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 SVM kernels we investigated), and then proceeded with a thorough analysis of the selected models.

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. We took care to both provide theoretical and experimental justification for our decisions, and connect our research to results in the literature.

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.

Another improvement we could have made would be to explore more model variants in greater detail. For example we could have investigated a model that takes advantage of the logical ordering of the class label such as Ordinal Logistic Regression, or continued investigation of SVM classifiers using polynomial and sigmoid kernels. Including a greater variety of model types in the stacking model may have led to better 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.

Statement of Authorship

R.H. and P.R. conducted the literature review. R.H. investigated and implemented the SVM-based models. P.R. investigated and implemented the logistic regression and stacking models. R.H. and P.R. wrote the report. R.H. and P.R. contributed equally to this assignment.