**CLASSIFICATION MODEL**

**1. Data preparation**

Dividing the dataset into two directories based on their classes.

**2. Data cleaning**

To remove special characters (e.g. \n,\xa0,\t), punctuation and to extract useful features pre-processing of text data will be done using techniques like:

* lowering the text,
* using string library to remove punctuation,
* removing stop words,
* using 're' to remove any unwanted characters
* stemming, to get all the root words.

**3. Converting strings into vectors**

Before feeding the data into the model Count vectoring and Tf-Idf transforming will be applied to get the vector representation of strings.

**4. Model**

Algorithm planned to be used:

Naive Bayes(Multinomial)

Random Forest

Linear SVC

Deep learning approach (LSTM)

**Feature Extraction**

To extract the specified features, the data will be cleaned by lowering text, removing stop words and punctuation and using 're' to remove any unwanted characters.

The cleaned data will then be tagged using POS tagging and chunking will be used to get the desired features.