**Assignment No: 4 Matriculation Number: 00387718, 00413018**

**Psychological Disorder Profiling through Social Media Mining**

**<Implementing through Scikit-Learn>**

**SUMMARY:**

We have focussed on Data cleaning and identifying features along with label mapping to the data. Also, we have implemented machine learning algorithms such as SVM, Random Forest & Logistic Regression using scikit-learn library

**Data Processing:**

The data set is further cleaned by removing the below impurities, missing and incorrect values.

* User Names from tweets
* English “Stop Words”
* Keyword “RT” indicating Retweets
* Hashtags such as #Depression, #OCD & #Anxiety
* Noise exclusion (Tweets having hyperlinks to other twitter pages instead of text)

Further the cleaned data was categorized as per disorder types into panda data frames format.

**Features & Labels:**

Features provide unique representation for sample data values. The most common used feature for text mining is “Word Frequency”. The same was implemented using TfidfVectorizer() method from scikit-learn library. Factors such as max\_df, min\_df, lowercase text have been taken into consideration while calculating TF-IDF frequency measures. The values of max\_df & min\_df were set to 0.8 & 3 respectively i.e. words that have occurred maximum in 80% of the tweets & the words that have occurred in minimum 3 tweets. The lower-case tweets are also taken into consideration.

Labels are used to assign samples to different classes. Afinn wordlist is used to score the sentiment of the sentences based on positivity, negativity & neutrality. LabelEncoder() method is implemented using scikit-learn library. So this method converts categorical Afinn score into numerical feature.

**Implementing Algorithms:**

The following Algorithms were implemented using scikit-learn showing their predictive accuracy

* SVM
* Logistic Regression
* Random Forest Classifier

**Cross-Data Validation:**

k-fold cross validation methodology is used to split the Data set into Training & Test data sets.

In this scenario “k” is selected as 5 where in the entire data set is split into 5 equal parts. Each “k-1” parts act as training set and the remaining part acts as test set. Each iteration ensures that each “k” part becomes test set at least once. Thus, all the data can be a test & training set in iterations. This helps to provide a better predictive accuracy. The predictive accuracy for the chosen algorithms has been listed below.

|  |  |
| --- | --- |
| Algorithms | Predictive Accuracy |
| SVM | 0.848 |
| Logistic Regression | 0.853 |
| Random Forest Classifier | 0.868 |

**Conclusion:**

It was found that the Random Forest classifier has the best accuracy for predicting outcomes when applied on 5-fold cross validation processed data set.