* **Accuracy, Precision, Recall and F1 Scores of the algorithms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **Scores** |
| **Naive Bayes algorithm** | 0.9025050778605281 | 0.9036562950570677 | 0.9069799700382843 | 0.9023793787177794 |
| **Discrete Naive Bayes Algorithm(Bernoulli Model)** | 0.5463777928232905 | 0.27318889641164523 | 0.5 | 0.3533274956217163 |
| **MCAP Logistic Regression algorithm(Bag of Words Model)** | 0.955991875423155 | 0.95538010087116 | 0.9592206254970501 | 0.9558623879319342 |
| **MCAP Logistic Regression algorithm(Bernoulli Model)** | 0.9546377792823291 | 0.9537935994246673 | 0.9574747082431707 | 0.9544834653629746 |
| **SGD Classifier (Bag of Words Model)** | 0.9580230196343941 | 0.9573690543004185 | 0.9612060515267529 | 0.9578963920388714 |
| **SGD Classifier (Bernoulli Model)** | 0.95937711577522 | 0.9585698299425804 | 0.962318518929516 | 0.9592421298953595 |

**Note:** All three data sets were combined as one for training and testing respectively.

* **MCAP Logistic Regression algorithm(Bag of Words Model)**

|  |  |
| --- | --- |
| Lambda : 1 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.955991875423155  Model Precision : 0.95538010087116  Model Recall : 0.9592206254970501  Model F1 Score : 0.9558623879319342 | Lambda : 0.5 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.955991875423155  Model Precision : 0.95538010087116  Model Recall : 0.9592206254970501  Model F1 Score : 0.9558623879319342 |
| Lambda : 0.1 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.955991875423155  Model Precision : 0.955296221570066  Model Recall : 0.9590939355268269  Model F1 Score : 0.9558558123601886 | Lambda : 0.9 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.955991875423155  Model Precision : 0.95538010087116  Model Recall : 0.9592206254970501  Model F1 Score : 0.9558623879319342 |
| Lambda : 10 Epochs: 300 learning Rate: -0.001  This may take some time  Training Completed...  Model Testing...  Model Accuracy : 0.9546377792823291  Model Precision : 0.9543175783714004  Model Recall : 0.9582348480645102  Model F1 Score : 0.9545236260330056 | Lambda : 15 Epochs: 300 learning Rate: -0.001  This may take some time  Training Completed...  Model Testing...  Model Accuracy : 0.9519295870006771  Model Precision : 0.9518372510909825  Model Recall : 0.9557565333185374  Model F1 Score : 0.9518213714877455 |

Finally: Lambda = 1 was chosen as it gave the maximum accuracy

* **MCAP Logistic Regression algorithm(Bernoulli Model)**

|  |  |
| --- | --- |
| Lambda : 1 Epochs: 300 learning Rate: -0.001  .953960731211916  Model Precision : 0.9531507402175707  Model Recall : 0.9568551295566776  Model F1 Score : 0.9538077472147408 | Lambda : 0.5 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.9546377792823291  Model Precision : 0.9537935994246673  Model Recall : 0.9574747082431707  Model F1 Score : 0.9544834653629746 |
| Lambda : 1.0 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.953960731211916  Model Precision : 0.9531507402175707  Model Recall : 0.9568551295566776  Model F1 Score : 0.9538077472147408 | Lambda : 0.9 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.953960731211916  Model Precision : 0.9531507402175707  Model Recall : 0.9568551295566776  Model F1 Score : 0.9538077472147408 |
| Lambda : 10 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.949898442789438  Model Precision : 0.9493298757317449  Model Recall : 0.9531376574377184  Model F1 Score : 0.9497546991478829 | Lambda : 15 Epochs: 300 learning Rate: -0.001  Model Accuracy : 0.949898442789438  Model Precision : 0.9493298757317449  Model Recall : 0.9531376574377184  Model F1 Score : 0.9497546991478829 |

Finally: Lambda = 0.5 was chosen as it gave the maximum accuracy

* **Answer to 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?**

The bag of words model provides the best result as highest accuracy, precesion as well as recall were found. This was found because the frequency of each word in the vocabulary was used in order to estimate the probability of the event.

1. **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.**

No, the Naïve Bayes does not perform better as the probability of each word in the vocab is considered independent. This does not consider the real world implications of natural language. Also the importance of each word in the vocab is also not considered.

Both the LR and SGD algorithms associates weightage to each word would will have real world language relations. Although, both the algorithms ignore the relationship between the words and their context, both the algorithms perform exceptionally well. Moreover the training as well as test data for the algorithms is quite large in volume. Therefore, the LR and SGD algorithms are better able to fit the data and give accurate predictions.

1. **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.**

No, the Discrete Naïve Bayes does not perform better as only the independent probability of each word is taken under consideration. In addition to this on the presence of the word in data is taken into consideration. The frequency of each in either Ham or spam is ignored resulting in algorithm performing the worst of the considered algorithms in the test cases

1. **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, both the algorithms perform very similarly as both use the same loss function. The difference being just the way weights are tweaked in SGD differs to LR which does not result in any real world improvement as an accuracy of around 95% is already good for the application case.