CS_6375 Homework-1

Algorithm: Multinomial Naive Bayes
Model: Bag of Words

Dataset: enron1
Accuracy: 0.9407894736842105
Precision: 0.8860759493670886
Recall: 0.9395973154362416
f1 score: 0.9120521172638436

Dataset: enron4
Accuracy: 0.8637200736648251
Precision: 0.975975975975976
Recall: 0.8312020460358056
f1 score: 0.8977900552486187

Dataset: hw1
Accuracy: 0.9623430962343096
Precision: 0.9
Recall: 0.9692307692307692
f1 score: 0.9333333333333333

Algorithm: Discrete Naive Bayes
Model: Bernoulli

Dataset: enron1
Accuracy: 0.9342105263157895
Precision: 0.8993288590604027
Recall: 0.8993288590604027
f1 score: 0.8993288590604027

Dataset: enron4
Accuracy: 0.9060773480662984
Precision: 0.9166666666666666
Recall: 0.9565217391304348
f1 score: 0.9361702127659574

Dataset: hw1
Accuracy: 0.9728033472803347
Precision: 0.9534883720930233
Recall: 0.9461538461538461
f1 score: 0.9498069498069497

Algorithm: MCAP Logistic regression
Model: Bag of words

Dataset: enron1
Accuracy: 0.9144736842105263
Precision: 0.9661016949152542
Recall: 0.7651006711409396
f1 score: 0.8539325842696628

Dataset: enron4
Accuracy: 0.9613259668508287
Precision: 0.9490291262135923
Recall: 1.0
f1 score: 0.9738480697384807

Model: Bernoulli

Dataset: enron1
Accuracy: 0.9342105263157895
Precision: 0.8789808917197452
Recall: 0.9261744966442953
f1 score: 0.9019607843137255

Dataset: enron4
Accuracy: 0.9613259668508287
Precision: 0.9490291262135923
Recall: 1.0
f1 score: 0.9738480697384807

Algorithm: SGD classifier
Model: Bag of words
Dataset: enron1
Accuracy: 0.9890350877192983
Precision: 0.9675324675324676
Recall: 1.0
f1 score: 0.9834983498349835

Dataset: enron4
Accuracy: 0.9852670349907919
Precision: 0.9799498746867168
Recall: 1.0
f1 score: 0.9898734177215189

Dataset: hw1
Accuracy: 0.9916317991631799
Precision: 0.9772727272727273
Recall: 0.9923076923076923
f1 score: 0.9847328244274809

Model: Bernoulli
Dataset: enron1
Accuracy: 0.9956140350877193
Precision: 0.9867549668874173
Recall: 1.0
f1 score: 0.9933333333333334

Dataset: enron4
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
f1 score: 1.0

Dataset: hw1
Accuracy: 0.9937238493723849
Precision: 0.9847328244274809
Recall: 0.9923076923076923
f1 score: 0.9885057471264368

Tuning parameters for MCAP logistic regression:

- λ value is selected from the list=[1000,100,10,1,0.1,0.01,0.001] for which the accuracy is maximum. Accuracy is calculated using the validation dataset, which comes from the training dataset. The training dataset is split into 70-30 as train data and validation data respectively. Generally larger the 'C' better the model fits the data, where C=1/λ. So, the lambda value chosen is 0.001. The number of iterations taken to tune the weights while training data to choose λ is taken as 50 to limit the run time.
- The number of iterations taken to tune the weights while training data after λ is chosen is 500.

Tuning parameters for SGD classifier:

alpha: The higher the value, the stronger the regularization
Max_iter: The maximum number of passes over the training data
Tol: The stopping criterion. Training will stop when (loss > best_loss - tol)
Learning rate: optimal, invscaling, adaptive
Eta0: default 0
All the parameters were learned using the grid search of sklearn.

For enron1 dataset

1. Which data representation and algorithm combination yields the best performance (measured in terms of accuracy, precision, recall, and F1 score) and why?

Ans. Overall SGD classifier with the Bernoulli model works better than any other classifier.

2. Does Multinomial Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall, and F1 score) than LR and SGDClassifier on the Bag of words representation? Explain your yes/no answer.
Ans. No. SGD>Multinomial NB> LR

3. Does Discrete Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall and F1 score) than LR and SGDClassifier on the Bernoulli representation? Explain your yes/no answer.
Ans. No. SGD> LR> Discrete NB

4. Does your LR implementation outperform the SGDClassifier (again performance is measured in terms of the accuracy, precision, recall and F1 score) or is the difference in performance minor? Explain your yes/no answer.
Ans. Difference in performance is minor.