Self-Reflection

COMP90044 – Semester 1, 2020 - Assignment 2

My approach to this problem was to see which classifiers were able to perform best when predicting the genre when fed a subset of the data set's features.

My approach can be broken into three parts. Firstly, the textual features are encoded using a TF-IDF scheme and the numerical features are scaled using a min-max scaler. Then the classifiers are optimised using a grid search over various hyper-parameters. Finally, the classifiers are scored using various metrics and their performance is analysed.

Overall, I think my report was well-written. The literature review was relatively short and succinct and utilised references from numerous sources. I investigated how my models were fitting and provided some insight into why they may be overfitting. Further, I feel that my general approach to the problem was sound and came to valid, falsifiable conclusions.

I feel my error analysis could have been improved upon. I did not investigate any specific instances of genres to figure out why they were incorrectly labelled. However, I think this is made up for in part by my overfitting analysis.

I would also have liked to investigate more options for tag/title processing, such as using the NLTK package. My inexperience with such a package and NLP in general made this quite difficult, however.

Perhaps with more time, also, I could perform larger, more complete grid searches, further optimising the models.

Also, I feel not being forced into a hold-out evaluation method would have been nice. I think this enforcement is quite restrictive in terms of the model validation we can perform.