Firstly, after I finished reading the spec given, I noticed that analyzing the text review will be crucial on predicting the rating labels. Since text is a subjective information, we need to find a way to quantify the text. After doing some research, I found two ways of vectorizing method. One is using Count vectorizer where you convert the text into "bag of words" format, ranking based most frequent words. Another is using TF-IDF vectorizer where you rank based on the least frequent words. The assignment also provides Doc2Vec representation, which is a higher level of vectorized representation which takes account the structure of the text per given document. Thus, I decided to devise the experiment with all the vectorization method to find which method performs best.

After figuring out the method of vectorizing, I looked at the meta data given. I concluded that only a voting information are useful and assumes that other attributes will not improve the classification performance. This assumption is also derived from assuming that the reviewer maintains complete unbiasedness when reviewing any restaurant. Due to the given time of the project, I only include the meta data to the doc2vec data successfully. However, the experiment devised tested both doc2vec with meta and without meta. For Count and TF-IDF vectorizer, feature selection with both mutual information and chi-squared analysis will be conducted to show which one has the better performance on each method.

Then all the mentioned data is trained to Linear Support Vector Machine, Logistic Regression, Multinomial Naive Bayes and Gaussian Naïve Bayes. Naïve bayes is chosen due to the robustness to most dataset with their assumption on conditional independence. Multinomial method was chosen since frequency based vectorizer such as Count and TF-IDF will be best analyzed with multinomial assumption. However, doc2vec is not able to be processed by multinomial assumption as it represents the correlation of the word with the structure, meaning that negative values will be present, which contradicts with the nature of multinomial. Thus, gaussian assumption is used when frequency vectorizer is compared to doc2vec representation. Linear support vector machine is chosen as the nature of maximizing margin is good fit for measuring different words for specific polarity. Same goes for logistic regression as the estimation of natural probability is a good fit to the problem. Both are extended to multiclass to fit this project.

There are several things that are considered but not implemented in the experiment. First is experiment with more hyperparameter values for the classifier, such as the C variable for logistic and svm classifiers to achieve the ideal hyperparameter. An experiment on the info gain of meta data should be implemented as well, as the bias of reviewer towards a restaurant may have a strong impact on classifiers. More classifier such as stacking and random forest was planned but not implemented as well. This could improve the experiment as the right base classifiers may reduce bias and variance. Network based classifier should also be looked at as it is able to represent a higher level of text representation. Analysis on report can also be improved as most graphs only shows line graph on performance. Confusion matrix and full classifier performance result of data can be shown and analyzed in the report.