EXPLAINATION OF CODE

1)The first cell loads all necessary libraries and modules.

2)The seed is set so as to ensure the same sequence of randomization on every run of the code.

3)The train csv is then loaded. To ensure proper conversion, the delimiter is used and the apt columns are taken.

4) The text is then cleaned to drop the entries with Nan values and conversion to lowercase and tokenization is done.

5)The counts of each category are checked so as to see if any category has extremely high or low number of entries. This is done to see any steps to prevent overfitting for any class are necessary.

The plan is to use Naïve Bayes and according to the documentation available, NB is more prone to give the correct answer for a category that occurs frequently than for one which doesn't (pertaining to train dataset).

6) Lemmatizaion was tested . This takes the most time and possibly a warning regarding incompatible dtype may pop. This can be ignored without any harm to the execution.

Lemmatization was replaced by stemming and almost no loss in accuracy was found. The reduction in computation time was large. During stemming, care was taken to not end up exluding the alphanumeric numbers. This might be crucial for accuracy.

7) The train test split is used.

8)Tfidf transformation is done.

9) The model for Naïve Bayes was implemented.

How to test:

1) Replace the path to train dataset in cell 3.
2) The last 2 cells show how test dataset is dealth with. The code expects test dataset to have a header with the name "text".
3) The same functions as executed for train dataset are called upon for the test dataset as well.
4) The final cell gives accuracy.