Answer 4

a) The final score for my Kaggle file was 0.0000 and my username is Yashika Bajaj. The prediction file is also attached along with the mail.

b) The error analysis of my final classifier is as follows:

```
    ...: print(confusion_matrix(Y_test, y_pred1))
[[725 204]
 [185 502]]
```

On using the decision tree model for the predictions on the test dataset, I checked the differences in the original dataset and the predictions. After analyzing the differences of my model predictions and the data in the training set, I saw that there were some values in "corr" in the training set which were true, but my model is predicting that they are false and vice versa. This made me to create an error matrix to check the overall ratio of correct and incorrect predictions. It can be seen in the error matrix that from the total of 1616 predictions, there are 389 wrong predictions and 1227 correct predictions.

This could be due to multiple reasons. One of the reason could be the efficiency of the model which was 75.93% therefore, some of the results were supposed to be wrong. This error could be because I have considered the length of the text of the column "text" to be a factor to analyze the prediction of "corr". But my understanding could be wrong and there can be instances in the training set where the length is small, and the answer is correct which is opposite to what I was thinking and hence made the model. And thus, the model made has this defect and the results predicted are wrong.

Say for example there is an instance in the data – row number 7935 where the length of the text is small but the "corr" is true however according to my model the "corr" predicted is false. This proves that length cannot be the only factor for deciding the "corr". Another row number 13444 which has large length of the text but the "corr" in the data is false is predicted as "true" by my model. Hence, there is a need to include some more parameters which will help in making predictions more accurately.