NAME : K SUNIL JOSHI

SOCIAL COMPUTING

HOMEWORK -2

“GENDER CALSSIFIER USING TWITTER DATA BY PERFORMING MACHINE LEARNING MODELS AND PREDICTING GENDER”

Problem Statement : Our aim is to predict the gender of a twitter user based on input features in the dataset.

Dataset :

<https://www.kaggle.com/crowdflower/twitter-user-genderclassification>

The dataset contains Twitter user information like Gender, Text , Tweet\_Id , location etc., which has 20050 rows and 26 columns.

1 ) Feature Engineering :

* Dataset is reduced only to few columns as most of the columns has redundant information
* Columns with text , gender and description are mainly used for model building.
* Features like Tokenizing, stopwords, text length , TF-IDF are used to analyze text and their frequency in order to see how male/female gender uses more repeated words and who used more tweet length which is one of the key factor to predict.
* Another features like sidebar\_color, link\_color is also used to test if we can predict the gender using these as input features to trin the model.
* The data is only considered for gender:confidence ratio of 100% which is the best data compared to less than 100%.
* Using the regular expressions the data is also cleaned from Special characteristics ,html tags, punctuations which may result in poor performance of the model, so they are removed from the text.
* Finally, the data is processed using above mentioned techniques and is fit into model to classify.

2) Classifier:

* The data is further split into test and train set , and modelled using Logistic regression , SVM ,Random Forest and Ensemble voting classifier as the task is to classify , so which falls under machine learning classification technique.
* From the output it is seen that Logistic Regression gives the best accuracy and SVM with linear kernel also gives the best accuracy.
* For this dataset we have to predict binary output so it’s a binary classification task, so these models are used in order to predict the gender.

3)Model Evaluation:

* I used model\_selection, train\_test\_split inbuilt Sklearn function to split data into train and test which splits data randomly based on (test\_size = 0.3) , also used metrics like accuracy\_score , classification\_report and confusion\_matrix to evaluate the model performance.
* By training and testing the data using the above models we finally came up with a classification task with highest accuracy and prediction and performance metrics like Accuracy , Precision and Recall.

4)Implementation:

* Data is preprocessed using Stopwords, tokenizing using NLTK inbuilt modules, removing special characters, html tags, punctuations and extra spaces using regular expressions (python re module).
* The data is extracted using NLTK inbuilt functions like PorterStemmer , TF-IDFstopwords , Tokenizer and corpus.
* The classifiers used are python’s Sklearn LogisticRegression , SVC and RandomForest modules in which the training data(x\_train , y\_train) is fed to build the model and to predict I used test data (x\_test, y\_test) in which x is text column and y is gender column from original dataset.
* The data set is partitioned into train and test using Sklearn model\_selection.train\_test\_split function , in each iteration the function randomly partitions into train and test based on 0.3 test\_size parameter .
* The metrics of the models are calculated using accuracy\_score, confusion\_matrix, classifier\_report inbuilt sklern.metrics module, in all of these the input arguments passed are y\_test and x\_test as the prediction is done on test data.

5) Results:

* The model accuracy and classifier report are calculated for each model and among all the Logistic Regression gives the highest accuracy , with 60.1% compared to 57.9% for other models.
* By using python seaborn module I plotted heatmap for confusion\_matrix from which it is evident how many of them are correct (true) predictions and which are incorrect (False) predictions.

