**Naïve Bayes Classifier on Test Dataset**

# Problem

## Input:

Text data

## Output:

Category of Text Data

## Goal:

Goal is to train the data onto the Naïve Bayes Classifier and report results

# Language Used for Implementation

* Python

# Platform Used

* Jupyter Notebook

# Steps of Project:

1. Dataset Creation
2. Implementation
3. Documentation

## Dataset Creation:

Dataset is created by using the text from different fields, from internet. 30 text files are created manually, having following labels based on the text inside the text files:

* Sports
* Entertainment
* Business
* Technology
* Politics

Each category has six text files, so there were total 30 files.

## Implementation:

Implementation of project is done in following steps:

* Import all required libraries
* Dataset exploration
* Split data into train test
* Apply count vectorizer on text dataset
* Train data on Naïve Bayes classifier
* Test data on trained model
* Report accuracy

Import all required libraries

In 1st step all libraries are imported that are required in python language to implement this project. Some of the libraries are about data exploration and some of the about the classifiers

Dataset Exploration

In this step all the text files are read and convert them into dataframe to further explore the data. Files of each category are explored individually and then convert all these into one dataset.

Train and Test Dataset

In this step, all dataset is split into two categories, train and test. We used train test split approach. 33 % is test dataset and other is train dataset. Train dataset is used for training the data on naïve bayes classifier while test dataset is used for evaluation of this trained model

Count Vectorizer

By using count vectorizer technique (Kulkarni & Shivananda, 2019), text data has been converted into the numeric to proceed the training

Training Phase

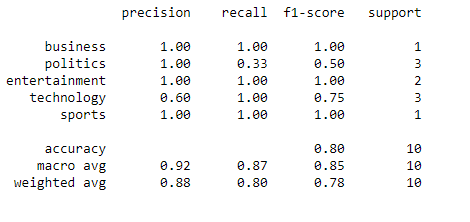
Training of train data is done by using the naïve bayes classifier (Rish, 2001)

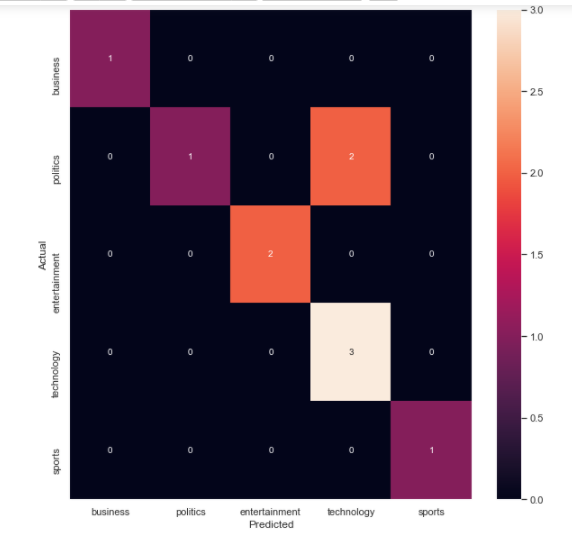
Testing Phase

Testing of trained model is done on the testing dataset

Evaluation

We achieved the accuracy of 80% by using this model. Figures below shows the classification report (Deng & Qi, 2016) and confusion matrix of test dataset





# References

Deng, X. & Qi, L., 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences.*

Kulkarni, A. & Shivananda, A., 2019. Converting text to features. In: *Natural Language Processing Recipes.* s.l.:s.n.

Rish, I., 2001. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence.* s.l.:s.n.