**CLASSIFICATION MODEL**

**1. Data preparation**

Divided 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 is done using techniques like:

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

All these techniques are wrapped into a single function which returns a list of root words for every given string.

**3. Converting strings into vectors**

Before feeding the data into the model Count vectoring and Tf-Idf transforming is applied to get the vector representation of strings. The vectors provide the information about how important the word is in the entire corpus.

**4. Model**

The data is split into two sets: Train and test with test size = 0.33% of actual dataset size

After this various algorithm are used for classification and their accuracy is compared.

Algorithm Accuracy Score

Naive Bayes(Multinomial) 0.8696

Random Forest 0.8393

Linear SVC 0.8878

Deep learning approach is also tested by using LSTM which gave an accuracy score of 0.81

**Feature Extraction**

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

Stemming is not done as it was changing the desired features to be extracted.

**APPROACH ONE**

POS tagging is used to tag all Nouns and digits and they are extracted. I was not able to separate the Employee and Organisation name using this approach.

**APPROACH TWO**

String is tokenized to get words which are tagged using POS tagging. After that Chunking is done. With this Employee name and Organization name is extracted.