Peer1

The author begins the experiment by figuring the best feature selection/combination to be trained to the classifiers. The provided meta data is analyzed with chi-squared and mutual information technique to measure its relevance on predicting the class. The author then proceeds with encoding of text data by two ways, “Bag of words” and vec2doc with multiple feature option representation. Both are experimented into the chosen classification models, which are zero-r, multinomial naïve Bayes, support vector machine, decision tree, KNN, logistic regression, random forest , and stacking system. The test was performed with 10-fold cross validation. The author then analyzes the single model’s accuracy showing that logistic regression performs highest with both bag of words and vec2doc representation, with bag of words having the higher accuracy. Then random forest classifier was trained which receives significantly lower accuracy. And finally, the two-stacking system was implemented. First stack consists of logistic plus multinomial naïve bayes with meta classifier of support vector machine which obtain overestimated accuracy. Second stack system consist of logistic plus knn with 10 nearest neighbors with logistic as meta classifier which receives a good accuracy as well. The author then evaluates the accuracy with Kaggle submission accuracy which shows very different results. The author then concludes with the difficulty with generating a generalizable solution to sentiment analysis.

The author did well with devising the experiment. Every step the author did has a reason and backed with experimental evidence. An example is how the author analyze mutual information and chi-squared value on meta feature to be included, whilst most people would just assume voting result is the most relevant feature to the other met features. The author has an informative and deep discussion which shows a good attempt of research. This was shown when the author mentioned about Colinear features having a tendency of lowering performance. It was further backed on how knowledgeable the author is with different factors affecting bias and variance of a classifier.

What the author needs improvement is devising a better experiment with the stacking model as it seems to be overfitted data compared to the training of other models. Literature summary of the project seems too long. Instead, the author could explain more evaluation and explanation to the result received. The author could experiment and explain more with K-Best feature selection, the function used (e.g. Mutual information or chi-squared) , and with different K level to achieve the most ideal K-best feature.

Overall, this is a good solid attempt on tackling the project.

Peer2

The author has devised the experiment by using Count vectorizer representation of the text review. The experiment assumes that the meta data is irrelevant to model prediction. Next, the author runs several simulation to figure out the best number of features for feature selection. After coming up with ideal feature number 400, the author runs the training set to several chosen classification which includes KNN, multinomial naïve bayes, logistic regression, support vector machine , and a stacking system. The author runs cross validation to measure accuracy with unspecified folds. Next, the author simulate an experiment of

tuning the hyper parameters for logistic and support vector machine which receives an ideal parameter of C = 0.7. Then, the author did some error analysis by looking at the confusion matrix of logistic and stack system classifiers. Essentially discussing how the models does not perform well in predicting neutral ratings. The author then concludes the experiment by stating that the difficulty in predicting a neutral polarity results in similar performances on all classifiers.

The author has done well in designing the experiment. Experiment on finding optimal parameters is often conducted, one with the ideal best feature number, another with tuning of hyper parameters of logistic and support vector machine. The author also shows good error analysis in the results when he analyzes the confusion matrix of the classifiers. Overall, the author shows a good attempt on the project.

Suggestions on improving the experiment includes using the doc2vec and other vectorizer representation compared to count vectorizer. The report shown was also lacking in critical analysis such as the number of cross validation folds, reasoning behind the accuracy received by models, and the reasoning behind tuning the hyper parameters. And also, an experiment could be devised about removing meta features from the dataset.