Student Name: Chan Jie Ho
Student Number: 961948

Reading the two reports, I have come to properly understand how large of a field NLP can be with so many different ways to tackling a single problem; there are so many ways to clean data, encode the text, and to classify them. I learnt that a clean data without null or missing values is not necessarily good data, especially if it was unstructured data computed from another dataset.

Reviewing a solo report, it reminded me of the importance of being able to visualise the data before diving into a problem. The student had assumed the ratings were numeric and thus approached it as a regression problem at first, which yielded them poor results. This shows how important it is to take time to understand the data you are working with, especially if it's unstructured data. They have also made me come to understand the importance of the balance between stating our observations, backed with the appropriate evidences, and justifications as to why such findings were observed.

Reviewing a group report, having done a solo report myself, it has also made me fully realise how much a lower word limit impeded a lot in terms of how much information I can put in. Reading through, my approach, thought process and findings were not dissimilar to the group's, but perhaps the extra 1000 words worked to their advantage as they had much more leeway in being able to go in-depth into their explanations. That said, I realised that I should have explicitly stated my thought process and hypothesise on how a classifier might behave, based on solid assumptions that would in turn be based off in-class theories. For example, since Multinomial Naïve Bayes is biased against negative values in the data, a Gaussian Naïve Bayes might work better when using Doc2Vec. This would have shown that I had a thorough understanding of the theories and the behaviour of the classifiers. While I had provided justifications for why some classifiers worked better than others by commenting on the data features and how they work with/against the classifier's strengths/weaknesses, I should have backed this up more using my own accuracies. At first, I wanted to add in cross-validated accuracy scores and compare it against the ones on Kaggle, but decided not to in the case that people could trace it back to which person submitted it, though it slipped my mind that I should have just left the local accuracies in.

Overall, while the group report has managed to solidify my conclusion that a Tf-idf encoding method with a LinearSVC classifier works best with the data, the solo report has reminded me that there are more ways than one to approach a problem. Having gotten a lower score using the stacking classifier than the SVM like the group, it shows that a model's simplicity/complexity does not define its ability to accurate classify instances. My improvements section also left quite a bit of room for improvement, as I could have expanded more on topics such as pre-processing (stemming, lemmatization, etc.) and deep learning, but I was bordering on my word limit, so it would not have been in-depth anyways.