Deliverable 2

Problem statement:

The problem identified is classifying fake news, this is a typical classification problem, with binary outputs “true” or “fake”.

Data preprocessing:

Due to time restraints, I have decided to not incorporate the title block in this iteration of the project. In data preprocessing, I dropped all columns except the body of the text of each instance of the news article. And added on a column of binary, (0,1) representing (fake, real) news. Then two CSVs are merged into one table. One problem that I have run into is that since two datasets are merged vertically, the combined datasets need to be shuffled first then to be split into the respective sets (training, validation, testing). The dataset is then cleaned by removing all unnecessary characters, lower-case all characters, and the keyword “Reuters” is removed.In addition to the “(Reuters)” keyword, I have discovered that the reported location such as Washington is often tagged with the keyword (Reuters), this may need to be addressed in future implementations.

Naive Bayes Classification:

I have implemented the Naive Bayes classification algorithm for the simplicity and its low computational requirement, the algorithm is trained on the 80% testing data set, giving out the probability of real news/fake news and the conditional probability of each vocabulary occurring given the nature of the news. In testing. Since the testing set is already vectorised in data preprocessing, we can input it directly to the Naive Bayes classifier for a clear classification. Finally, for this initial implementation, we will simply use the percentage accuracy score of (prediction / true prediction) to get a broad overview of the result, therefore it is easy to implement changes in the future. Regarding the fitting problem, from the calculated initial accuracy scores, we can see that the validation accuracy and the testing accuracy is almost identical, we conclude that this preliminary implementation has an accurate fit.

Preliminary results:

Running the Naive Bayes algorithm on both the validation set and the test set yields around 80% accuracy, which is rather accurate for our current iteration.

Next step:

Currently I am hoping to implement the following in the future iterations of this project:

1. Hyper-parameter tuning, potentially grid searching for better accuracy
2. Implement the “Location keyword removal”
3. Implement “glove” or “word2vec” in addition to just “bag of words”