Detection of Employee Stress Using Machine Learning

In this paper author is describing concept to detect employee stress by using machine learning algorithms such as SVM and Random Forest Algorithms. To detect stress author is using social media dataset such as tweets where employee can share their views and by analysing this views we can identify whether employee is in relax or stress mood but by analysing this views manually may take lot of human efforts so author using machine learning algorithms and the experiment with this algorithms show stress detection accuracy more than 90%.

To implement this project we are using twitter tweets dataset and this dataset is pre-process with NLTK (natural language tool kit) to remove stop words and special characters. Processed dataset will be splitted to train and test part where application used 80% dataset for machine learning model training and 20% dataset is used to evaluate or test machine learning prediction accuracy.

Random Forest: This algorithms first select random values and then start looking for class which is close to that random values and assign that class to that values and this process continues till all classes assign with closer random values.

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

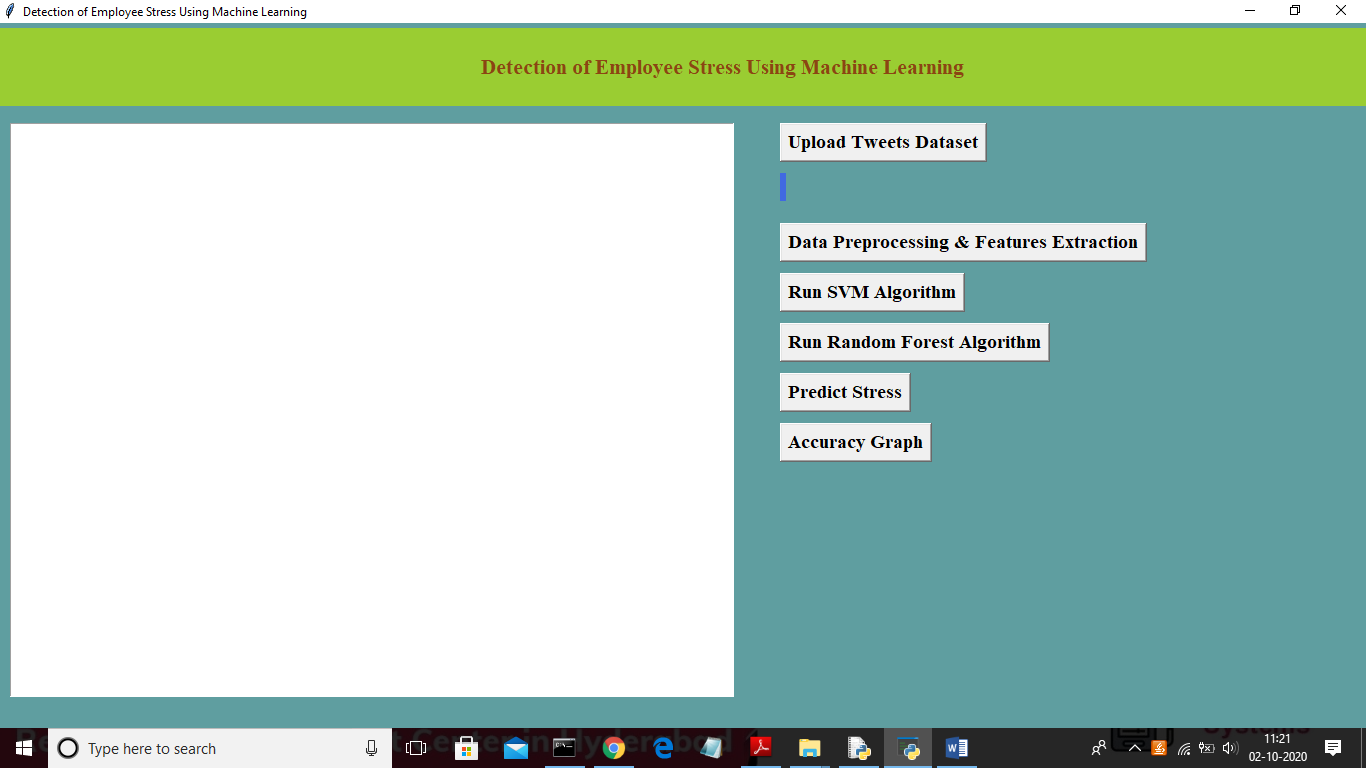
How do we find the right hyperplane?

Or, in other words, how do we best segregate the two classes within the data?

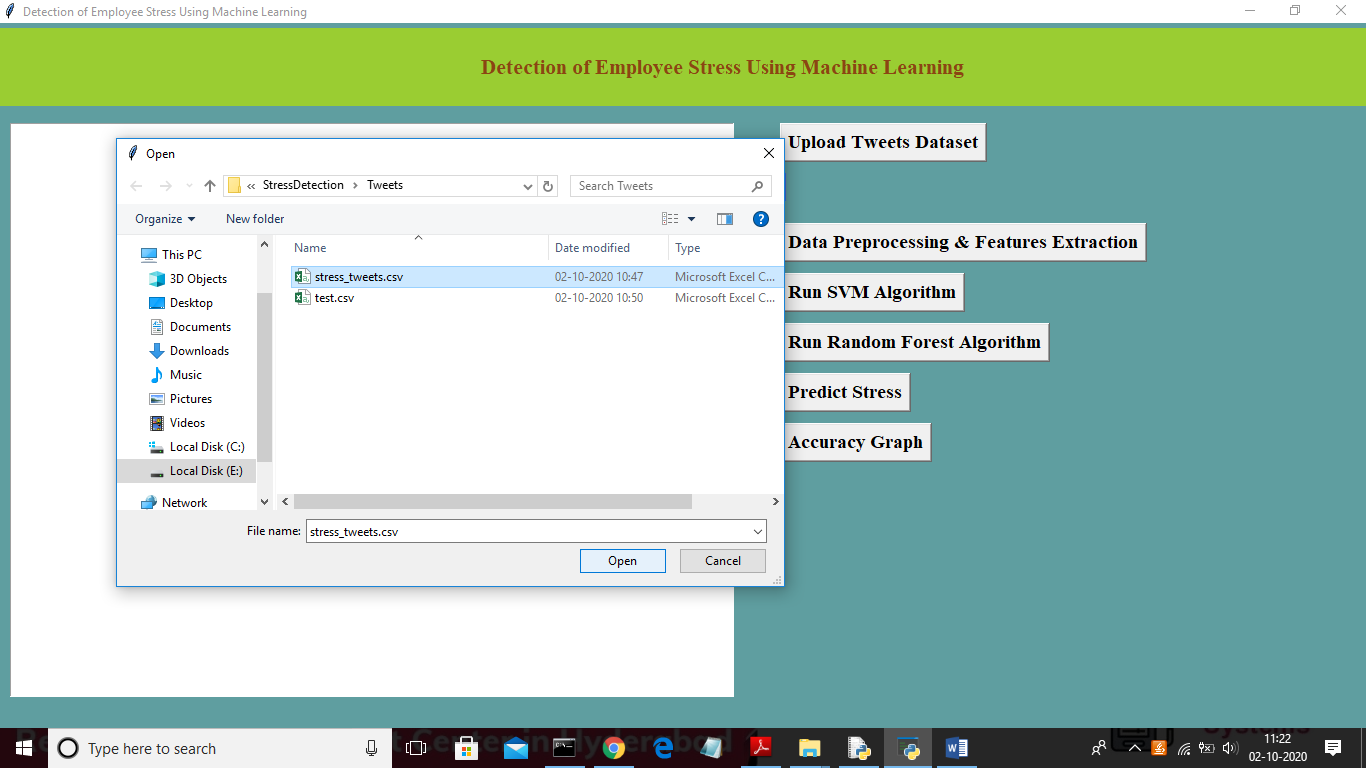
The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

SCREENSHOTS

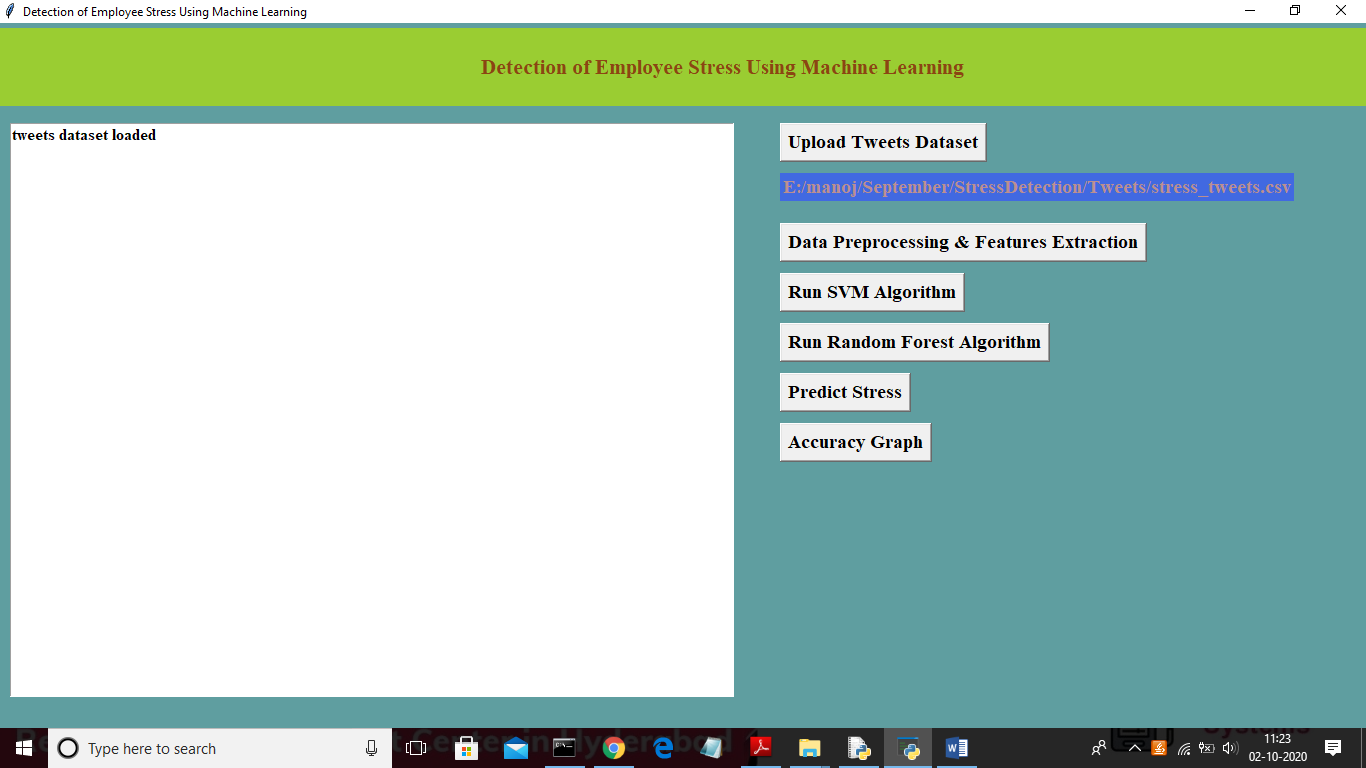
To run project double click on ‘run.bat’ file to get below screen



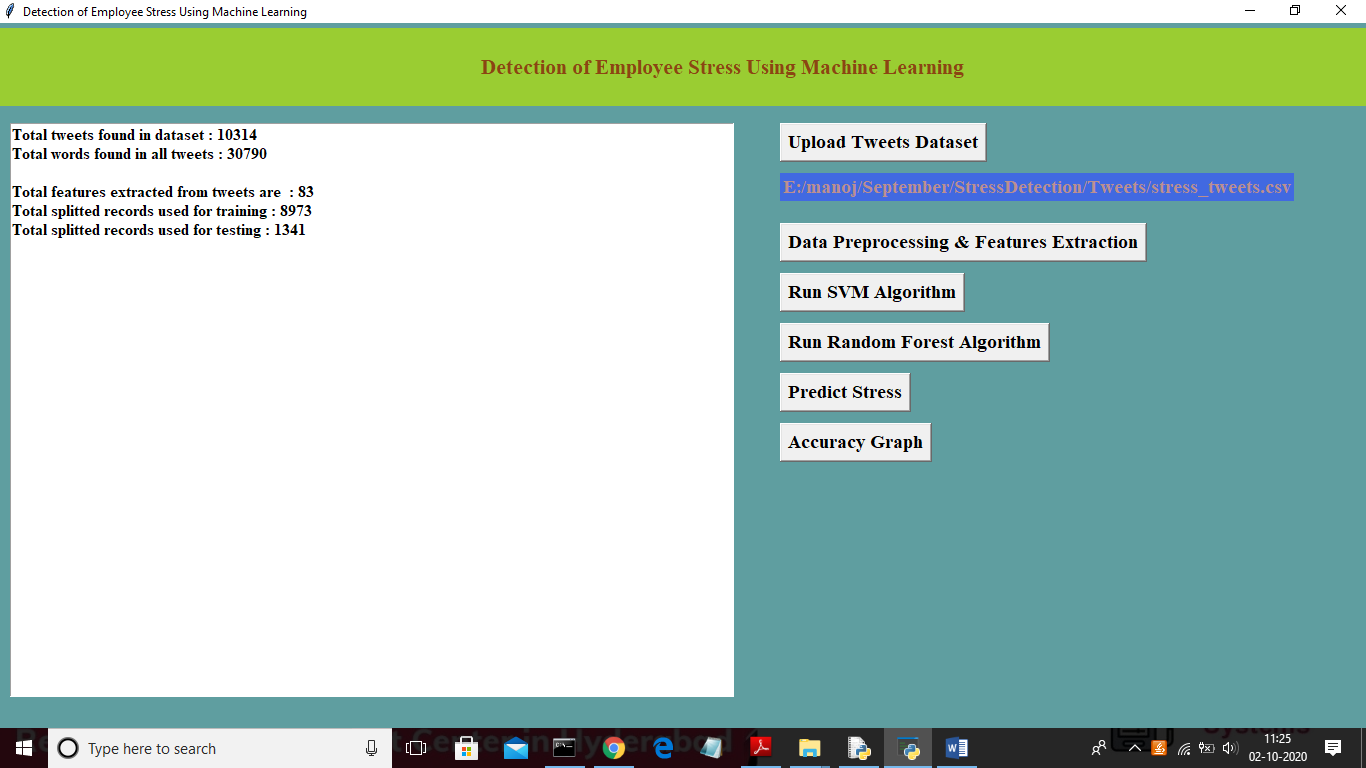
In above screen click on ‘Upload Tweets Dataset’ button to load dataset



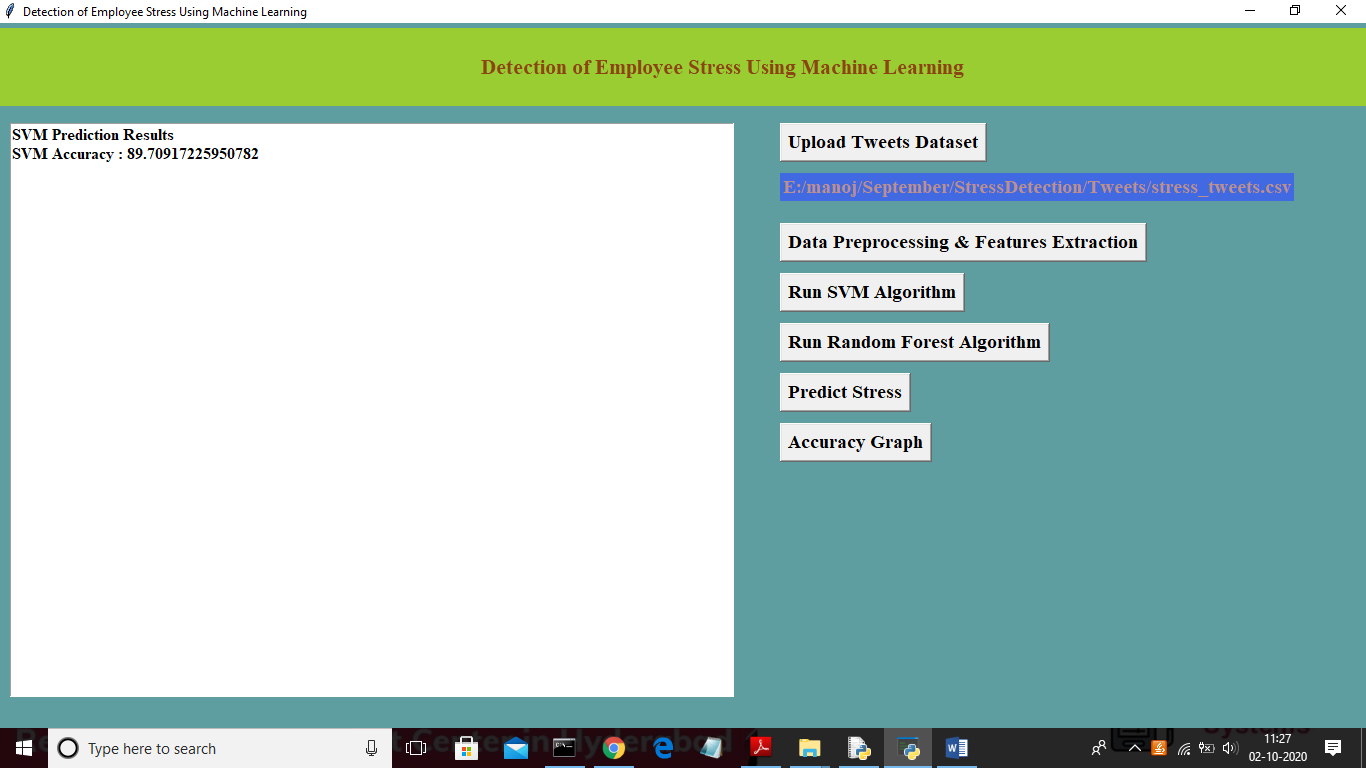
In above screen select ‘stress\_tweets.csv’ dataset and then click on ‘Open’ button to load dataset and to get below screen



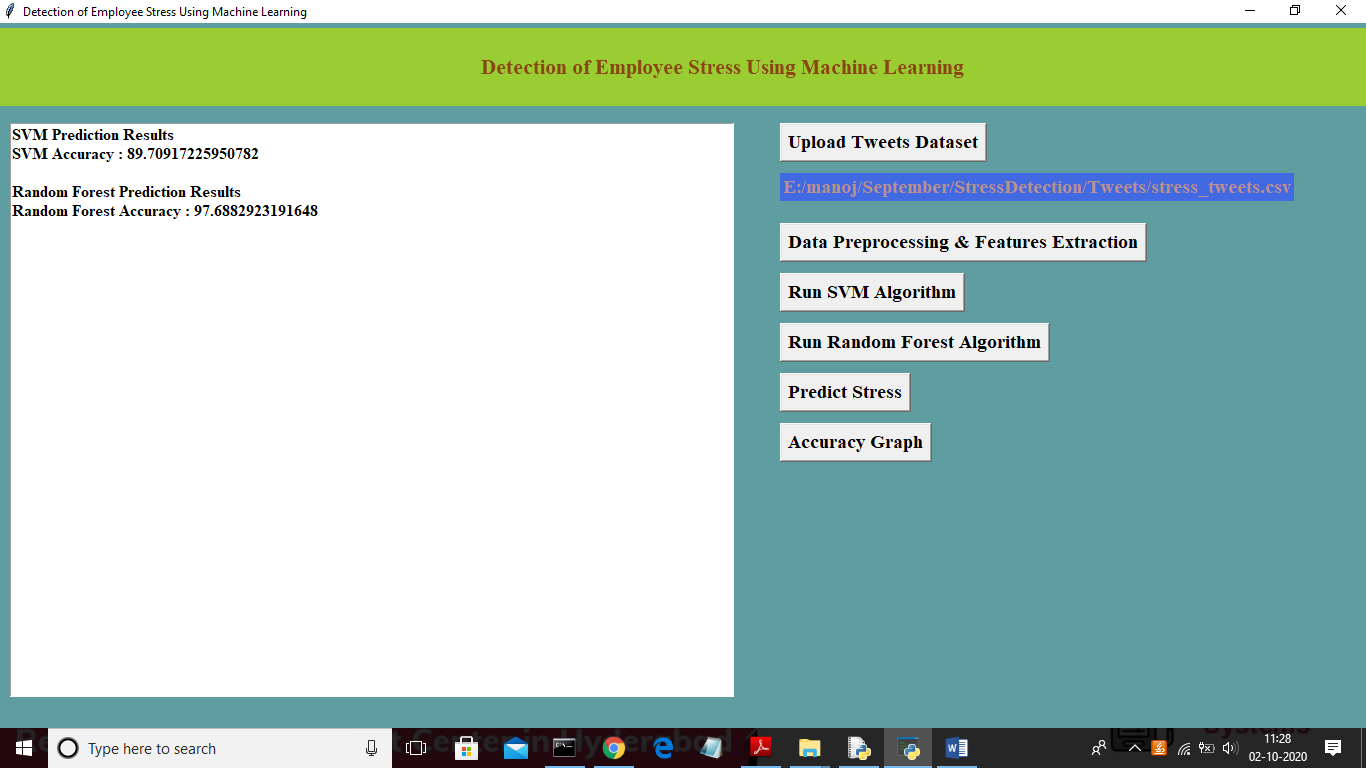
In above screen click on ‘Data Preprocesing & Features Extraction’ button to read dataset and to clean and extract features such as words from dataset and find total records in dataset, total words and application using how many records fro training and testing



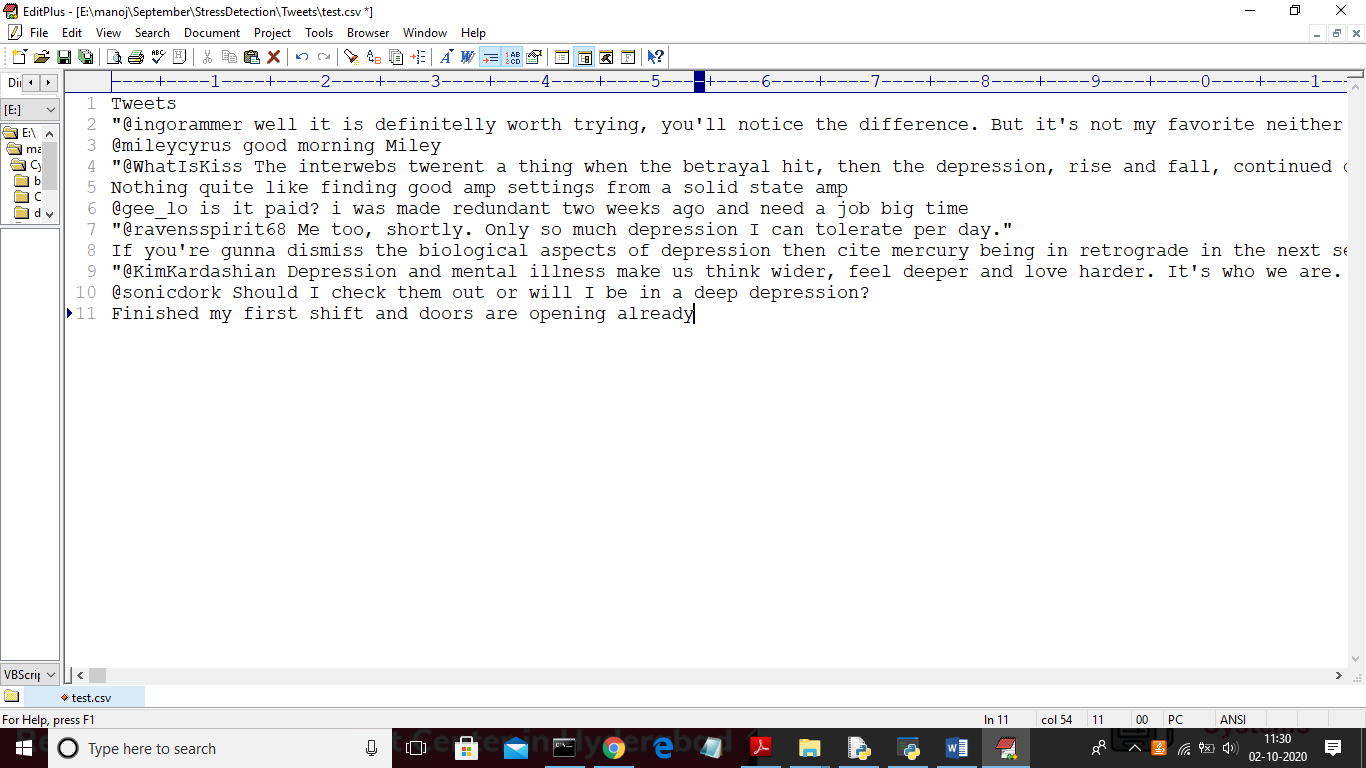
In above screen dataset contains total 10314 tweets and all tweets contains 30790 words and total unique words are 83 and application using 8973 records for training and 1341 for testing. Now both train and test data is ready and now click on ‘Run SVM Algorithm’ button to trained data using SVM machine learning algorithm.



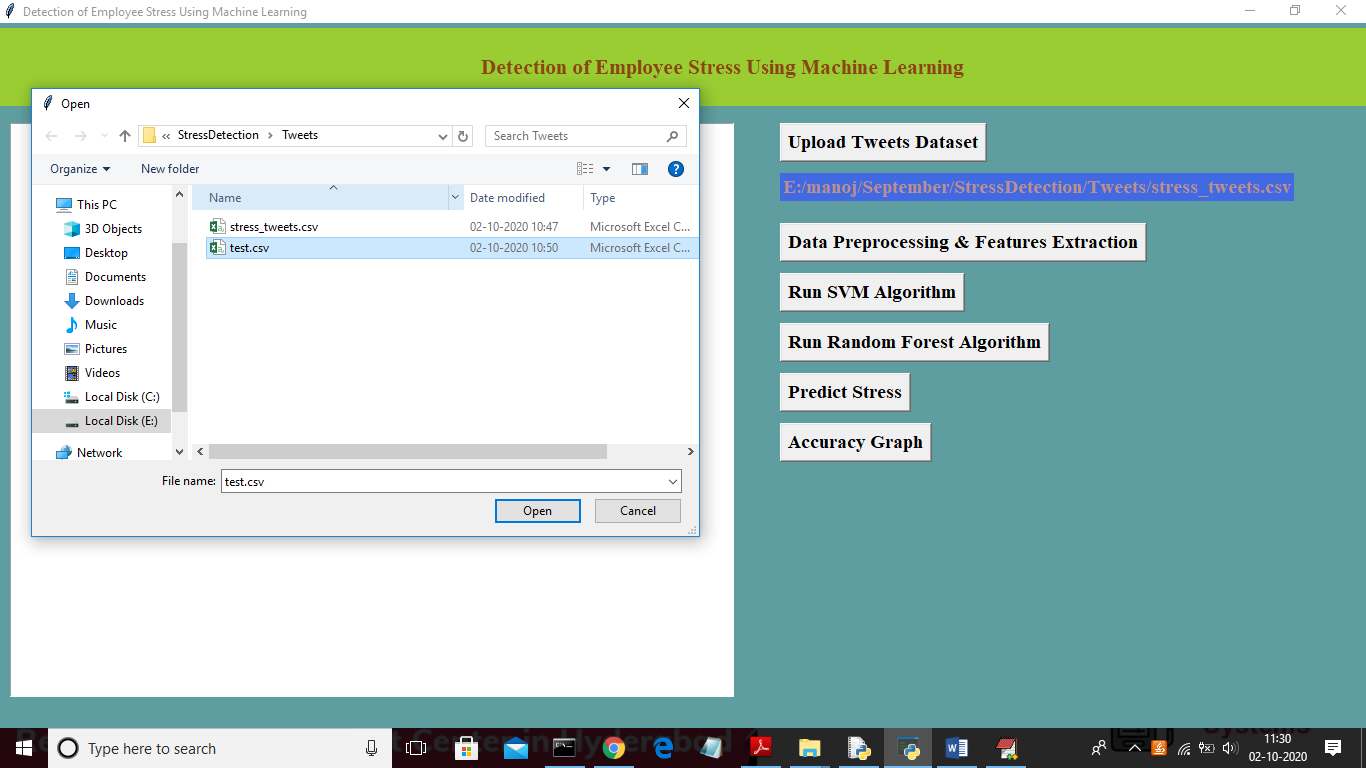
In above screen SVM got 89.70 correctly predicted accuracy from test data and now click on ‘Run Random Forest Algorithm’ button to calculate its accuracy



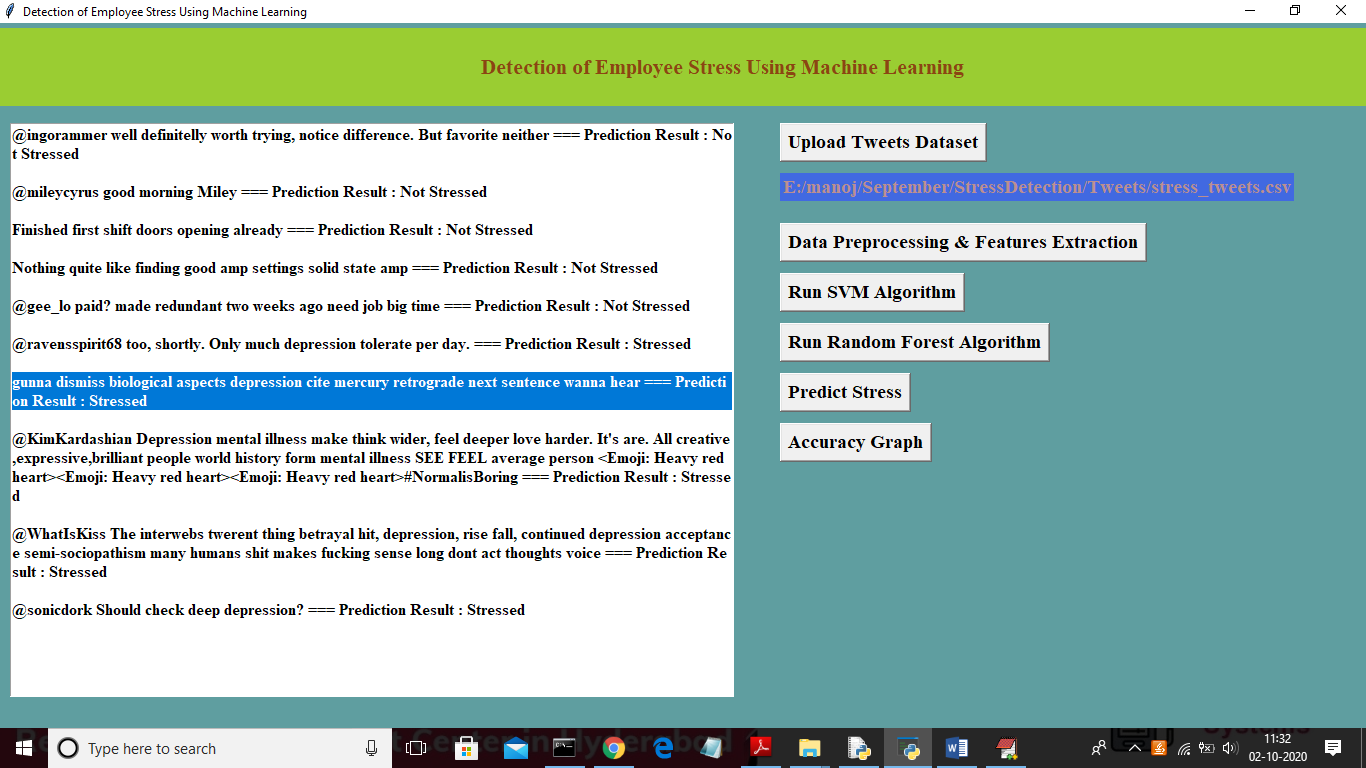
In above screen random forest got 97.68 correctly prediction accuracy and now click on ‘Predict Stress’ button and upload test file which contains tweets and by analysing those tweets machine learning algorithm will predict whether tweets contains any stress data or not. Below is the screen shots of test tweets which we upload in next screen



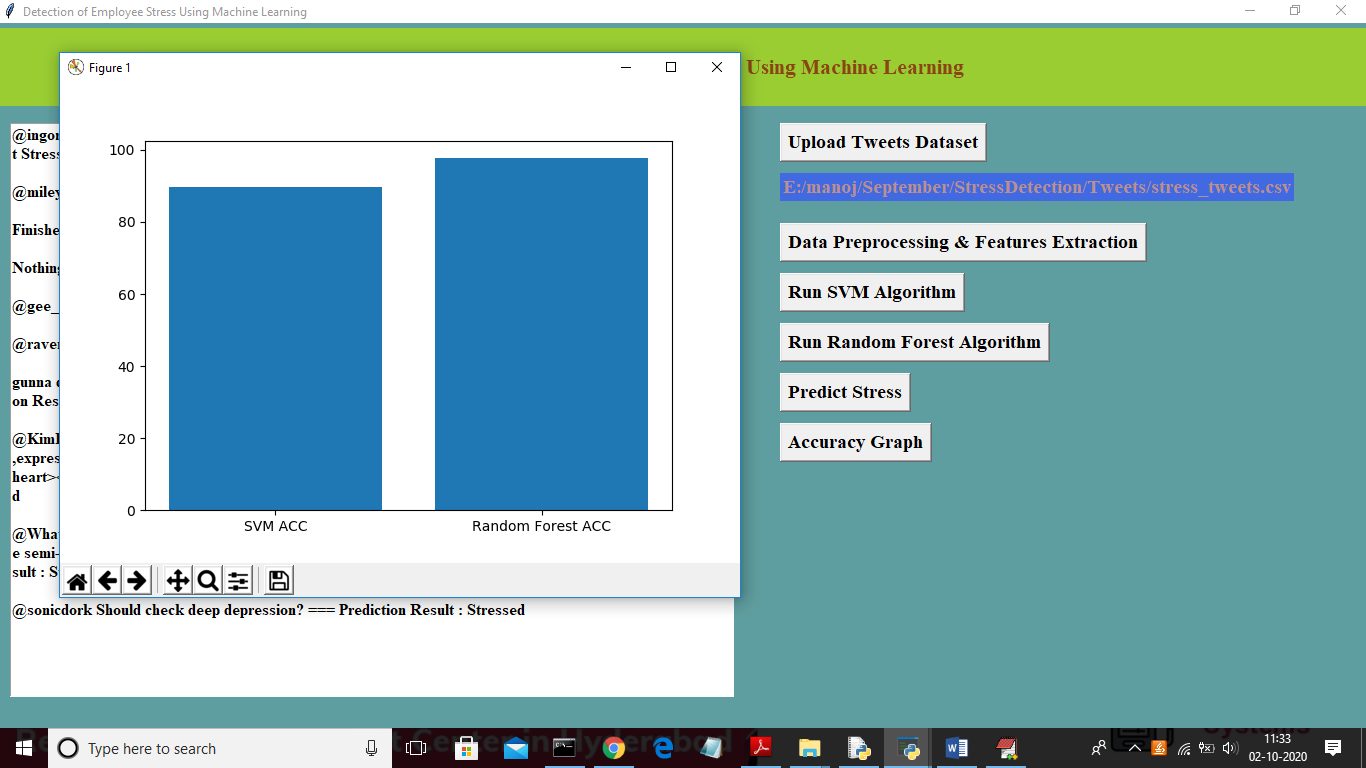
In above screen we can see some tweets are there in test .csv file and now we upload this tweets to predict stress by clicking on ‘Predict Stress’ button



In above screen uploading ‘test.csv’ file and now click on ‘Open’ button to predict stress



In above screen beside each tweet we can see predicted result as Stressed or Not stressed. From above screen we can see application detecting stress successfully from messages and now click on ‘Accuracy Graph’ button to get below comparison graph



In above x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can say random forest is better than SVM