Assignment 2: Reflection

Wanxuan Zhang (1079686)

For this project, I first split the data “recipe_train” into train and development set. By visualising the n_steps and n_ingredients by using a boxplot, we find that the numeric attributes of the number of steps and number of ingredients might not be appropriate to fit the model. We then take the text features, combine the text features together and use count vectorizers to process the words. Under the baseline model using multinomial Naïve Bayes, the combination of name and steps will give a relatively highest accuracy hence we chose to use these two features to process the other models. Other than the MNB, we also fit the training data to Logistic Regression, Linear SVC, Ensemble Model using max voting and Random Forest. We train the classifiers by using the 30000 instances for training and test on the 10000 development set. Counter vectorizer and tf-idf are applied to process the data to find which works better for each classifier on this dataset. We perform a 5-fold cross-validation procedure for them to reduce some bias and get a more reasonable result. Moreover, we investigate the problem of overfitting by comparing the difference of the accuracy between the training set and the development set to see if the models have a much higher accuracy on the training set. To find the best parameter that generalises the dataset well, we tune the parameters using grid search with the SDG classifier. It simulates Linear SVC and Logistic regression through different loss functions. The LinearSVC is shown to have a better performance, hence I further perform another grid search only on the Linear SVC and plot a heatmap to find the optimal parameters for the model. By comparing the models with tuned parameters and then use those top-performing models to test the actual test set, we find that the ensemble model using max voting with attribute steps and name vectorising using tfidf will obtain the highest accuracy.

Something that I am satisfied with is that I have presents the results for each model with different attributes or vectorizing methods in a table where we can compare the performance for each model. Also, since I divide the data into train and development set, it is easier for me to identify if there is any overfitting. I have use grid search and illustrates in plots which helps me to tune the parameters. The use of k-folds validation reduce some bias and makes my results more reliable.

On the other hand, something that can be improved is that I should use the Linear SVC model with the tuned parameters in the ensemble model to achieve a better result. I should also tune the parameters for other models to see if the performance can be improved. I could use a more advanced algorithm (such as bagging and stacking) to build the ensemble model instead of simply using max voting. In addition, In the future, I could evaluation my result in more details. For example, I can use different Multiclass Evaluation Methods to calculate the overall precision and recall (eg. Macro-Averaging, Micro-Averaging, Micro and Weighted Averaging) and compare the results to get a deeper understanding of the performance of the model.

We are work in a group of two, I train the models and tune the parameter, whereas my partner describes the functionality of each model and the methods that we use to vectorise the text in the report. We both work together to analyse and evaluate the results and reflect on some further improvements that we could make in future to build better models.