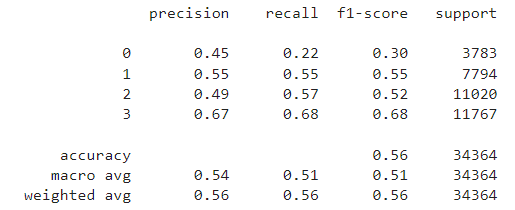
Figure 11: Top-DT Classification Report (Emotion)

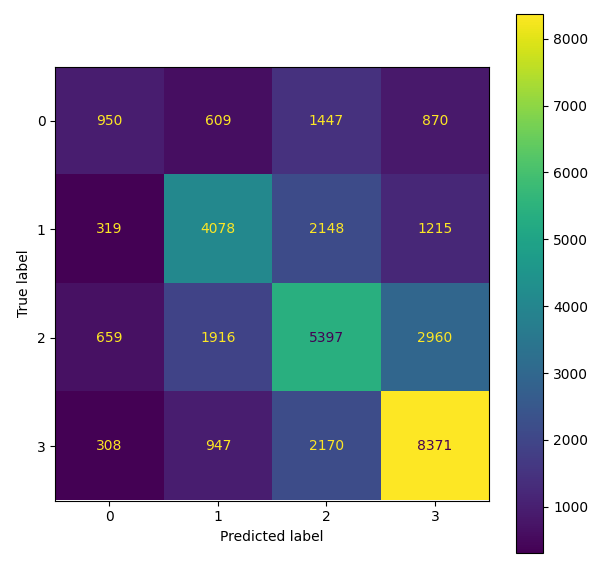
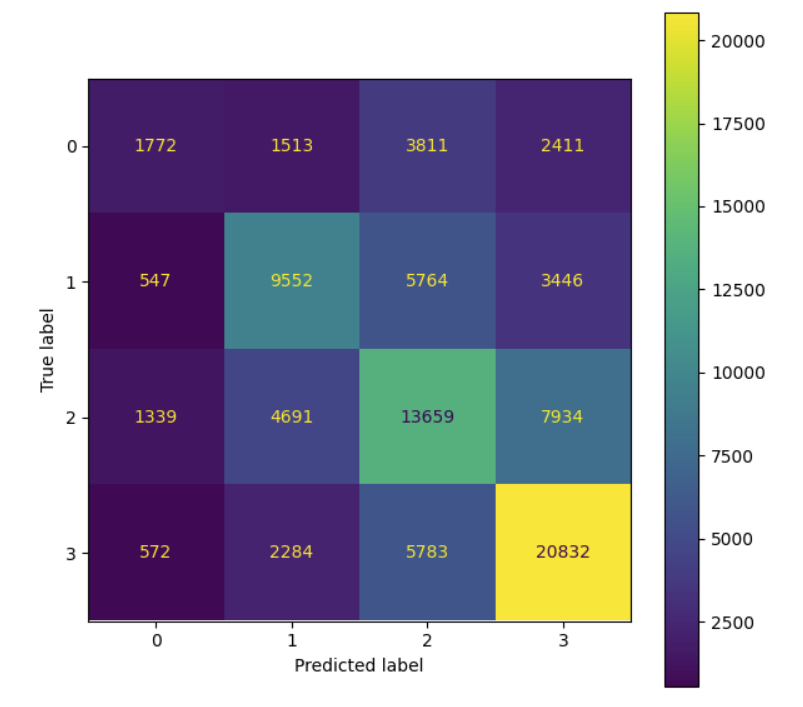
Figure 12: Top-DT Classification Report (Sentiment)

Figure 13: Top-MLP Classification Report (Emotion)

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 more 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 are 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 in order 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.

**4.3**: 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.