1. What are the text cleaning methods you tried? What are the ones you have included in the final code?

Ans:
I did text cleaning to remove any numbers from file, all the hyperlinks from the comments file, digits from the file, I removed the stop words from the comments file. Out of all other methods, the last method proved out to be the most useful and impactful.

2. What are the features you considered using ? What features did you use in the final code?
Ans: I worked with both TfIdfVectorizer and CountVectorizer but CountVectorizer was giving better results , thus I tried using CountVectorizer with different parameters, max_df, min_df, and by using analyser to use English stemmer , and also strip_accents{“ascii”} to take just ascii values
3. What optimizations did you add in your code ?
Ans: I performed data cleaning by removing any numbers, hyperlinks, new lines, tabs. I also tried to perform stemming using English Stemmer. That really improved my results.
Once I got good results with Logistic Regression, I tried tuning Logistic Regression’s hyperparameters - C, Dual. Also Tried to get the optimum value of C using GridSearchCV.
4. What are the ML methods you tried out, and what were your best results with each method? Which was the best ML method you saw before tuning hyperparameters?
Ans: I tried out MultinomialNB, SVM, LinearSVC, Logistic Regression, Random Forest Classifier, Multiperceptron
MultinomialNB: 0.932
SVM: 0.942
Random Forest: 0.726(maxdepth = 2)
0.813(Maxdepth =5)
0.833(Maxdepth=8)
0.874(n_estimators=20)
0.879(n_estimators=25, max_depth=10)
0.924(n_estimators=25, max_depth=20)
0.934(n_estimators=80, max_depth=30)
0.942(n_estimators=100, max_depth=50)
0.952(n_estimators=200, max_depth=100)

MLPClassifier:
Default =0.500
Logistic Regression (with tfidfvectoriser): 0.954
Logistic Regression(with countvectorizer): 0.957
Logistic Regression(With countvectorizer and max_features=100000) : 0.959

I also tried using GridSearchCV with Logistic Regression, but It decreased my score from 0.962 to just 0.957

The ML method that gave the best results was Logistic Regression.

5. What hyperparameter tuning did you do ?
I tried tweaking the C value of 0.1, 100, 1000 But all of this declined the roc score from 0.962.
I also tried using GridSearchCV. But that also didn’t help in increasing the score.
When I included the hyperparameters dual=True and penalty=’l2’ , Even tried using dual =False, But it just doesnot increase score at all. It was stuck at 0.962.

6. What did you learn from the different metrics? Did you try cross-validation?
Ans: precision and recall values shows the measure of relevance of the system. And support helped me know about the number of samples of true response that lies in that class.

7. What are your best final Result Metrics? By how much is it better than the strawman figure? Which model gave you this performance?
Ans: My best result was ruc score 0.962. Which is 0.005 better than strawman figure. Logistic regression used with snowball stemmer and countvectorizer where max features are 1000000 gave me that result.

8. What was the hardest thing to do ?
Ans: The hardest thing was having no background about Machine learning and thus struggling to figure out the ML Algorithms and understanding what it does was really challenging.