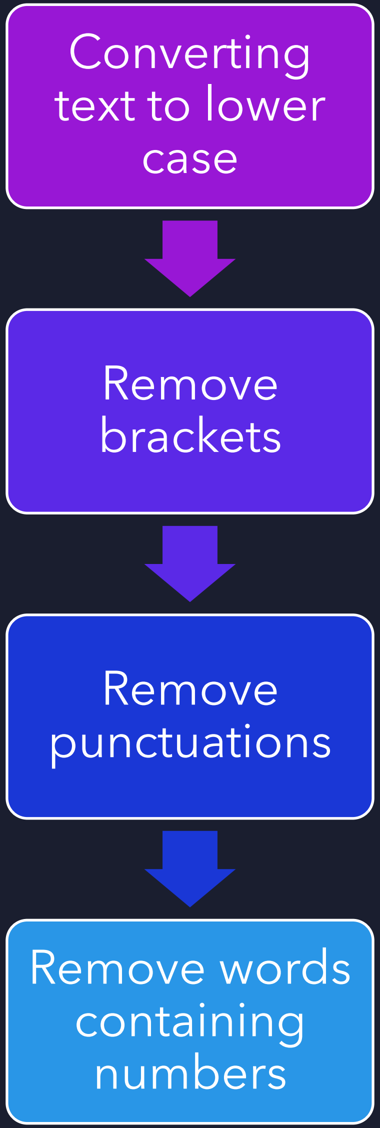
Yahoo Troll Question Detection

Team Karandeepak

- Karanjit (IMT202003)
- Netradeepak (IMT2020014)

Early Preprocessing steps (Round-1)



Converting text to lower case

Remove brackets

Remove punctuations

Remove words containing numbers

Extra Preprocessing Steps (Round-2)

Punctuation	• Get rid of additional punctuations
Lines	• Removing "\n" escape sequenes
URL	• Remove hyperlink URLs
HTML	• Remove HTML tags

Word Tokenization

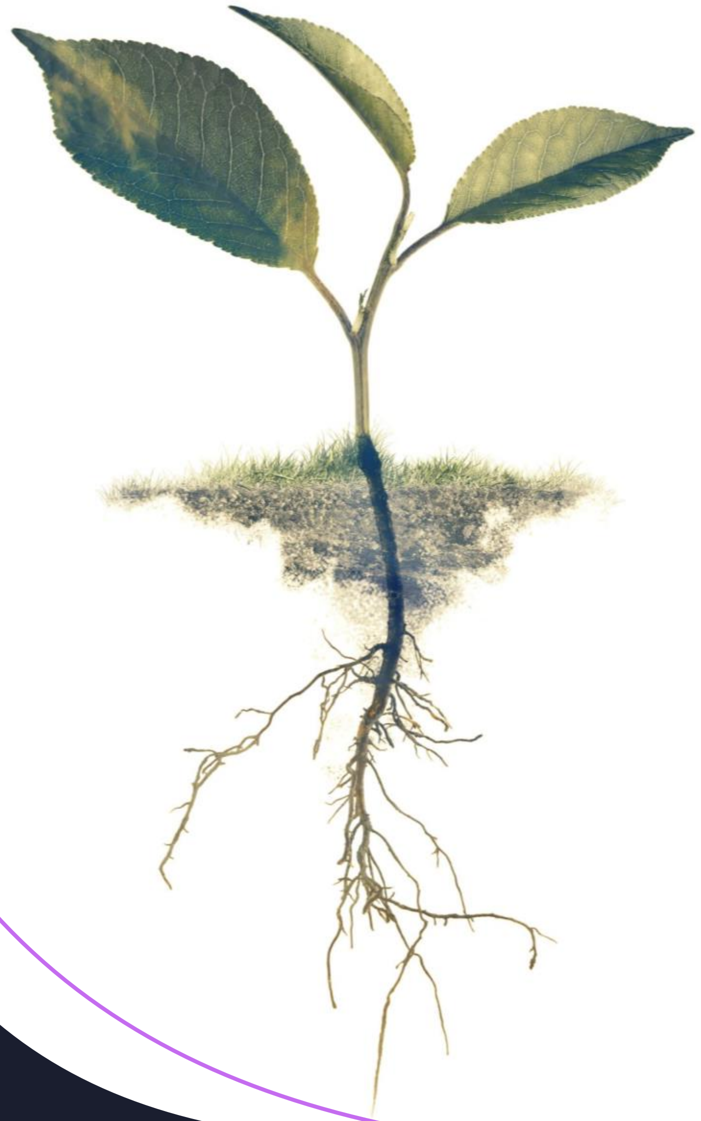
This technique breaks down a given sentence into a list of words which can be then processed upon independently making stuff like the upcoming lemmatization and stop word removal easy.

Stop Words Removal

Some words tend to amass a large percentage of the sentence but contain little to no important information regarding the question at hand. This process removes them

Lemmatization

This process converts words having the similar root meaning into the same root. For example, "go, went, gone" all will be converted to "go"



Vectorization

- As text strings are hard to use, the strings are converted to numbers. Here we are using the Count vectorizer.

- Here we used the n_gram parameter as well to consider phrases of lengths varying from 1 to 3.

Train –test split

We used train_test_split of sklearn library to split our dataset into training and testing dataset so that we take care to not overfit the dataset with any model.

Here we have split the dataset in the proportion train:test:: 0.7:0.3.

Here we also used the stratify parameter to make our training data less biased.

Models Used

1) Multinomial Naïve Bayes

We used multinomial Naïve Bayes model from sklearn library.

Metrics used:-

We have used f1 score as a metric for measuring the output of the model.

Our Observations:-

train f1 score: 0.8663474656987669

test f1 score: 0.4531286815443132

2) Logistic Regression

We used multinomial Naïve Bayes model from sklearn library.

Metrics used:-

We have used f1 score as a metric for measuring the output of the model.

Our Observations:-

train f1 score: 0.9501582278481012

test f1 score: 0.6029784943480379

3) XG Boost Classifier

We used multinomial XGBClassifier model from xgboost library.

Metrics used:-

We have used f1 score as a metric for measuring the output of the model.

Our Observations:-

train f1 score: 0.6551203410442477

test f1 score: 0.5341063045464571