Reflection

To begin with, me and my teammate read the instruction carefully and then decided for the works we need to be done. Then, I have read the provided datasets in detail, and considering the characterises of the data as well as the thing we might need to do to process the data into useful information. We want to build a baseline model with MNB first so we can do the feature selection as some features might not help us with classification. After building the baseline model, we test to see if some features would increase the prediction accuracy. Therefore, I use trial and error to test out that “name” and “steps” have the highest accuracy. After that, we build more models such as SVC, Logistic Regression and Random Forest using the selected features and process the feature by counter vectorizer and tf-idf separately. We still want to obtain higher prediction accuracy, so grid search and other methodologies are applied to tune the hyperparameters. Finally, we build a max voting ensemble model of the above classification algorithms. During the process of writing codes, we have talked to each other a lot to make sure we understand each other’s code and discuss any ideas. After the coding, we started to write the report. We discussed the approach, result, evaluation and other things, which some graph and statistic as supporting evidence.

Reflect on this project, I think that we have done well in produce various plots and tables to visualise the data and results, such as the boxplots used to remove some useless features during feature selection, the heap map generated for grid search to find the optimal hyperparameters, and the tables listed to present the results. They really help us with illustrating the behaviour of models and the decision we made. Also, we used two methodologies (bag-of-word, tf-idf vectorizer) to process the texts into numerical data in order to obtain a higher accuracy for different models.

However, there are also a few points that I believe should be improved. Start from the basics, I found that we have some obvious typos and some grammar errors that could be found by reviewing the report carefully. And I would like to add some page numbers in the report to help others read. Besides these, I realized that we need to improve our explanation of why a particular model/technique perform good or bad. This something I did not really notice when I was writing the report, for most of the times, I just simply listed all the features of different methods without a proper connection to our project. Lastly, I think the way we implement the feature selection is not efficient, so I would like to try to use other methods like sequential forward selection or filtering to engineer feature in a smarter way.

I worked on the project with my teammate Wanxuan. And I reckon that we have a pretty fair amount of contribution throughout the task. For example, she has processed the data into a usable format, then I did the feature selection and then build the classification models. She performs the grid search and tunes the parameters, followed by writing the ensemble models and test the result. For the report part, I have written the introduction, conclusion, future works and reference. Then we did the “approach”, “results” and evaluation parts together. Overall, I think we work well as a group since we have a clear division of things to do, as well as motivated mindsets. Also, we are willing to solve the problem together and helping each other through the work.