Classification of tweets.

Overview:

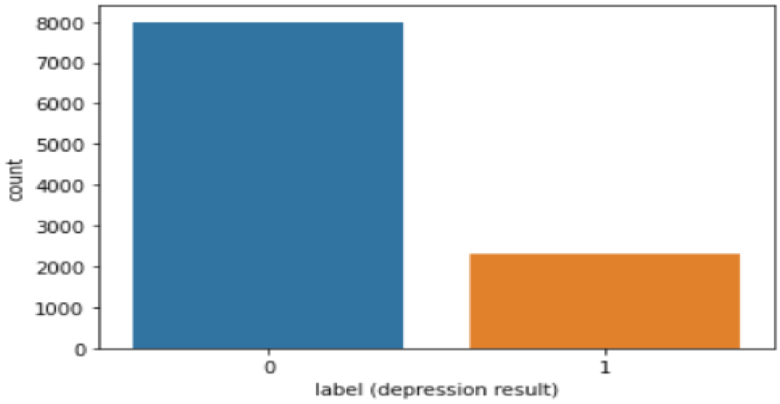
The dataset contains 10314 samples 3 features containing the text of the tweets and the sentiment of tweets whether they are depressed or not.

First import required libraries for pre-processing and then load the dataset using Pandas.

Data Analysing:

The data contains a feature of text (tweets) and a feature named label has sentiment values of depressed or not (1-depressed,0-not depressed).

The distribution based on value of depression is plotted as:



Text Pre-processing:

1. Removed the URLs, mentions from the tweets.

2. Tokenization, Stemming, Stop-Word Removal, Lemmatizing is performed to obtain the processed text data.

3. Applied TfidfVectorizer to the data to convert text data into vectors.

3. Splitting the data – Using train_test_split we split 20% of data as test data and the remaining as train data.

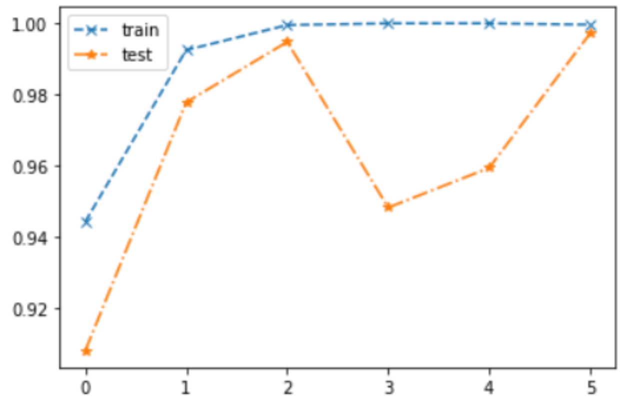
Modelling:

By using different models train the dataset to get better accuracy. These include:

Multinomial Naïve Bayes, Logistic regression, SVC, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier.

Now train the data using all the above models.

The accuracies obtained by above models for train and test data is plotted as follows:



Here,

0-Multinomial Naïve bayes,1-Logisticregression,2-SVC,3-Decision tree,4-Random forest, 5-Gradient boosting classifier.

Therefore the better model we obtained is Gradient Boosting Classifier.

Final-model:

From above all the models we select final model for the data.

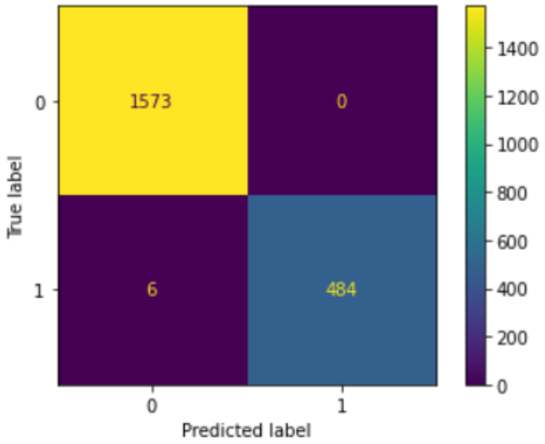
The metric values are:

Accuracy train: 0.999

Accuracy test: 0.997

f1 score: 0.997

Confusion matrix as below:


True label
Predicted label