Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms

In this paper author is using Bio-Inspired features selection algorithms such as PSO and Genetic Algorithm to extract important features from SPAM EMAIL dataset and this important features will be input to various machine learning algorithms such as Random Forest, Decision Tree, Naïve Bayes, SVM and MLP to train SPAM DETECTION MODEL.

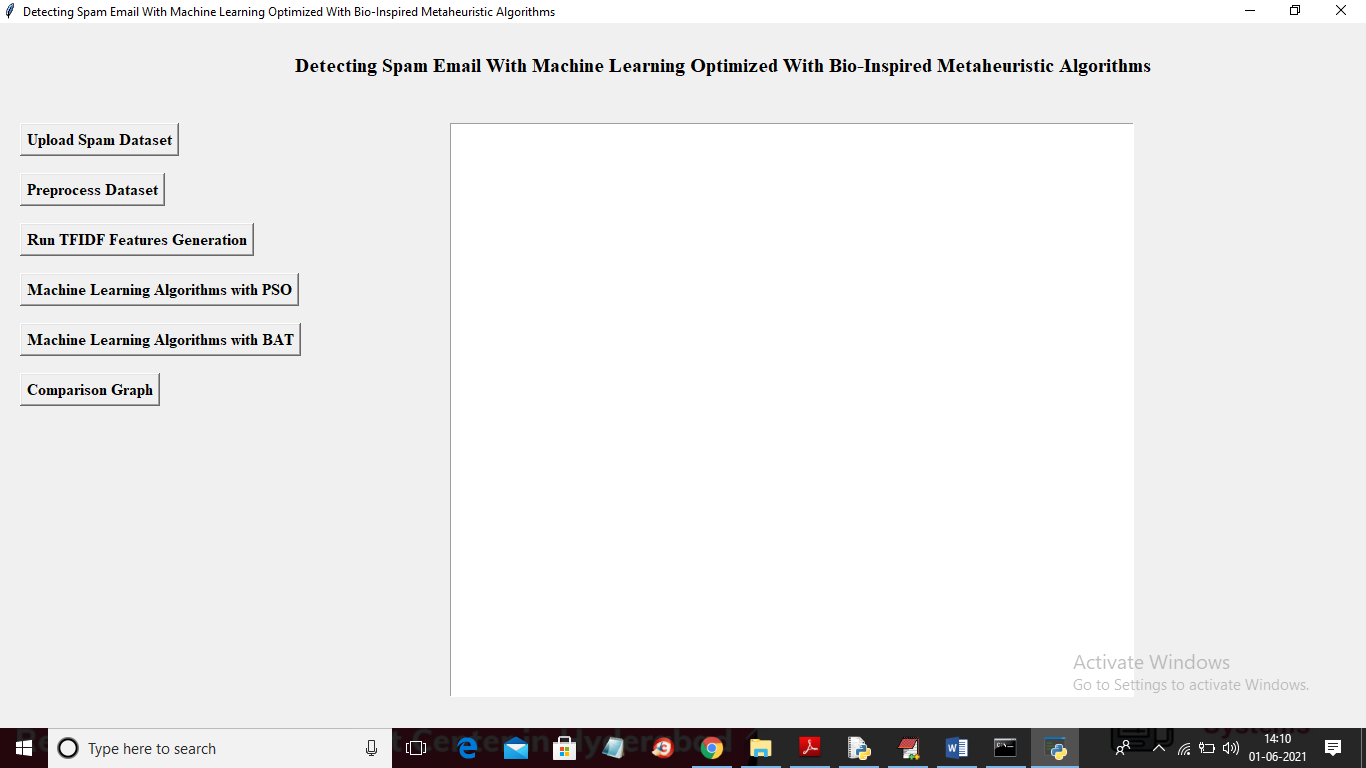
All dataset consists of features/column values and in all dataset all features are not important so we can optimize those features by apply various features selection algorithms such as PSO, PCA, BAT and many more.

In propose work we are using PSO and BAT as features selection algorithms and to implement this project we have designed following modules

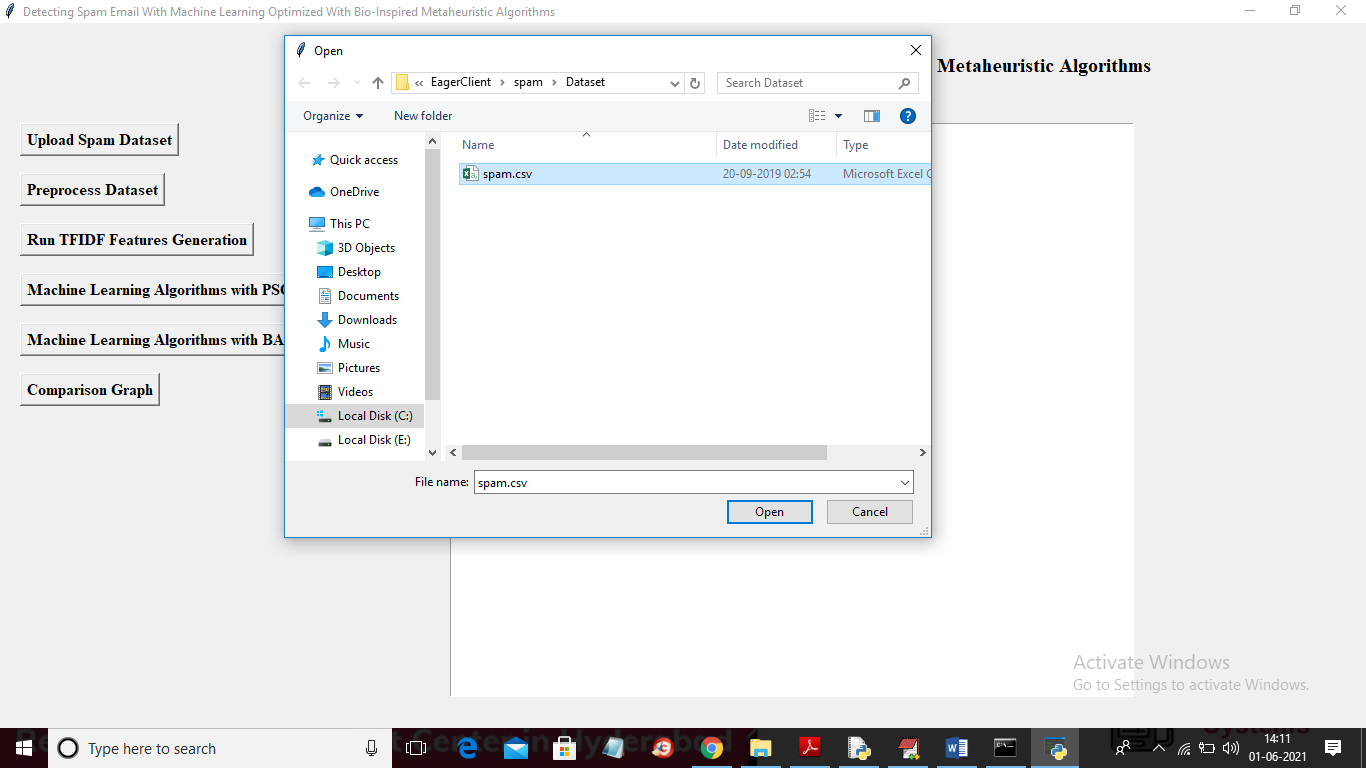
1. Upload Dataset: using this module we will upload dataset to application
2. Preprocess Dataset: using this module we will clean all dataset messages by removing stop words and special symbols
3. TFIDF feature engineering: using this module we will convert all words into vector where all words will be arrange as vector column names and the count of each word will be arrange as vector rows
4. Machine Learning Algorithms with PSO: using this module we will train all 5 above describe machine learning algorithms by using optimised features from PSO
5. Machine Learning Algorithms with PSO: using this module we will train all 5 above describe machine learning algorithms by using optimised features from BAT
6. Comparison Graph: will plot accuracy, precision, recall and FSCORE graph for all algorithms with PSO and BAT.

SCREEN SHOTS

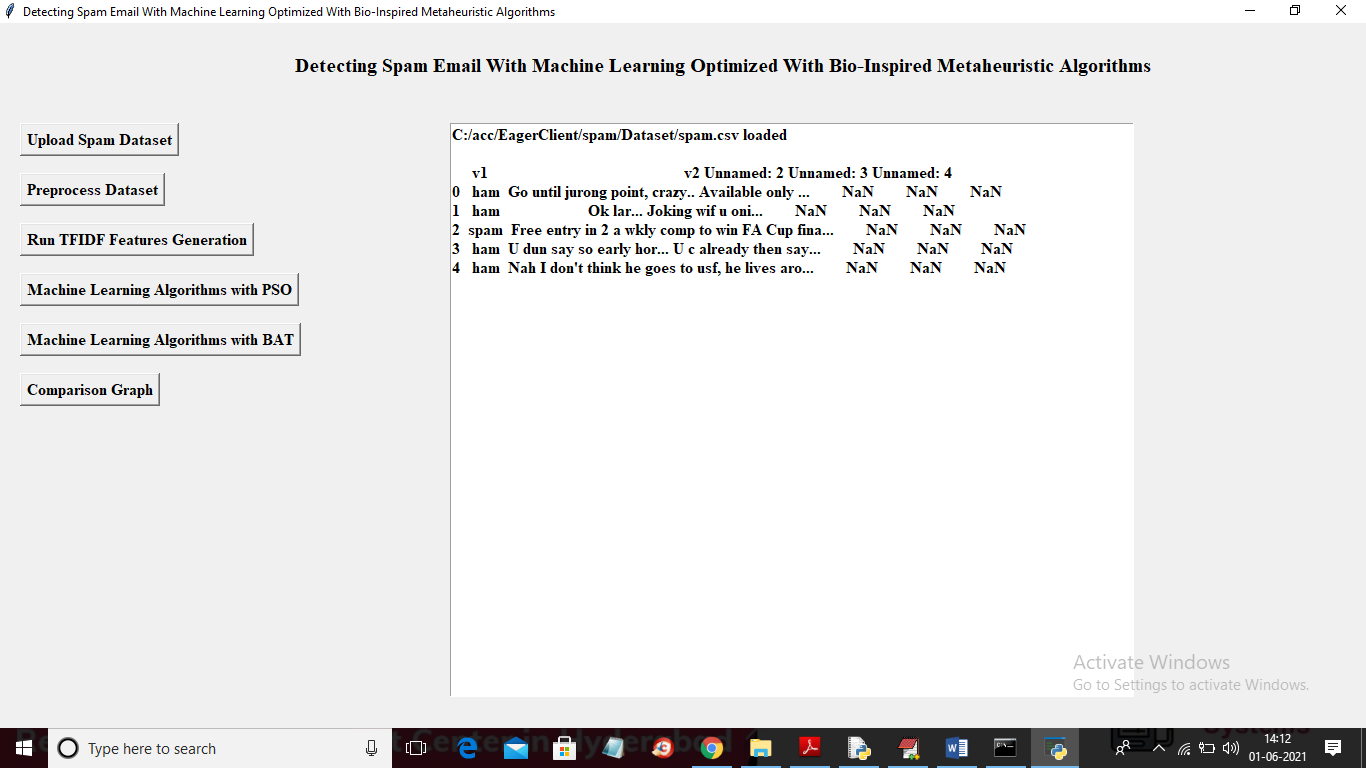
To run project double click on ‘run.bat’ file to get below screen



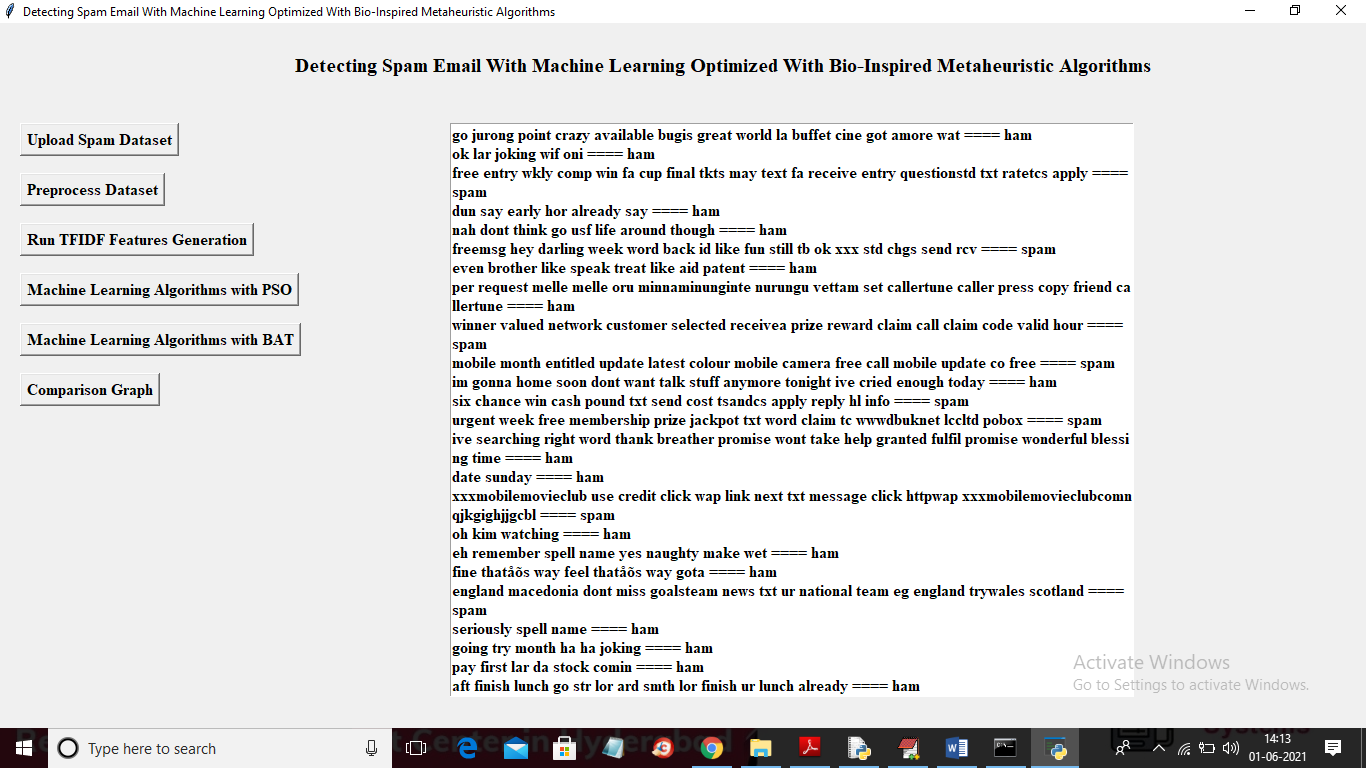
In above screen click on ‘Upload Spam Dataset’ button to upload dataset



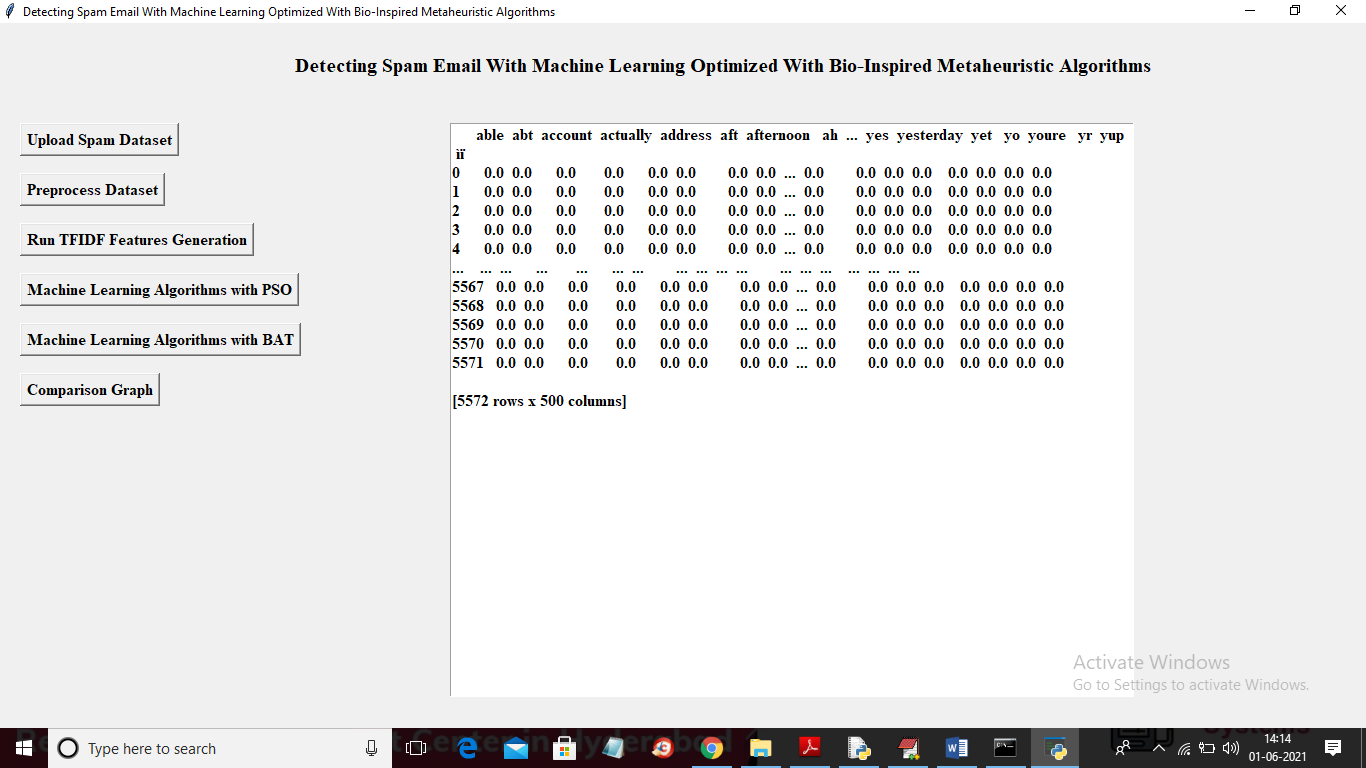
In above screen selecting and uploading ‘spam.csv’ file and then click on ‘Open’ button to load dataset and to get below screen



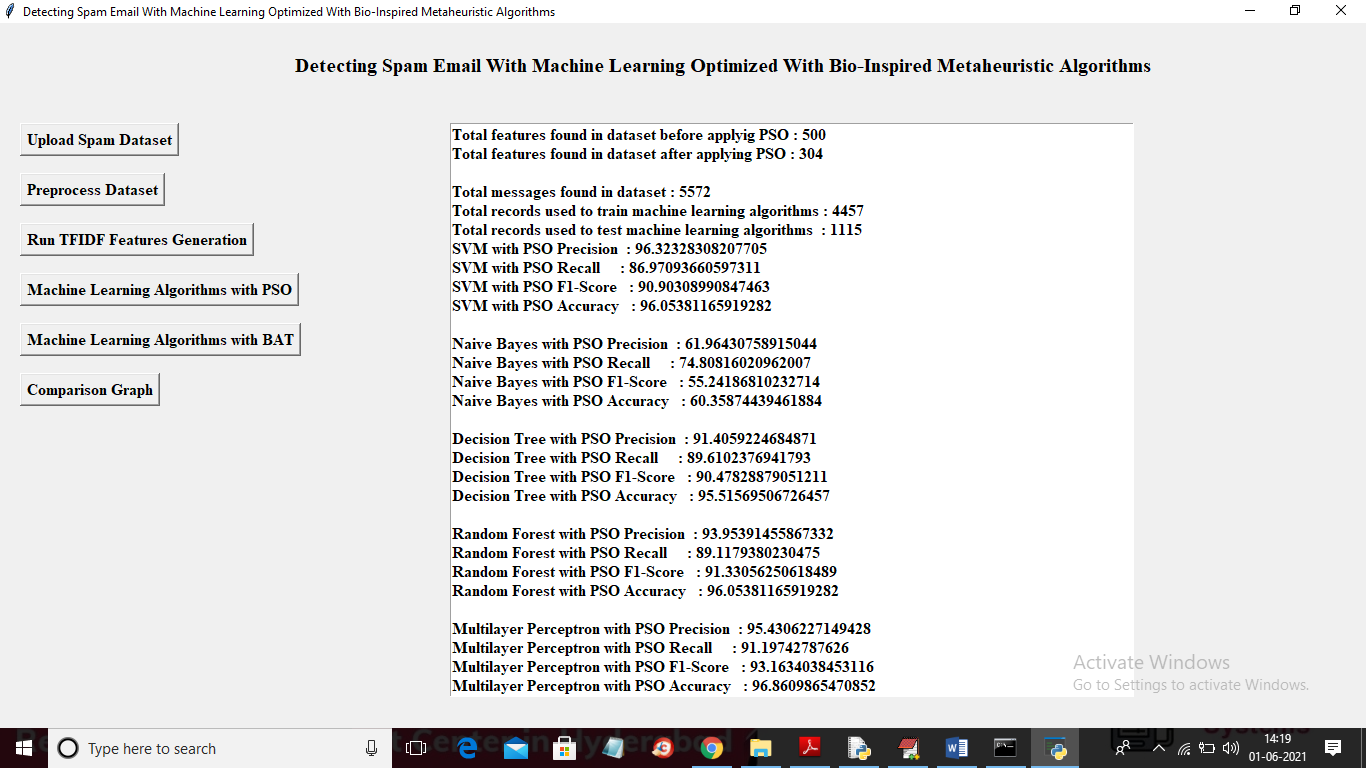
In above screen dataset loaded and displaying few values from dataset and this dataset has two columns where first column contains message label as ‘spam’ or ‘ham’ and second column contains email message. Now click on ‘Preprocess Dataset’ button to clean dataset



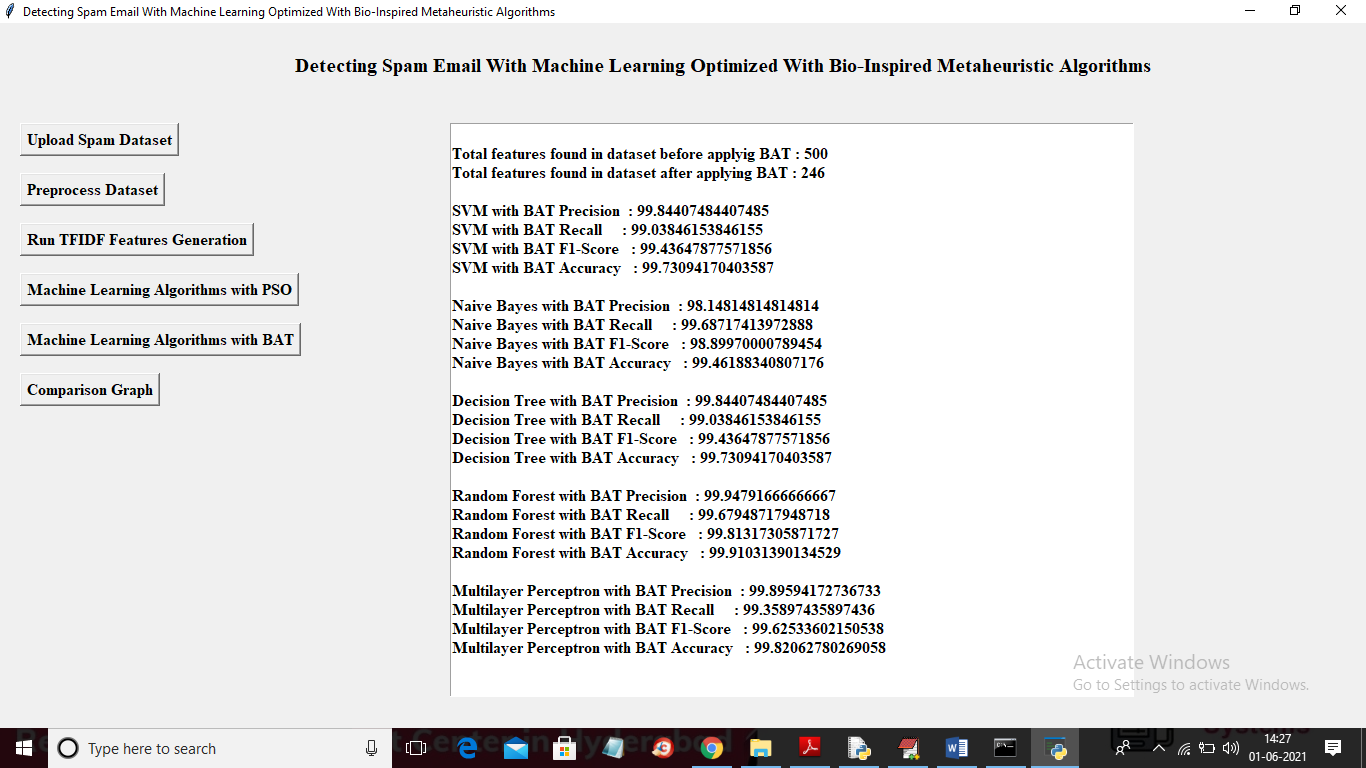
In above screen from each message special symbols and stop words are removed and after equal symbols we can see message label. Now click on ‘Run TFIDF Feature Generation’ button to convert messages into TF-IDF count vector.



In above screen we can see first column contains words from messages and remaining rows contains count of that word. If word available then that row will have count of word else 0 will be put. I am displaying few vector values in above screen. Now click on ‘Machine Learning Algorithms with PSO’ button to train ML with PSO optimize features



In above screen in first line we can see dataset contains 500 features and after applying PSO features has been reduced or optimized to 304 and then train each algorithm with optimize features and almost all algorithms got more than 95% classification accuracy on test data. In above screen we can see dataset contains more than 5500 records and for training application using 4457 and for testing classifier accuracy application using 115 records. Instead of confusion matrix I calculated precision, recall and FSCORE. Now click on ‘Machine Learning Algorithms with BAT’ button to train ML with BAT optimized features



In above screen we can see dataset contains total 500 features and BAT algorithm selected 246 important attributes and then train all ML with 246 optimized features and then for each ML we got 99% classification accuracy. Now click on ‘Comparison Graph’ button to get below screen



In above graph x-axis represents algorithm names with PSO and BAT and y-axis represents accuracy, precision, recall and FSCORE. All algorithms gave better accuracy with BAT features