CSCI 544- Applied Natural Language Processing

HW 1

Question 1)

- Average length of reviews before and after data cleaning (with a comma between them)

```
In [18]: print("Average length of reviews before and after data cleaning:",dt_before_clean,',',dt_after_clean)
         Average length of reviews before and after data cleaning: 269.4664 , 261.4241666666667
```

Question 2)

Average length of reviews before and after data preprocessing (with comma between them)

```
In [25]: print("Average length of reviews before and after data preprocessing:",dt_before_preprocess,',',dt_after_preprocess)
         Average length of reviews before and after data preprocessing: 261.4241666666667 , 251.40098333333333
```

Question3)

Precision, Recall, and f1-score for the testing split in 4 lines (in the order of rating classes and then the average) for Perceptron (with comma between the three values)

```
----------------------------------Perceptron Classification----------------------------------------------------------------
class 1 Precision: 0.7201382033563672 , Recall: 0.7256901268341208 , F1 score: 0.7229035055122012 , average: 0.7229106119008963
class 2 Precision: 0.6285034373347436 , Recall: 0.6080839089281146 , F1 score: 0.6181250812638149 , average: 0.6182374758422243
class 3 Precision: 0.7959673547767643 , Recall: 0.8147420147420148 , F1 score: 0.805245264691598 , average: 0.8053182114034589
```

Question 4)

Precision, Recall, and f1-score for the testing split in 4 lines (in the order of rating classes and then the average) for SVM (with comma between the three values)

```
----------------------------------Support Vector Machine Classification----------------------------------------------------
class 1 Precision: 0.7509930486593843 , Recall: 0.7523004227804029 , F1 score: 0.7516461672257424 , average: 0.7516465462218432
class 2 Precision: 0.6531625718766335 , Recall: 0.639293937068304 , F1 score: 0.6461538461538462 , average: 0.6462034516995945
class 3 Precision: 0.8210323203087313 , Recall: 0.8363636363636363 , F1 score: 0.8286270691333981 , average: 0.8286743419352552
```

Question 5)

Precision, Recall, and f1-score for the testing split in 4 lines (in the order of rating classes and then the average) for Logistic Regression (with comma between the three values)

```
-------------------------------------------Logistic Regression Classification-------------------------------------------------
class 1 Precision: 0.7474368592148037 , Recall: 0.7433474260134295 , F1 score: 0.7453865336658354 , average: 0.7453902729646895
class 2 Precision: 0.6449257114077058 , Recall: 0.655154771041187 , F1 score: 0.65 , average: 0.6500268274829643
class 3 Precision: 0.8267990074441688 , Recall: 0.8186732186732186 , F1 score: 0.8227160493827161 , average: 0.8227294251667011
```

Question 6)

Precision, Recall, and f1-score for the testing split in 4 lines (in the order of rating classes and then the average) for Naive Bayes (with comma between the three values)

```
-------------------------------------------Naive Bayes Classification---------------------------------------------------------
class 1 Precision: 0.7722342733188721 , Recall: 0.7082815220094504 , F1 score: 0.7388766376961993 , average: 0.7397974776748405
class 2 Precision: 0.5786831434639892 , Recall: 0.7666922486569455 , F1 score: 0.6595510563380282 , average: 0.6683088161529875
class 3 Precision: 0.8962655601659751 , Recall: 0.6899262899262899 , F1 score: 0.7796751353602667 , average: 0.7886223284841772
```

During the Data Cleaning phase, I have implemented the following actions on the dataset-

Lower case all the strings in the dataset, while strip is used to remove the extra spaces before and after the strings, Contractions is used to remove all the words which are shortened by dropping letters and replacing them by apostrophe.

The contraction step comes before the punctuation removal step because if we use the punctuation removal step first, the apostrophe, which is a characteristic feature of removing contractions, will be lost, and the contraction step will be useless and will not affect the dataset, resulting in random words in the strings that make no sense.

After contractions are removed, the main non-contributing things are the HTML and URL tags, which are removed using Regex library of Python and the extra white spaces are also removed between words.

After the Data Cleaning step, Data Pre-processing is used to remove the non-essential words which do not contribute much to the sentiment of the review,

For this purpose, firstly I removed the stop-words, but the precision come out to be less, around 63% but when I tried to perform the models without removing the stop words, the precision came out to be better compared to with removing stop words around 70%.

After that, Lemmatization is used to extract the base words, which contribute more towards the sentiments rather than using the actual words which can be deceiving towards extracting the sentiments of the sentences.

After the Data Pre-processing step, TF-IDF feature extraction is used get the important features out of all the sentences, which directly contribute getting the exact sentiment of the sentences.

Then comes the step to implement all the Machine Learning models namely- Perceptron, Support Vector Machine, Logistic regression and Naïve Bayes, and the result are as follows-

Perceptron-

	precision	recall	f1-score	support
class 1	0.72	0.73	0.72	4021
class 2	0.63	0.61	0.62	3909
class 3	0.80	0.81	0.81	4070
accuracy			0.72	12000
macro avg	0.71	0.72	0.72	12000
weighted avg	0.72	0.72	0.72	12000

Support Vector Machine-

	precision	recall	f1-score	support
class 1	0.75	0.75	0.75	4021
class 2	0.65	0.64	0.65	3909
class 3	0.82	0.84	0.83	4070
accuracy			0.74	12000
macro avg	0.74	0.74	0.74	12000
weighted avg	0.74	0.74	0.74	12000

Logistic Regression-

	precision	recall	f1-score	support
class 1	0.75	0.74	0.75	4021
class 2	0.64	0.66	0.65	3909
class 3	0.83	0.82	0.82	4070
accuracy			0.74	12000
macro avg	0.74	0.74	0.74	12000
weighted avg	0.74	0.74	0.74	12000

Naïve Bayes-

	precision	recall	f1-score	support
class 1	0.77	0.71	0.74	4021
class 2	0.58	0.77	0.66	3909
class 3	0.90	0.69	0.78	4070
accuracy			0.72	12000
macro avg	0.75	0.72	0.73	12000
weighted avg	0.75	0.72	0.73	12000

While Doing the Homework, I have learnt how to do real-life sentiment analysis on the reviews which are on some business platform like Amazon reviews.

The main task of any sentiment analysis is to remove the unnecessary words which doesn't contribute much to the task in hand, while keeping the important words or part of sentences which do contribute. Performing tasks like removing stop words, lemmatization really help me understand the intricacies of how to perform the sentiment analysis perfectly and handle real-life dataset.

While going about the Precision, Recall, and F1-score, I learned about the different metrics which help evaluate the performance of the models and also learned how to increase those metrics values.