Approach of solving the problem

1. Imported the dataset using the pandas library.

2. Checked whether the dataset is balanced or not. It was a balance both the labels had almost equal instances.

3. Checked in “boilerplate” it was having title and body. Separated both the title and body using json.loads and dictionary and created two columns title and body from them.

4. Cleaning the text-
   - ➔ Convert the text to lowercase.
   - ➔ Removed the special characters using the regex module.
   - ➔ Converted the text into tokens.
   - ➔ Removed the stopwords present in the text using nltk.
   - ➔ Performed spelling correction of misspelled words.
   - ➔ Performed stemming of words using PorterStemmer.

5. Combined the cleaned title and body column to form the boilerplate column again.

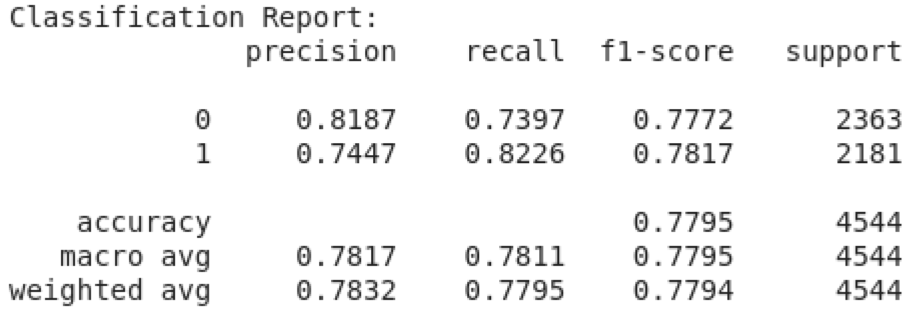
6. Saved the cleaned dataset to a csv file.

7. Used the pretrained GloVe embedding (**glove.6B.300d.txt**) to convert words into vectors.

8. Using Pytorch, trained the dataset using LSTMClassifier.

9. Created a custom loss function to improve the precision and recall while training the dataset.

10. With all these steps I was able to achieve the following results.



```
Classification Report:
              precision    recall  f1-score   support

           0     0.8187    0.7397    0.7772      2363
           1     0.7447    0.8226    0.7817      2181

    accuracy                         0.7795      4544
   macro avg     0.7817    0.7811    0.7795      4544
weighted avg     0.7832    0.7795    0.7794      4544
```

Nikhil Sharma
(nikhilsharma2296@gmail.com)