Abhijay Gupta
Student ID: 260708548

Comp 551 Assignment 3

P.S. If the metrics reported do not match those obtained from running against the saved data in Question 1, it is because of re-runs of the code. So, although the results MAY or MAY NOT exactly match, they will be in the same range/area so this shouldn't be an issue.

Bernoulli Naive Bayes Decision Trees Linear SVM
Gaussian Naive Bayes Frequency representation Binary representation

Problem 2

(a) (b) (c) (d) <<uploaded on myCourses>>

(e)

All the classifiers amongst Bernoulli Naive Bayes, Decision Trees and Linear SVM have similar performance. Linear SVM was the one that fetched the best result with a F1 score of 0.4435 on the test set. The penalty parameter (C=2) and the tolerance parameter (tol = 0.01) helped secure the best performance for Linear SVM.

Problem 3

(a) (b) (c) <<uploaded on myCourses>>

(d)

The classifiers Gaussian Naive Bayes, Decision Trees and Linear SVM had varying performances with the Frequency Bag-of-Words representation. Linear SVM had the best result with an F1 score of 0.5135 on the test set. Its parameters that led to this result were 'C' = 10 and 'tol' = 0.0001. The penalty term C here is high and

the tolerance is also low - these parameters were well tuned on the validation set and therefore led to a good result on the test set.

(e)

The classifiers in Binary BoW had very similar performance to one another while there were variations for Frequency BoW. It's interesting to see that the mean of the F1 scores (for the test sets) for the three classifiers for Frequency BoW is almost equal to the F1 score value obtained by the classifiers in Binary BoW.

Performance:

BernoulliNB	>	GaussianNB
Binary Decision Trees	~	Frequency Decision Trees
Binary Linear SVM	<	Frequency Linear SVM

(f)

Going by individual statistics, the Frequency representation seems to be better since we can choose the Frequency Linear SVM which gave the best performance overall. On the other hand, if we are open to taking any approach and want to obtain similar results for the different models, then the Binary representation is better.

Problem 4

(a) (b) (c) (d) <<uploaded on myCourses>>

(e)

The three classifiers overall had good performances. Bernoulli Naive Bayes and Linear SVM had very similar performances but BernoulliNB did better by a small margin. We chose a large range of values for the hyperparameter alpha when we tuned the BernoulliNB over the validation set. This allowed us to scope over a

large range and then select the best possible value that works out for us. The best value chosen for alpha, the Laplace/Lidstone smoothing parameter, was 0.02 over a range spanning from [0.02, 0.04, 0.06, …. 0.98, 1.00].

Problem 5

(a) (b) (c) <<uploaded on myCourses>>

(d)

Linear SVM outperformed the other two classifiers by a fair margin. It achieved an F1 score of 0.85648 over the test set. The values of the hyperparameters that led to this result were 'C' = 8 and 'tol' = 0.01. The penalty term was high while the tolerance term was lenient when compared to the penalty term. Their combination helped prevent the model from overfitting and produced a good result on the test set after being tuned on the validation set.

(e) (f) (g)

Overall, the Frequency representation perfomed better than the Binary representation. There was more variability in the results for Frequency but we obtained much higher F1 scores with it and this would be a good choice for representing our data when we are interested in performing classification.