COL774 ASSIGNMENT 2

Khyateeswar Naidu Nalla

2019CS10376

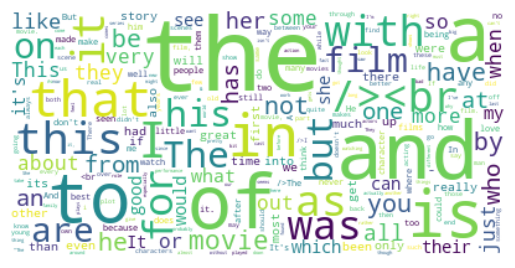
Question 1:

a)

*The accuracy over training data set by the naïve bayes model we constructed is 94.79%*

*The accuracy over testing data set by the naïve bayes model we constructed is 80.62%*

*The word cloud for positive class:*

**

*The word cloud for negative class:*

**

b)

*Test set accuracy for case of random guessing of classes=49.5%*

*Test set accuracy if we predict each class positive = 66.67%*

*The prediction accuracy in the model we used naïve bayes is 80.62%. So, this model has improved by 31.1% over the random prediction model and 13.95% over the only positive prediction model.*

*c)*

*Confusion matrix for part a:*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *7704* | *611* |
| *Negative* | *2296* | *4389* |

*Confusion matrix for part b – random prediction model:*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *4922* | *2496* |
| *Negative* | *5079* | *2503* |

*Confusion matrix for part b – positive prediction model:*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *10000* | *5000* |
| *Negative* | *0* | *0* |

*In all the three confusion matrices above, positive category has the highest value of the diagonal entry.*

*The more positive corrects may be due to the fact that the data has more positive category.*

*The pattern is that if more number of positives are predicted, the more number of positives are correct results and similar is for the negatives.*

*d)*

*After stemming and stop words removal,*

*Accuracy for the training data is 94.6*

*Confusion matrix – train data:*

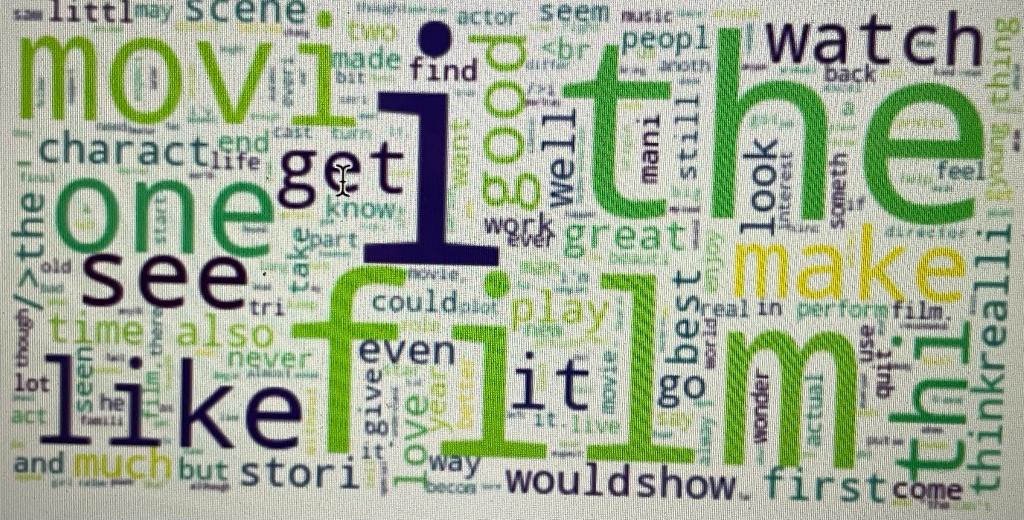
|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *11476* | *322* |
| *Negative* | *1024* | *12178* |

*Accuracy for the test data is 80.28*

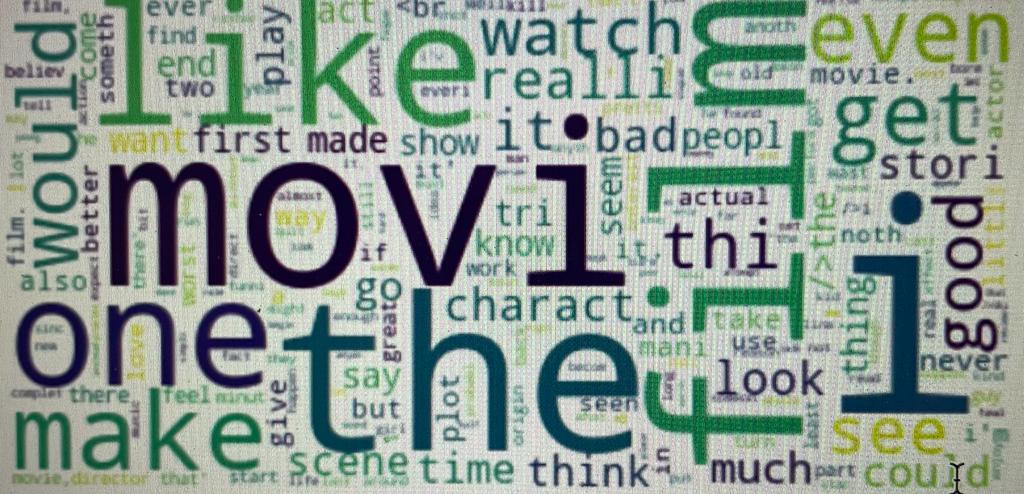
*Confusion matrix – test data:*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *7543* | *500* |
| *Negative* | *2457* | *4500* |

*Word cloud for positive class*

**

*Word cloud for negative class*

**

*e)*

*Accuracy for the test data is 56.08*

*Confusion matrix – test data:*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *3739* | *326* |
| *Negative* | *6261* | *4274* |

*Added feature is lemmatizer*

*Accuracy achieved is 56.17*

*Confusion matrix – test data*

|  |  |  |
| --- | --- | --- |
| *Predicted \ True Value* | *Positive* | *Negative* |
| *Positive* | *3755* | *329* |
| *Negative* | *6245* | *4271* |

f)

*part a:*

*precision – 92.65*

*recall – 77.04*

*F1-score – 84.12*

*part d:*

*precision – 93.7*

*recall – 75.43*

*F1-score – 83.61*

*part e:*

*precision- 91.98*

*recall – 37.39*

*F1-score – 53.16*

*The best is F1-score because accuracy, precision or recall is best when the no of positive testcases and negative testcases but not with the case of F1-score so it is the best*

Question 2:

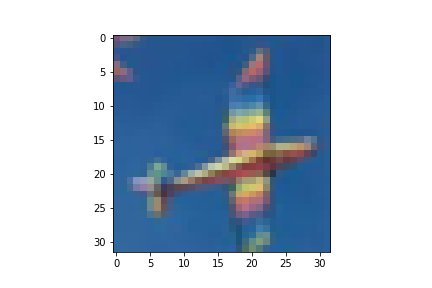
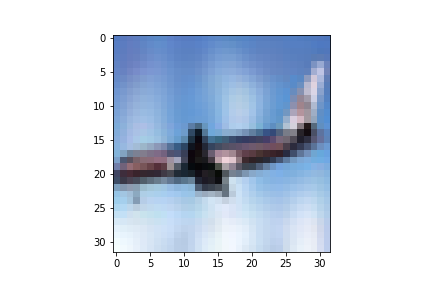
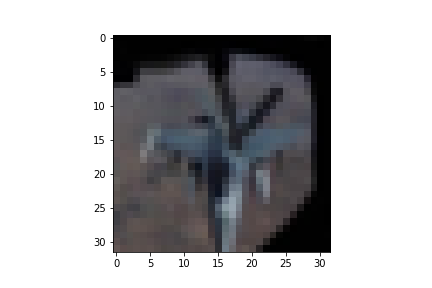
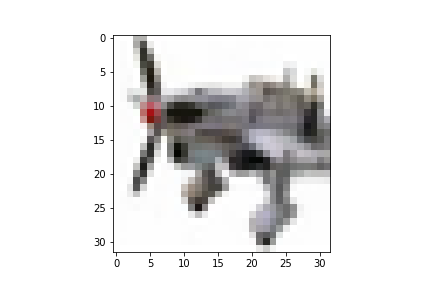
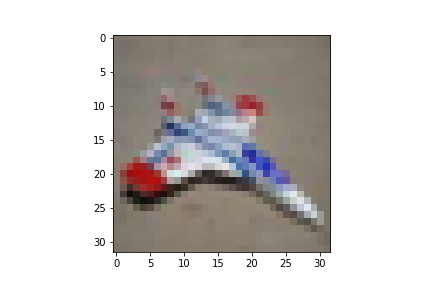
a)

*My entry number is 2019CS10376. So I am taking the image from classes 1 and 2. The number of support vectors I got in this case is: 4000*

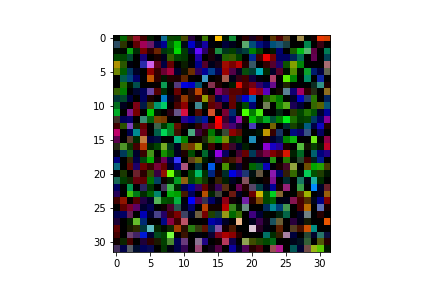
*100% of the training examples constitute support vectors.*

*After calculating w, b and classifying the test data, accuracy of test data I got is 78.15%*

*Time taken = 65.32 s*

**

*Weight plot*

**

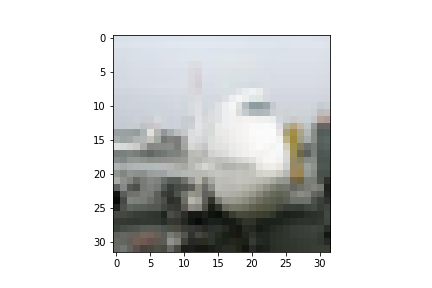
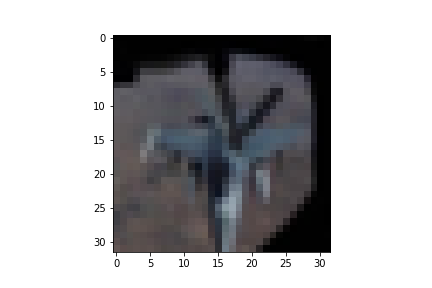
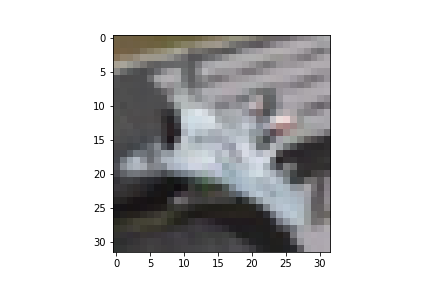
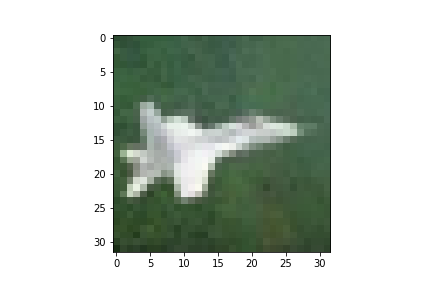
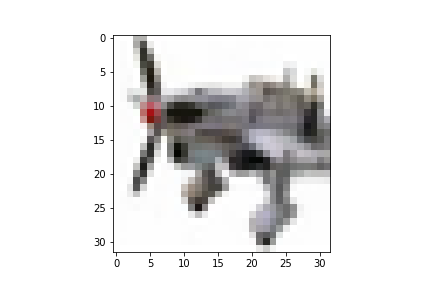
b)

*The number of support vectors I got in this case is: 1872*

*After calculating w, b and classifying the test data, accuracy of test data I got is 87.5%*

*Time taken = 61.42 s*

*No of similar support vectors to linear is 1161*

**

c)

*Linear:*

*The number of support vectors I got in this case is: 1541*

*After calculating w, b and classifying the test data, accuracy of test data I got is 78.05%*

*Time taken = 26.19 s*

*No of similar support vectors to linear is 1161*

*Gaussian:*

*The number of support vectors I got in this case is: 1865*

*After calculating w, b and classifying the test data, accuracy of test data I got is 87.4%*

*Time taken = 15.76 s*

*No of similar support vectors to linear is 1159*

Question 3:

*a)*

*accuracy = 59.04*

*time taken = 625.86*

*Confusion matrix*

predicted

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *738* | *74* | *84* | *54* | *50* |
| *118* | *721* | *42* | *84* | *35* |
| *150* | *60* | *412* | *128* | *250* |
| *96* | *97* | *123* | *561* | *123* |
| *99* | *45* | *217* | *119* | *520* |

b)

*accuracy = 59.3*

*time taken = 187.60*

*Confusion matrix*

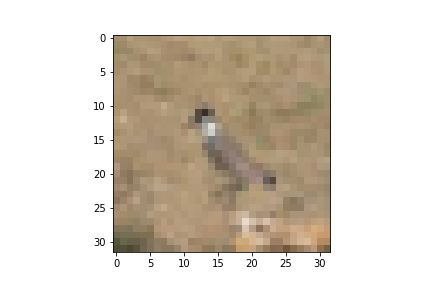
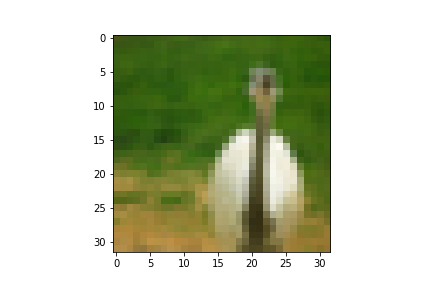
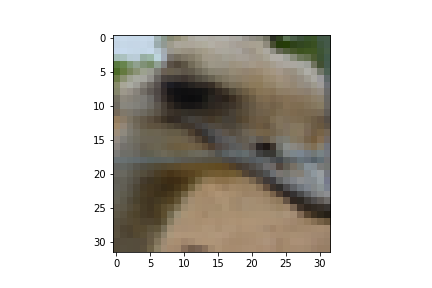
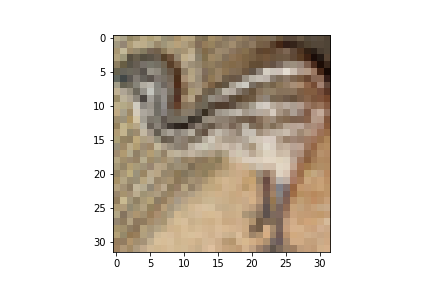
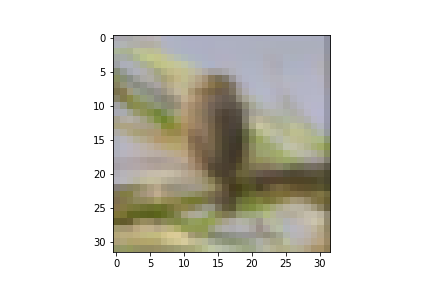
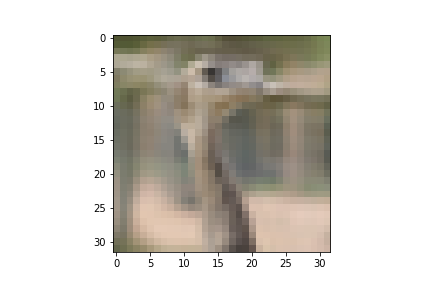
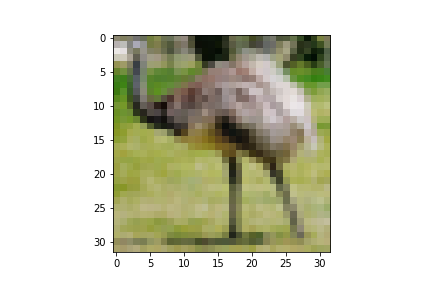
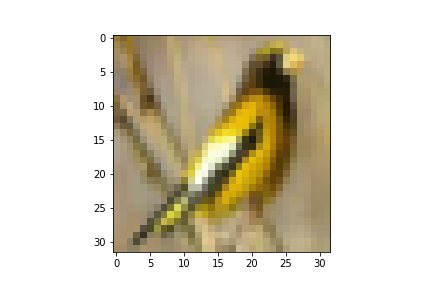
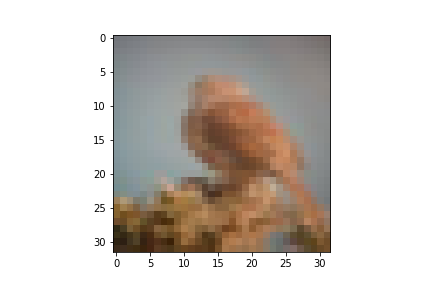
predicted

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *729* | *83* | *78* | *61* | *49* |
| *100* | *731* | *42* | *92* | *35* |
| *145* | *61* | *409* | *133* | *252* |
| *82* | *97* | *123* | *572* | *126* |
| *98* | *48* | *212* | *118* | *524* |

c)

Most Misclassified images

There are labeled 2 but they were predicted 4



Libraries used – numpy,nltk,sklearn,cvxopt,glob,pickle,sys,time,matplotlib,word cloud,PIL,math,random