In this report, I attempted to apply external knowledge such as hashtags, trending words and slang into feature engineering the best possible token words in the dataset. In addition to feature selection, I also attempted to analyse and compare the performance of classifier combinations to simple supervised classification learners.

In the implementation, I built several models using the sci-kit learn library and fitted my pre-processed data into the models in order to generate benchmark accuracies for me to make comparisons and consequently, improvements to the training and development data such as selecting the best features and attempting to apply significant weights to specifically location indicative words.

I believe in my report, I succeeded in developing a general consensus over the best performing model thus greater analysis in what could be improved in the best performing classifier, that was RFC. This was a result of my thorough analysis of a variety of classification machine learners and recognition of differences in the probabilistic models that they built. Moreover, I was able to identify limitations in each model which lent a helping hand to selecting models based on their underlying assumptions of the data. Additionally, I performed hyperparameter tuning to the model well, by determining the optimal size of the random subset and the number of trees which generated a higher accuracy than the base model. Ostensibly, I wrote in a concise and informative writing style alongside visuals to demonstrate my thinking process and the overall development of the task. The visuals were well labelled and provided more insight for readers.

Although there are several positive aspects in my report, I do believe there are some areas of improvement. First addressing the lack of depth in the task. Although, I have pointed out throughout the report that the amount of data is a significant limitation to the task, I could have improved my analysis of the best performing classifier by developing the skeleton of a small scale neural network in order to thoroughly explore viable machine learner options. Apart from the lack of depth, a deeper insight into the feature engineering section of the report could have been explored by stating the approach and algorithm/method applied in performing feature selection and experimentation - especially with hashtags, slang and location indicative words.

In this report, the author focused on the complement naïve Bayesian (CNB) learner and made alterations and improvements on the training and development data based on this learner by making comparisons to the multinomial naïve Bayes learner, bagging classifier and linear SVM. Additionally, the report explores the balance large datasets and efficiency in order to develop to make the CNB learner more cost efficient in terms of memory usage and computational power.

By focusing on improving the performance of a single learner – CNB – the author succeeds in creating an in-depth analysis on feature selection and hyperparameter tuning in regard to CNB. By identifying the underlying assumptions in the baseline and benchmark classifiers, the author demonstrates a moderate level of abstract though in thoroughly analysing limitations of the learners. Additionally, the author has clearly understood the theoretical properties of the methods applied to refine CNB such as exploring the concept of computational efficiency alongside best feature engineering.

An area of improvement would be to include visualisations of the accuracies generated by the learners to ease readers' understanding and ability to view the comparisons the author has mentioned. Apart from simple visualisations, has ultimately performed a thorough exploration of short text location prediction under this subject's limitation.

In this report, the author explores several supervised classification learners and attempts to perform hyperparameter tuning for each learner in order to compare and analyse improvements in accuracy. The author has performed pre-processing on the raw text data, built a Zero-R baseline model, multinomial naïve Bayes, support vector machine and logistic regression, and lastly analysed the accuracies and shortcomings of each learner in regard to applying it to short text location prediction.

There is a thorough analysis of feature selection and hyperparameter tuning in the report, which is also supported by visualisations. The author succeeded in identifying hyperparameters using grid search to optimise learner performance. Additionally, a thorough exploration in feature selection and the process underlying the methods was well written.

There are certain areas of improvement in regard to formatting and writing style. Although the report structure is logical and flows, the author did not format the report template perfectly such as the font style. In addition to formatting, the author should improve on being more concise and focus on grammatical errors. Apart from the report quality, the author should seek to better link and mention more theoretical properties of the methods they performed vis-à-vis their observations.