First Review:

The authors have done an excellent job in determining the best combination of encoding method, feature engineering, and classification method. They have shown thorough understanding of the data – critically analysing the underlying format and characteristic of the features and weighing it against each classifier's strengths and weaknesses. Not only did the authors go out of their way to manipulate and encode the raw reviews using their own encoding methods when they could have easily just used the pre-computed ones, but their understanding is also exemplified by their critical analysis of what each encoding method considers and does as well as their output, thus allowing them to make informed hypotheses of which classifier works best with which encoding method. The authors are also able to properly justify their feature selection choices, such as opting for chi-squared over the mutual information, with correlations and time efficiencies. Clear justification for hyperparameter tuning choices with accuracies, and kudos for ensuring minimal model bias (no overfitting).

Linking all these together, the authors did a marvellous job in experimenting with different combinations and basing their hypotheses on solid assumptions. Even further down in the report, the authors did well to relate each part of the report to previous hypotheses and providing concise justifications – validating and invalidating their previous hypotheses – obviously facilitated by a very smooth flow throughout the report. Also, error analysis was nothing short of insightful throughout the report while explaining thought process behind each choice as well as the consideration of the distribution of the classes itself. There was also a very good attempt at evaluation analysis evaluation methods were on point, confirming the reduction of evaluation bias and variance using the k-fold cross-validation method.

However, while the assumption that NB has that attributes are independent is correct, I feel that it contradicts the justification used when using Td-idf and also in the conclusion when talking about n-grams – would word combinations not imply that the words are slightly dependent on each other? Like the authors have said in the conclusion, they can improve by the considering the use of n-grams, or perhaps by performing some pre-processing on the raw reviews itself. E.g. normalisation, standardisation, stemming, lemmatization.

Overall, excellent report that I could barely find any parts to fault. Authors did an excellent job in showcasing their application of theories and models learned in-class as well as showcasing their fair share of external research.

Second Review:

Good attempt at determining the best combination of feature engineering, classifier, and hyperparameter tuning. There were examples of good hypotheses, however, there are a fair bit of inaccuracies too; for example, to say that sparse matrix is hard to feature engineer would be inaccurate as one can perform feature selection using chi-squared or mutual information methods. Good understanding of generic classifier performance towards feature correlation and good application of in-class theory using the LASSO.

Shortlisting **untuned** classifiers based on their accuracies would be unwise as it should instead be weighed against the distribution of the data, and how the data is represented, as these models may work well with proper hyperparameter tuning. Also, star ratings are nominal, not "inherently numeric", which would've been the correct justification as to why regressors had done poorly, as this is ultimately a classification problem. That said, your application based on your assumptions shows good understanding of in-class theories. Great that your claims on which classifiers did better are backed with accuracies, but you could've done better by hypothesizing the reasons based on theories. Hyperparameter tuning had good observations, but again, no analysis as to why they worked. Good hypothesis that the hyperparameters are correlated with one another.

In the future, try visualising the data and understand the distribution of the data first. Also, the statement that the data is already very clean is not entirely true; while there won't be missing values, the underlying data that the Doc2Vec was computed is text, which will include spelling errors and high frequencies of insignificant words, like "restaurant". The raw reviews could've been pre-processed using methods such as stemming and lemmatization. Try to work with the raw reviews itself with methods of representing the data such as Tf-idf vectorizer or tuning the n-grams hyperparameter in the vectorizers. Also, it would be faster to create another bag of words representation when tuning the vectorizer itself than creating another Doc2Vec, but since you are working on the pre-computed data, this option is out of the bag (pun intended).

While you have shown great application and ample observations with appropriate evidences, you did not provide justifications. Too much emphasis on what has been observed, rather than why this has been observed. There was also hardly any attempt at error and evaluation analysis. Cross validation was used but you failed to relate it with how this helps to reduce evaluation bias/variance.