Reflection:

In order to complete this project, various investigations were undertaken to understand the context behind the task. The first investigation done was on the CountVectors on the provided dataset. After comprehending and understanding the theory and reason behind why this preprocessing was needed, the code required was also learned. Even though there was a provided sparse matrix dataset by countvectorization, I decided that since I was not sure of the sorts of specific preprocessing that were undergone, making a personalised sparse matrix should be easier to understand and manipulate. But there was a lack of exploration into different preprocessing techniques such as tf-idf vectorizers or Doc2Vec, which may have led to a more accurate model. For the feature selection, I merely used a manual sequential forward selection method which tried to provide the best set of features and seemed to be sufficient. However, I could have recognized beforehand that some features did not have high contribution to the target feature, which may have been more time efficient. Furthermore, perhaps trying out more than two classifiers may have been able to provide a new classifier with better accuracies than a simple Naïve Bayes or Decision Tree classifier. Also, when providing context to the decisions I have made, I should have provided more data on the steps that were deployed, such as using graphs or accuracy evaluations of different methods tidied in a table with titles. Another point that could be improved on was not merely relying on the fact that the dataset is large and that the holdout method would provide an acceptable evaluation accuracy, but actually trying out the cross-validation method first to check if the resulting evaluation is similar. For the error analysis, the argument that the error due to an unbalanced dataset seems viable and that it will not provide a perfect prediction model for the label 3 in the actual test data set.