Question 8)

In this part, our task is to implement Naïve Bayes classification and multiclass SVM classification (with both One VS One and One VS the rest methods) for classifying multiple document classes. Respective measurements and confusion matrices are shown below.

table. Measurements of Naïve Bayes, SVM one VS one, SVM one VS rest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | F-1 Score |
| Naïve Bayes | 0.7412 | 0.7389 | 0.7407 | 0.7260 |
| SVM one VS one | 0.8715 | 0.8708 | 0.8743 | 0.8715 |
| SVM one VS rest | 0.8722 | 0.8715 | 0.8710 | 0.8712 |

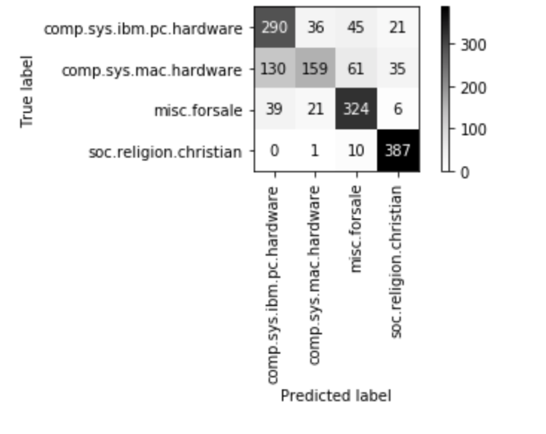


figure. Confusion Matrix of Naïve Bayes

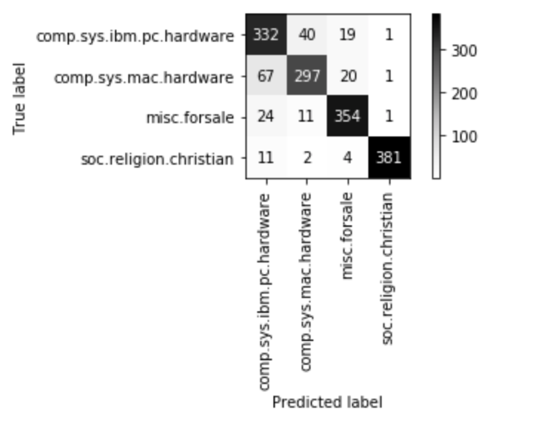


figure. Confusion Matrix of SVM one VS one

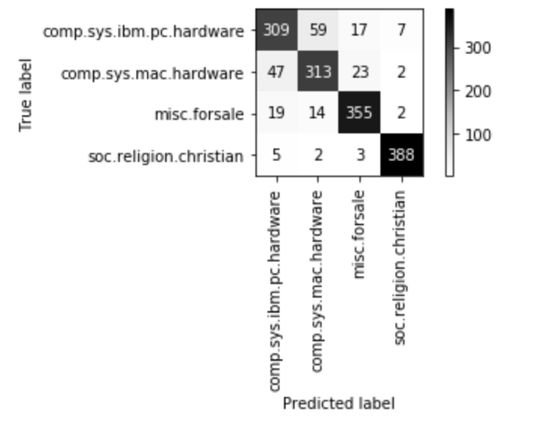


figure. Confusion Matrix of SVM one VS rest

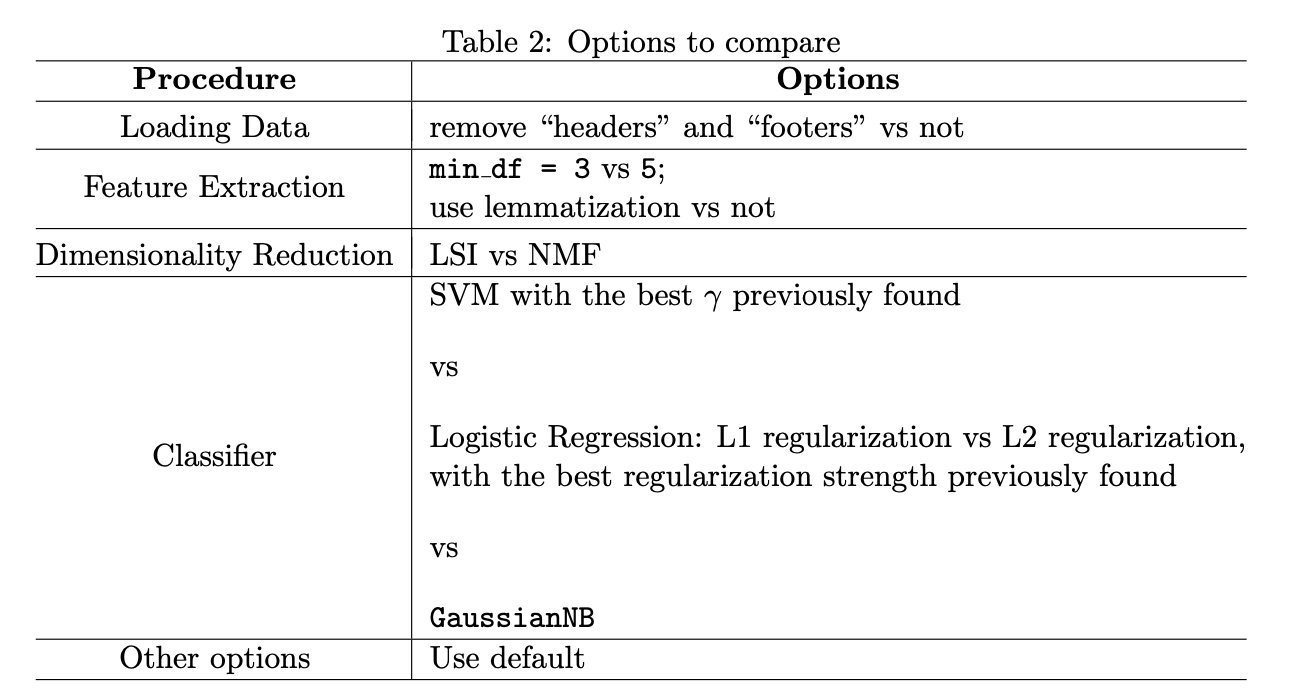
**Note**. For the data we acquired from the program, SVM (both one VS one and one VS rest) performs better then Naïve Bayes. Since Naïve Bayes is based on independent assumption, and the dataset may have items that are not independent to each other. On the other hand, the assumption of Gaussian distribution may also cause inaccuracy.

Although the difference between SVM one VS one and one VS rest is not so big, it is possible that with the number of classes increasing, the difference may be more obvious.

Question 7)

In this part, out task is to tune the parameters of the algorithms and compare the outcome of the program. We construct a pipeline implement 5-fold cross-validation to compare the options we are given, and find out which is the best combination. The options are shown below.

table. options to compare



Since the outcome can be very complicated and the data size is huge, so we check the head and tail of the outcome form, to find out the best solution of parameter combination.

table. head of the outcome

|  | with\_headers\_footers | min\_df | lemmatized | Reduce dim | classification\_method | mean\_test\_score |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | True | 3 | True | LSI | L2 Logistic | 0.975063 |
| 1 | True | 3 | True | LSI | L2 Logistic | 0.975063 |
| 2 | True | 5 | True | LSI | L1 Logistic | 0.974852 |
| 3 | True | 5 | True | LSI | L2 Logistic | 0.974429 |
| 4 | True | 3 | True | LSI | SVC with Parameter 33 | 0.974219 |

table. tail of the outcome

|  | with\_headers\_footers | min\_df | lemmatized | Reduce dim | classification\_method | mean\_test\_score |
| --- | --- | --- | --- | --- | --- | --- |
| 59 | True | 5 | True | LSI | Gaossian NB | 0.892853 |
| 60 | False | 5 | True | LSI | Gaossian NB | 0.797129 |
| 61 | False | 5 | True | LSI | Gaossian NB | 0.787193 |
| 62 | False | 5 | False | LSI | Gaossian NB | 0.772400 |
| 63 | False | 3 | False | LSI | Gaossian NB | 0.761192 |

**Note:** According to the outcome data, we can see the best combination of parameters is with headers and footers, using lemmatization, min\_df=3, the reducing dimensional method is LSI and using L2 Logistic. The mean test score is 0.975063, which is very high in our opinion.

Question 6