Name: Harsha Priya Daggubati
NetId: hxd220007

Performance Measure	Dataset 1	Dataset 2	Dataset 3
Multinomial Naive Bayes on the Bag of words model	enron	enron1	enron2
Accuracy	96.22641509433962	94.05286343612335	87.1268656716418
Precision	89.92805755395683	88.46153846153846	97.2972972972973
Recall	96.89922480620154	93.87755102040817	84.375
F1 Score	93.28358208955223	91.08910891089108	90.3765690376569

Performance Measure	Dataset 1	Dataset 2	Dataset 3
Discrete Naive Bayes on the Bernoulli model	enron	enron1	enron2
Accuracy	97.48427672955975	93.3920704845815	90.85820895522388
Precision	96.06299212598425	89.79591836734694	91.56327543424317
Recall	94.57364341085271	89.79591836734694	96.09375
F1 Score	95.31249999999999	89.79591836734694	93.7738246505718

Performance Measure	Dataset 1	Dataset 2	Dataset 3
Logistic Regression on the Bernoulli model	enron	enron1	enron2
Accuracy	86.16352201257862	87.22466960352423	90.85820895522388
Precision	85.39325842696629	98.9010989010989	94.42970822281167
Recall	58.91472868217055	61.22448979591837	92.70833333333334
F1 Score	69.72477064220184	75.63025210084033	93.5611038107753

Performance Measure	Dataset 1	Dataset 2	Dataset 3
Logistic Regression on the Bag of Words model	enron	enron1	enron2
Accuracy	91.82389937106918	89.86784140969163	88.80597014925373
Precision	94.11764705882353	94.69026548672567	88.02816901408451
Recall	74.41860465116279	72.78911564625851	97.65625
F1 Score	83.11688311688312	82.30769230769232	92.5925925925926

Performance Measure	**Dataset 1**	**Dataset 2**	**Dataset 3**
SGDC on the Bag of Words Model	enron	enron1	enron2
Accuracy	98.74213836477987	98.89867841409692	97.94776119402985
Precision	95.55555555555556	96.71052631578947	97.21518987341772
Recall	1	1	1
F1 Score	97.72727272727273	98.32775919732442	98.58793324775353

Performance Measure	**Dataset 1**	**Dataset 2**	**Dataset 3**
SGDC on the Bernoulli model	enron	enron1	enron2
Accuracy	99.79035639412998	1	1
Precision	99.23076923076923	1	1
Recall	1	1	1
F1 Score	99.61389961389961	1	1

Parameter Tuning for Logistic Regression: Eta

- First, divide the training data into train data and validation data.
- Set learning rate *Eta* to 0.01
- Set initial lambda value to 2.
- For lambda values ranging 1 to 10 with an increment of 2, calculate the model weights using the learning rate and weights of each feature in the validation data. Train the model for 25 iterations.
- For each document in validation data, calculate the conditional loglikelihood using the above model weights.
- If the conditional log likelihood is maximized, set the corresponding lambda value as the final lambda parameter.

Parameter Tuning for SGD Classifier:

- Eta0, the initial learning rate is 0.1 or 0.5
- The stopping criteria **tol** is set to 0.001 or 0.005
- Alpha value set to 0.01 or 0.05
- Max iter ranges from 500 to 3000
- Learning rate is optimal, invscaling and adaptive, which calculates the value of eta0
- All these hyper parameters are fed to GridSearchCV, and the classifier is fitted on the validation data.
- The classifier is then used to train the model on the training data, and the prediction is done on test data.

Answer the following questions:

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

As observed from the above results, Stochastic Gradient Descent Algorithm on Bernoulli Model and the Bag of Words model performed the best and gave the best accuracy. SGD performs the best since hyper parameter optimization is used, and it works better than logistic regression on large data.

2. Does Multinomial Naive Bayes perform better than LR and SGD Classifier on the Bag of words representation? Explain your yes/no answer.

Yes, Multinomial Naive Bayes performed better than LR for smaller datasets (enron and enron1). For large dataset enron2, LR performed better than MNB.

SGD performed better than MNB for bag of words model.

3. Does Discrete Naive Bayes perform better than LR and SGD Classifier on the Bernoulli representation? Explain your yes/no answer.

Yes, Discrete Naive Bayes performed better than LR on Bernoulli model. Naive Bayes assumes that the features are conditionally independent. So, it performs better on the dataset. LR predicts probability using a function form while Naive Bayes figures out how data was generated given the results, which works out better on the Bernoulli model.

SGD Classifier performed better than Discrete Naïve Bayes on Bernoulli representation

4. Does your LR implementation outperform the SGD Classifier (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.

No, SGD outperforms LR in terms of accuracy, precision, recall and F1 score. By default, SGD does not perform well, but after hyper parameter tuning and setting appropriate learning rates and number of iterations, SGD gave the best results. The difference in performance is large, and as the size of the dataset increases, SGD will continue to outperform LR, as observed in the enron2 dataset.