COMP30027: Assignment 2 Reflection

Rohan Hitchcock (836598) and Patrick Randell (836026)

We approached the task predicting the star rating of restaurant reviews by using a selection of support vector machine (SVM) and logistic regression classifiers, which use paragraph vector (doc2vec) encodings of the review text as features. We identified this as a sentiment analysis task, and begun by reading literature on sentiment analysis techniques. We focused our attention on review and survey articles which provided a comparison of many different sentiment analysis methods, and on this basis chose to pursue SVM and logistic regression based classifiers. We also researched paragraph vector encoding so we could better understand how the features it generated would work in our model. We then performed initial tests to narrow down the number of variables we needed to consider (such as to reduce the number of kernels we were investigating).

We believe we presented a comprehensive investigation and analysis of the models we selected which was informed both by an understanding of the literature and the results we observed in the process of completing the assignment. Extend this

One key aspect we could have improved, which was briefly discussed in our report, was add a preprocessing step to linguistically simplify text prior to the paragraph vector encoding step. This could have involved: using a spell checker to automatically correct typos and misspellings, attempting to simplify negated words (such as converting "not good" to "bad"), and identifying common phrases and idioms. Although paragraph vector encoding is theoretically able to account for this, it likely requires many examples of synonymous phrases in order to map them properly to the feature space.

We could have investigated a model that takes advantage of the logical ordering of the class label such as Ordinal Logistic Regression. Including a greater variety of model types (They were mostly linear) during ensembling may have led to a greater performance?

We could have also improved the overall presentation and usability of our code. As it stands, changes need to be made in the source code to select different models and hyperparameters. This flexibility was helpful while we were experimenting with different ideas, but it is currently not very usable outside of that setting.

Research was conducted by both members prior to any developement. Each member investigated a class of models; Rohan explored SVMs while Patrick did preliminary feature selection and data exploration before moving onto Logistic Regression. When each of us were satisfied with the extent of our investigation and results had been collected, Rohan moved onto collating all our work into a consistent and presentable manner and started the report. Patrick investigated the stacking of models during this time. The final draft of the report was refined together. (Please change any of this how you want)

Discuss (400 – 600 words):

- Critical reflection on the process of completing this project.
- What were we satisfied with?
- What could be improved?
- What are the individual contributions?