Assignment 2

Q1 Text Classification

a.i)

Training Data Accuracy: **93.81%**
Test Data Accuracy: **82.67%**

a.ii)

Positive model word cloud

Negative model word cloud

b.i)

Test Data Accuracy random guessing: **49.52%**

b.ii)

Test Data Accuracy predicting each sample as positive: **57.48%**

b.iii)

Compared to random guessing our trained model is **65.35%** more accurate.

Compared to predicting each sample as positive out model is **42.39%** more accurate.

c.i)

Format of Confusion matrix:

[TP | FP]

[FN | TN]

Test Data Confusion Matrix:

[[8011. 1989.]
 [611. 4389.]]

Test Data Confusion Matrix random guessing:

[[4309. 3211.]
 [4313. 3167.]]

Test Data Confusion Matrix predicting each sample as positive:

[[8622. 6378.]
 [0.   0.]]

c.ii)

For our original test data confusion matrix **True Positive** has the highest value in the diagonal elements also because there are more positive examples then negative ones in our test data

For our original random data confusion matrix **True Positive** has the highest value in the diagonal elements

For our original test data confusion matrix **True Positive** has the highest value in the diagonal elements also because we are classifying each test data as positive one

c.iii)

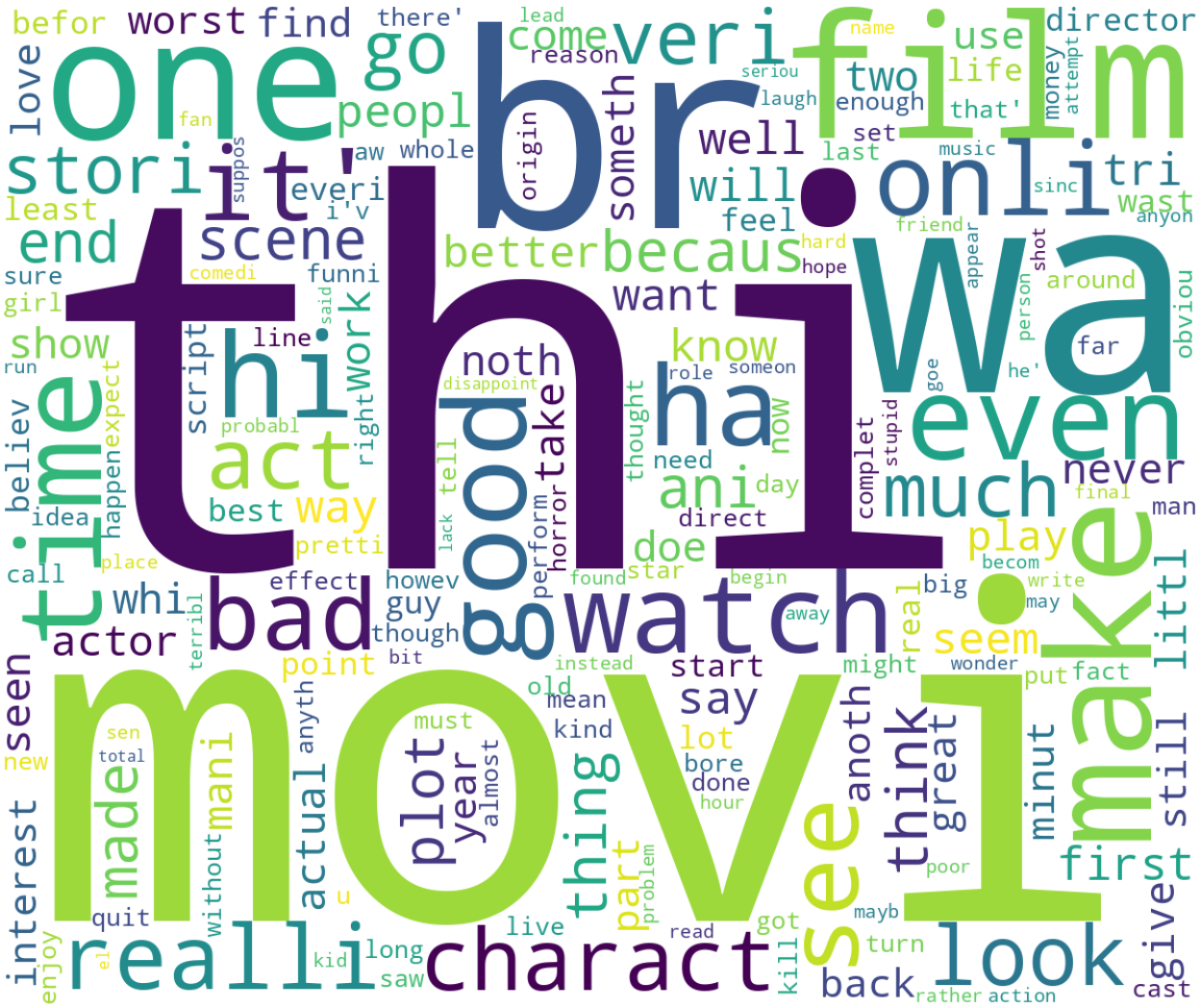
It doesn't seem that the confusion matrix is following any kind of pattern other than that the highest value is always for the True Positive index, which is just luck in case of random prediction so no pattern.

d.i)

Implemented in program

d.ii)

Negative model word cloud



Positive model word cloud

d.iii)

Test Data Accuracy: **84.74%**

Test Data Confusion Matrix:

[[8572. 1428.]
[861. 4139.]]

d.iv)

Accuracy only slightly increased from **82.67%** to **84.74%**

Maybe removing more stopwords can increase the accuracy further.

e.i) **Bigrams**

Training Data Accuracy: **99.66%**

Training Data Confusion Matrix:

[[12432. 68.]
[13. 12487.]]

Test Data Accuracy with Bigrams: **84.97%**

Test Data Confusion Matrix:

[[8585. 1415.]
[840. 4160.]]

e.ii) **Bigrams and Trigrams**

Test Data Accuracy with Bigrams and Trigrams: **84.95%**

Test Data Confusion Matrix:

[[8584. 1416.]
[841. 4159.]]

e.iii) **Improvements**

Test Data

Original test data accuracy was **82.67%** after performing stemming and removing the stop-words and adding new bigrams and trigrams Features accuracy slightly improved to **84.95%**

Training Data

Original training data accuracy was **93.81%** after performing stemming and removing the stop-words and adding new bigrams and trigrams Features accuracy improved significantly to **99.66%**

f.i)

Original Test Data

Positive model

Precision: 0.80
Recall: 0.93
F1-score: 0.86

Negative model

Precision: 0.88
Recall: 0.69
F1-score: 0.77

After stemming and stopwords removal

Positive model

Precision: 0.86
Recall: 0.91
F1-score: 0.88

Negative model

Precision: 0.83
Recall: 0.74
F1-score: 0.78

(BEST PERFORMING MODEL)

After stemming and stopwords removal and adding bigram features

Positive model

Precision: 0.86
Recall: 0.91
F1-score: 0.88

Negative model

Precision: 0.83
Recall: 0.75
F1-score: 0.79

Q2 Binary Image Classification

a.i)

Number of support vectors: **1505**
75.25% of training sample constitute the support vectors

a.ii)

w and b values:
w: [-0.40311351 -0.09715971 -0.99773958 ... -0.48383299 0.04012729
-0.53011341]
b: 1.5880446236357777

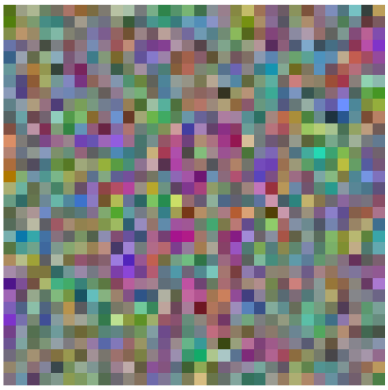
Accuracy
Test set accuracy: **79.00%**

a.iii)

Top-5 negative class coefficients

Top-5 positive class coefficients


W(weight vector)



b.i)

Number of support vectors: **1754**
Common support vectors between linear and gaussian model: **1126**

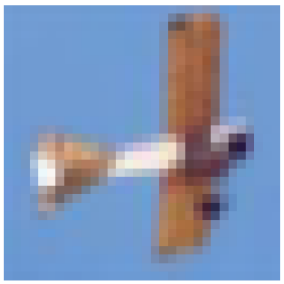
b.ii)

Accuracy
Test set accuracy: **85.75%**

b.iii)

Top-5 class coefficients






b.iv)

Test set accuracy with linear kernel: **79.00%** where as test set accuracy with gaussian kernel: **85.75%**

c.i)

Linear kernel

Number of support vectors in **sklearn** model: **1494**
Number of support vectors in **cvxopt** model: **1505**
Common support vectors between **cvxopt** and **sklearn** model: **1494**

Gaussian kernel

Number of support vectors in **sklearn** model: **1743**
Number of support vectors in **cvxopt** model: **1754**
Common support vectors between **cvxopt** and **sklearn** model: **1743**

c.ii)

Root Mean Squared Error W: **6.498026331617066e-06**
Absolute Error B: **0.0230558785460373**

c.iii)

Accuracy sklearn, linear model: **79.10%**
Accuracy sklearn gaussian model: **85.80%**

c.iv)

Execution time in seconds

LINEAR

sklearn- **21.284178495407104**
cvxopt- **62.71004605293274**

GAUSSIAN

sklearn- **12.902670860290527**
cvxopt- **64.00805234909058**

Q3 Multiclass Image Classification

a.i) Test set accuracy of **cvxopt** multiclass gaussian model: **59.0%**

b.i)

Test set accuracy with **sklearn** multiclass gaussian(ovo) model: **59.3%**

b.ii)

Training time in seconds **sklearn** model**: 141.9613401889801**
Training time in seconds **cvxopt** model: **1617.354914284545**

c)

Confusion Matrix cvxopt **Linear model:**

[[792, 212],
[208, 788]]

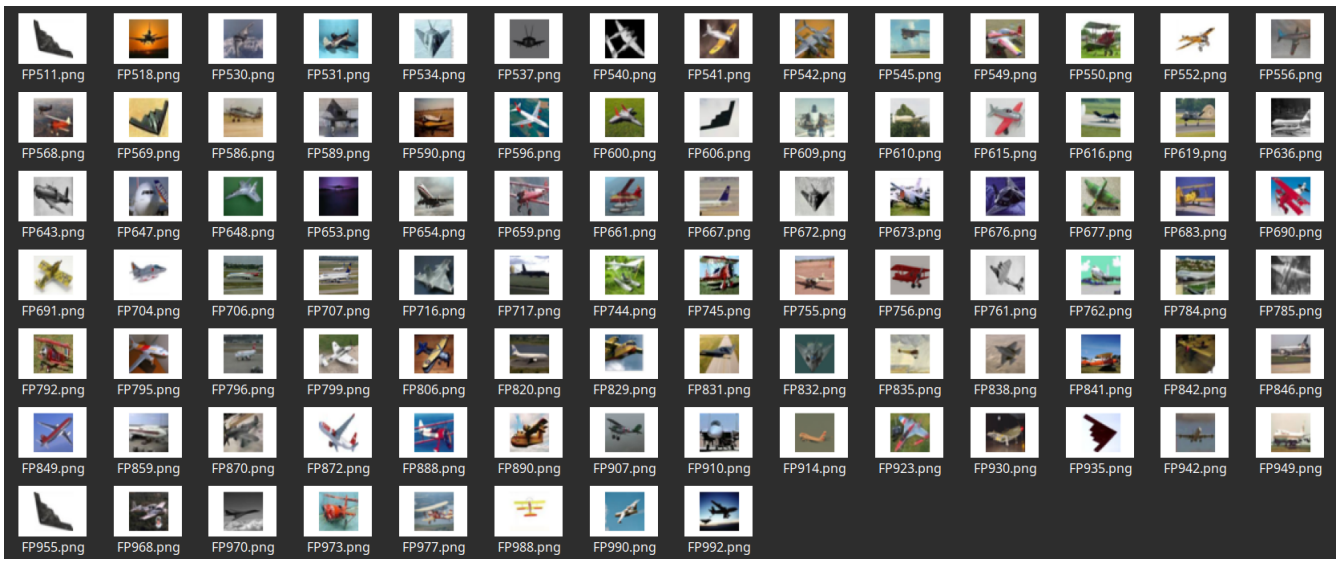
Confusion Matrix cvxopt **Gaussian model:**

[[858, 143],
[142, 857]]

Observation:
Almost equal number of positive and negative images are correctly Classified

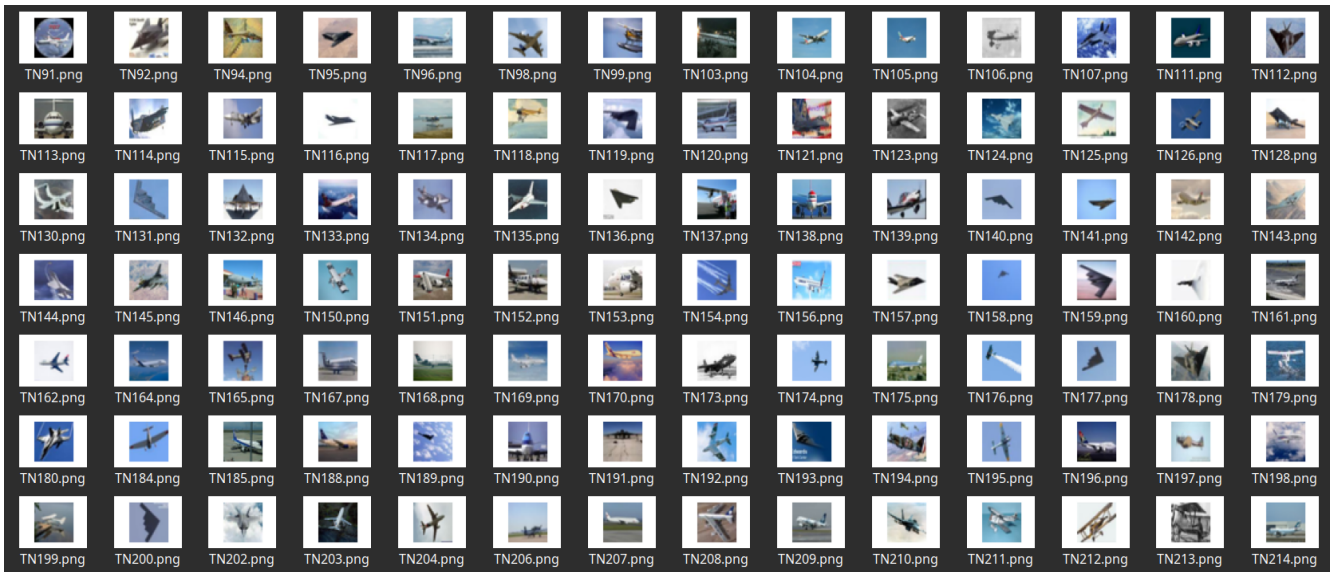
These are **False Positive**

I.e These are wrongly classified as deer class



These are **True Negative**

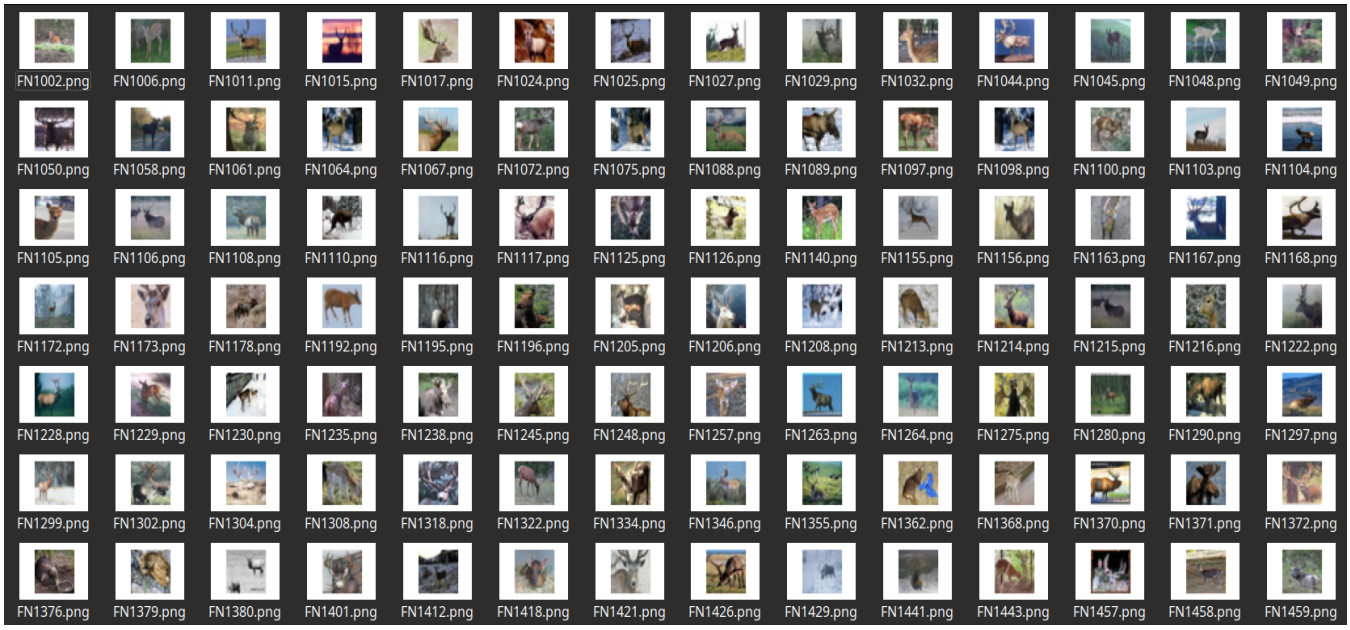
I.e These are correct plane class according to our model



We can clearly see that the background of the truly classified images are clear and mostly blue

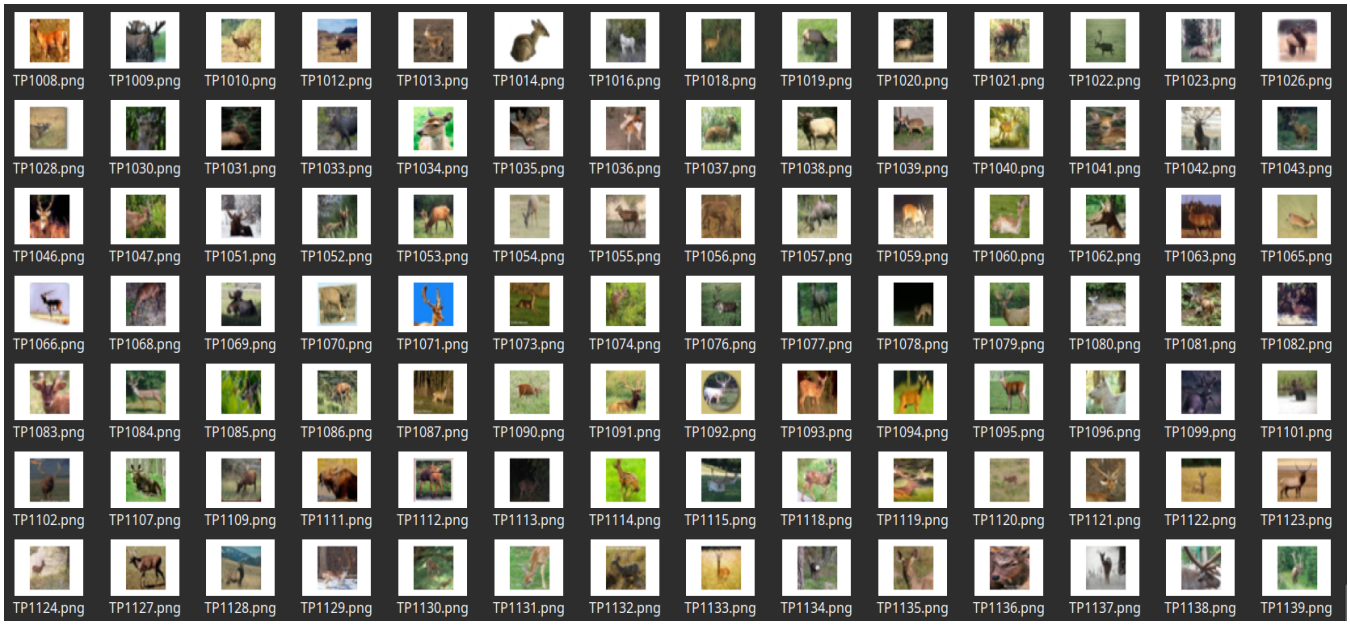
These are **False Negative**

I.e These are wrongly classified as plane class



These are **True Positive**

I.e These are correct plane class according to our model



We can clearly see that the background of the truly classified images are clear and mostly green