4. I have chosen the decision tree model as the best classifier based on my findings and will be using this to make predictions on the test dataset.

   a. I have uploaded the file on kaggle where my user name is Saurabh Thakrani and got a score 0.00. I have attached the prediction file in the mail also.

   b. Error Analysis-
      For the decision tree model, I have created an error matrix and is shown below.

```
In [52]: print(confusion_matrix(Y_test, y_pred1))
[[737 192]
 [184 503]]
```

      Here, we can see that there are 192 False positives and 184 False negatives. Overall, out of 1616 predictions, 1240 predictions are correct and 376 are incorrect. So, there can be several reasons for this. According to my knowledge, the first thing is the accuracy of the model. My model has an accuracy of 76.73% and hence some error is expected. This may happen because there are only a few tournament years where there is a big difference in the no. of correct and incorrect answers. Also, the length of the text cannot define a clear threshold to separate correct and incorrect answers. Sometimes the length of the text may be small and sometimes large for the correct answer. Also, the answer type "work" has more correct answers but the small amount of incorrect answer cannot be neglected. All these things lead to some error in the prediction by the model.
      I analyzed a few rows that were predicted incorrectly. In row 7008, the text length is large, tournament year is 2007 and answer type is work. All the parameters give a hint of the answer being False and it is False in the dataset, but the model shows as True.
      In row 5758, the text is small, tournament year is 2004 and answer type is people. All the parameters give a hint of the answer being True and it is True in the dataset, but the model shows as False.
      Thus, we can conclude that these parameters are not enough to make correct predictions as there is no clear threshold to separate correct and incorrect answers and therefore, some additional features needs to be added to improve the accuracy of the model.