Twitter user gender classification

Objective:

Predict user gender based on Twitter Profile information.

Data Source:

The Data has been extracted from Kaggle. The dataset consists of 20050 rows and 26 columns. Among 26 columns there are 25 predictor variables and 1 target variable which is gender in this case.

The link to the data source is given below.

Link- DataSet

Methodology:

Step 1- Install Classifiers and import all the important libraries.
Step 2- Read Data from CSV file.
Step 3- Show the shape of Data, and check the Information of it.
Step 4- Dropped redundant columns from the DataSet.
Step 5- We check for null values in the DataSet, if there are any, we drop them.

Step 6- Count the variables of the ‘gender’ column.

Step 7- Encode the ‘male’ as 1 and ‘female’ as 0 using the replace() function.

Step 8- text cleaning.

Step 9- Tokenization and Lemmatization of Cleaned column and remove stop words(English) using NLTK library.

Step 10- Add new column to DataSet of “DescriptionList”.

Step 11- Bag of Words (we will take top 5000 feature)

Step 12- (make an array from Counter vector of the "DescriptionList")

Step 13- Split the Data into train and test Data.

Step 14- Applied classifiers (Decision tree,

Random Forest,

LogisticRegression)

to get the ACCURACY.

Step 15- Show the Highest ACCURACY.

Step 16- View the classification report for test data and predictions.

Step 17- Show The Result.