COL774 Assignment 2
Aryan Dua

Libraries Used: cvxopt, sklearn, spacey, nltk, wordcloud, PIL, pickle

Q1. a. i) Training Set Accuracy - **91.64%**
Test Set Accuracy - **80.14%**

ii) Word Cloud for positive reviews - wordcloud1.jpg in plots folder
Word Cloud for negative reviews - wordcloud2.jpg in plots folder

b. i) Test Set Accuracy with random predictions - **50.62%**
ii) Test Set Accuracy with positive predictions - **66.67%**
iii) It improves a lot compared to the positive prediction baseline and even more compared to the random prediction baseline. The exact percentages are **13.81%** and **29.52%**.

c. i) General format of a Confusion Matrix -

TP	**FP**
TN	**FN**

Confusion Matrix for Part a) -

7446	**424**
2554	**4576**

Confusion Matrix for Part b) i) -

5058	**2465**
4942	**2535**

Confusion Matrix for Part b) ii) -

10000	**5000**
0	**0**

ii) The first element in the first row for each confusion matrix has the highest value of diagonal entry. This represents the True Positive Cases. It implies that the program is classifying more amount of positive labeled data correctly, but we should note that in the data, we have more amount of positive data than negative data(2:1 ratio).

iii) The sum of elements along a column equals the number of examples of that label present in the test data. The number of diagonal elements represents the correctly classified data, and the non-diagonal elements represent the incorrectly classified data.

d. i) Stemming and stopword removal performed.
Training Set Accuracy - **93.24%**

ii) Word Cloud for positive reviews - **wordcloud3.jpg** in plots folder
Word Cloud for negative reviews - **wordcloud4.jpg** in plots folder

iii) Test Set Accuracy - **80.27%**

iv) Accuracy on the test set increases from before. This is because the stopwords like if, but, else, they act as noise to the data. They increase computation but do not provide any meaning. Given the word “If”, even humans cannot make any conclusions of whether it is a word representing a positive review or a negative. Such words will spoil the model's accuracy, which is why it is better to remove them.

e. i) Test Set Accuracy with Bigrams, without stopword removal - **80.36%**
Test Set Accuracy with Bigrams, with stopword removal - **80.22%**
Both these models show close to 100% training accuracy, which is a clear sign of overfitting.
ii) Test Set Accuracy with Trigrams - **79.99%**
iii) Yes, they help in improving overall training accuracy, but the test accuracy more or less remains the same. With increasing ‘n’ of the n-grams, it gets more and more familiar with the given data and how the position of the words are relative to each other. In other words, overfitting increases with n. If n is made 3(trigrams), then the training accuracy reaches the 100% mark.

f. i)

	Training Data	Test Data
Precision	0.999	0.942
Recall	0.98336	0.7435
F1-score	0.9911	0.831

ii) I think f1 score is a more suitable metric for this kind of dataset because the samples don't have an equal number of positive and negative labels. There are 10000 positive labeled examples compared to only 5000 negative labeled examples. Accuracy is a direct measure of examples classified correctly, it does not take into account the classes they come from. F1 score is derived from precision and recall, so this matter of concern is covered.

Q2. a)

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^i y^j (x^i)^T x^j - \sum_{i=1}^{m}\alpha_i$$

This is the dual problem we have to solve using the CVXOPT package. It will take inputs based on these equations and constraints:

$$\alpha^T P\alpha + q^T\alpha + d$$
$$G\alpha \preceq H$$
$$A\alpha = b$$

Note - My Entry Number ends with 5.

i) I am getting **1082** support vectors. This is equivalent to **27.05%** of the training samples.

ii) Test Set Accuracy - **77.8%.**

b = **1.608**.
It was found using - avg(y-Xw) over relevant alphas.

w was found by summing alpha*y*X.

w =

```
[[0.4952489 ]
 [0.2547612 ]
 [0.15746791]
 ...
 [0.381453  ]
 [0.23654375]
 [0.50731099]]
```

iii) The images corresponding to the top5 coefficients have been plotted and are stored in the plots directory by the names - **linear_best5_i.jpg**, where i = 1,2,3,4,5

b)

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^i y^j \phi(x^i)^T\phi(x^j) - \sum_{i=1}^{m}\alpha_i$$

This is the dual SVM Problem we have to solve using kernels. The only difference from the previous part is that instead of x, we have ϕ(x).

i) nSV for Gaussian Kernel - **4000**. The best accuracy reported was when all the alphas were considered. Therefore, I have considered all of them to be the support vectors. The linear case support vectors are a subset of these.

ii) Test Set Accuracy - **85.9%**

iii) The images corresponding to the top5 coefficients have been plotted and are stored in the plots directory by the names - **gaussian_best5_i.jpg**, where i = 1, 2, 3, 4, 5

iv) The test set accuracy of the Gaussian Kernel version is much better than the linear approach, and this was quite intuitive since we are expanding our feature space to infinite dimensions in this case, vs. a linear space of features in the earlier case.

c) i) nSV for Gaussian Kernel using LIBSVM Package - **1972**
nSV for Linear Kernel using LIBSVM Package - **1621**

ii) Weight obtained using Linear Kernel - **vector of size 3072**
Bias obtained using Linear Kernel - **1.71**

iii) Test Set Accuracy using Gaussian Kernel - **85.8%**
Test Set Accuracy using Linear Kernel - **78.1**

iv) Training Times:
CVXOPT, Gaussian - **396.655**
CVXOPT, Linear - **43.896**
LIBSVM, Gaussian - **45.9**
LIBSVM, Linear - **51.03**

Q3. a) i) Test Set Accuracy - **59.38%**

b) i) Test Set Accuracy - **59.3%**

Training Time - CVXOPT	Training Time - LIBSVM
86.53 minutes	**14.12 minutes**

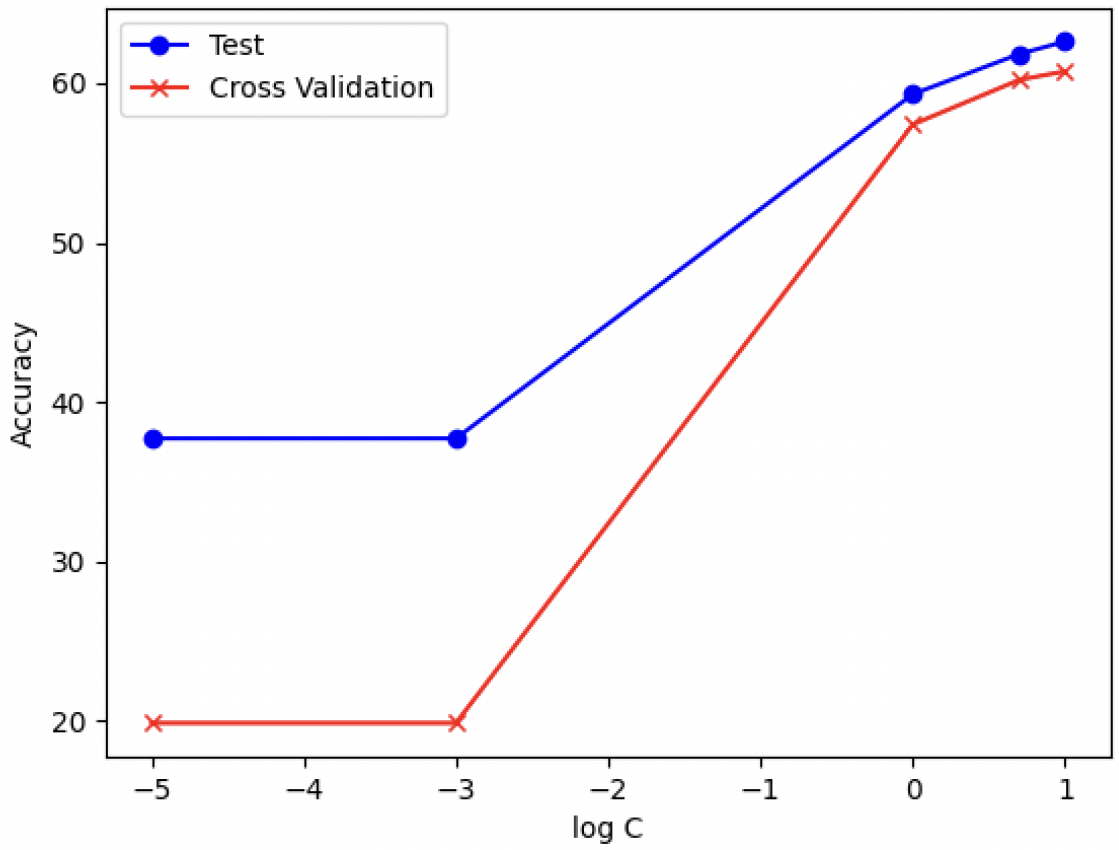
ii) LIBSVM is much faster than CVXOPT(**6 times as fast**), and their accuracies are quite similar as well.

c) The Confusion matrices for both the parts, as well as the ten misclassified examples, are saved in the plots folder of this question. Class 4 is often misclassified as class 2, and vice versa. Similarly, Class 3 and 4 are also often misclassified. This was based on observing the non-diagonal entries in the confusion matrix. Yes, the results make sense because, visually, the classes are quite similar to the human eye as well.

d)

C	5-fold Cross-Validation Accuracy	Test Accuracy
1e-5	19.88	37.72
1e-3	19.88	37.72
1	57.4	59.3
5	60.2	61.82
10	60.73	62.6

Plot of Log(c) vs Accuracies:



Observations -

C=10 gives the maximum value of both cross-validation and test accuracy. The values of C tend to be below the test accuracy. This shows that the model has an underfitting characteristic. The graph tendency shows that the accuracy is reaching a maximum, and therefore, on increasing C further, it will tend to overfit the training data. On other systems, the maxima occurred at C=5, but in mine, the accuracy at 10 is higher than at 5.