Minor Project Report On

Personality Detection from Text

Submitted in partial fulfilment of the requirements for the award of degree

of

Masters of computer application (MCA(SE))

Guide:	**Submitted By:**	
Dr. C.S. Rai	B N Rishi	01516404518

Department of Computer Science

University school of information, communication & Technology

Guru Gobind Singh Indraprastha University, New Delhi (2018-2021)

CERTIFICATE

DECLARATION BY CANDIDATE

This is to certify that the project entitled "***Personality detection through text***" is a bonafide record of independent project/research work done by me under supervision of ***Dr. C.S. Rai*** sir and submitted to Guru Gobind Singh Indraprastha University in partial fulfillment for the award of the Degree of MCA(SE). I Certify the content of the project are authentic and original

Date: 23.04.2020 B N Rishi (01516404518)

CERTIFICATION BY MENTOR

This is to certify that the project entitled " ***Personality detection through text***." is a bonafide record of independent project/research work done by B.N.Rishi bearing enrollment number 01516404518 under my supervision.
To the best of my knowledge and believe work done by candidate is original and has not been submitted for award of any other degree.

Date: 23.04.2020

C.S.Rai
(professor)
University school of information,
communication and technology

ACKNOWLEDGEMENT

I would like to express my sincere thanks to those who have contributed significantly to this report. It is a pleasure to extend the deep gratitude to my guide ***Mr C.S Rai*** sir for his valuable guidance and support to continuously promote me for the progress of the report. I hereby present my sincere thanks to him for his valuable suggestions towards my report, which helped me in making this report more efficient and user friendly.

I thank each and everyone's efforts who helped me in some or the other way for small and significant things.

B N Rishi
(01516404518)

LIST OF FIGURES

Table of Contents

ABSTRACT

The project is about how the personality can be derived from the list of essays which can be listed among any of the 5 big personality traits accompanied by the list of features or reflecting the nature of the essay from 10 features. We would be using list of essays with its 5 big traits being already implemented as a dataset for the proceedings. In the list of essay dataset every word is mined to match words with the dictionary dataset along with the charged values and creating a separate list of words with its charged values. Training the data using dictionary list efficiently by iterating every essay and generating list of 15 elements for every essay. Data list is classified using various algorithms like Multi-layer perceptron, Support vector machine, Random forest, K-nearest neighbor to evaluate accuracy by displaying the features of the input essay along with 5 big traits.

INTRODUCTION

The project is about how the personality can be derived from the list of essays which can be listed among any of the 5 big personality traits reflecting the nature of the essay from 10 features.

In recent times, the interest of the scientists is leaning towards personality recognition which is growing quickly. There are applications that can make use of personality recognition are social network, recommendation/review systems, deception detection, authorship attribution, sentiment analysis/opinion mining, among others. It is being proven from previous researches that personality is correlated with many parts of life, like job success, happiness, negativity, depression or anxiety.

Personality is an important human characteristic and it also portrays the individuality. It is one of the basic aspects, by which we get to know various types of people and their selfless in a better way. It is considered to be one of the long-term goals and a difficult task for psychologists to evaluate human selfless and its effects on human nature. A person's reaction to a certain conditions plays a major role on how the person is depending on the situation. But, in most of the time, people react with respect to their personality or nature. It is possible to extract someone's personality traits by text samples to automatically identify personality and predict their reactions and behaviour. Researchers around the world are working on this domain especially computational linguistics such as machine learning, natural language processing predominantly in artificial intelligence.

Personality detection from text means to resolve and extract the certain characteristics of a person who have written the text. Various domains which are using in daily life like job recruitment, psychologists can use the personality detection model that adapts the interactivity according to user's nature such that can arrive to conclusion about how to deal with certain human personalities.

The sub-objectives are as follows: -

1. We would be using list of essays with its 5 big traits being already implemented as a dataset for the proceedings.
2. After Training the data using dictionary list efficiently by iterating every essay and generating list of 15 elements for every essay.

3. Data list is classified using various algorithms like Support vector machine, Random forest, K-nearest neighbour to evaluate accuracy by displaying the features of the input essay along with 5 big traits.

The Big five personality traits: -

1 **Extroversion (EXT):** Is this person outgoing, talkative, and energetic or is he reserved and solitary?

2 **Neuroticism (NEU):** Is this person is sensitive and nervous or is he secure and confident?

3 **Agreeableness (AGR):** Is this person trustworthy, straightforward, generous, and modest or is he unreliable, complicated, and boastful?

4 **Conscientiousness (CON):** Is this person efficient and organized or is he sloppy and careless?

5 **Openness (OPN):** Is this person inventive and curious or is he dogmatic and cautious?

The working in this field is beneficial for many activities that are performing by means of online facilities on a daily basis like customer care support, and suggestions of services and products, etc.

Problem Statement

In recent times, the interest of the scientists is leaning towards personality recognition which is growing rapidly. There are applications that can make use of personality recognition such as social network, recommendation/review systems, deception detection, authorship attribution, sentiment analysis/opinion mining, among others. It is being proven from continuous researches that personality is correlated with many parts of life, like job success, happiness, negativity, depression or anxiety.

The working in this field is beneficial for many activities that are performed by the means of online facilities on a daily basis like customer care support, suggestions of services and products etc.

In India, if government opens vacancies for government jobs, millions of applications are dropped which is beyond the possibility of a human to analyse or read each and every applications and there may be chances that a deserving candidates may get rejected because human brains tends to get tired during manual processing of applications. Similarly, thousands of job applications are to be analysed by the HR teams of companies to map them to eligibility criteria required to be fulfilled by selected candidates. In the meantime the developers of the e-commerce resources are steadily improving various algorithms to help the customers obtain/select products and services that maps the needs more accurately and representing the products in a more reachable way to increase sales. These tasks may require a crucial step of mental or through a user's pattern analysis of user personality. Personality detection models may be useful in various domains such as e-learning, information filtering and e-commerce which can use these ways humans are interacting and to process or resolve problems according to some coherent patterns.

LITERATURE SURVEY

Personality detection from text means trying to identify what kind personality traits does an individual possesses by inspecting something written by him/her in the form of paragraph or an essay. Deep leaning is a technique that can be used very efficiently for this purpose which has several algorithms like random forest , KNN , MLP etc. for doing the job. Knowing someone's personality type is very crucial as it plays a very major role in determining how the person would react to certain situations and conditions. This is particularly helpful for the recruitment teams of offices, the psychiatrists , psychologists among others. Though the results can be misleading in the case where the person knows he will be judged on the basis of what he is writing and so he writes something fake to appear to be of a certain personality type.

There are many research work published in the area of personality detection from text using deep learning. Alexander Gelbukh et al., [1] have worked in this area and presented a method that uses convolutional neural network (CNN) for fetching the major traits of personality of an author from paragraphs or essays written by him by training 5 different networks for the five personality traits. The output of each network gave a binary value which represented the presence or absence of a particular trait.

Xiangguo Sun et al., [6] have presented their work to show that the structure of texts can play a lead role in detecting personality from texts. CLSTM is the name of their proposed model, which detect user's personality using structures of texts. CLSTM is nothing but concatenation of LSTMs (Long Short Term Memory networks) which are bidirectional in nature with CNN (Convolution Neural Network). Two different kinds of datasets were used for evaluation which had long texts and short texts respectively.

Di Xue et al., [6] have used online social network posts of users to predict their personalities through deep learning. For learning deep semantic features from every user's posts which are textual in nature, they utilized a deep neural network which is of the form of hierarchy that comprised of a variant of the Inception structure and their self – developed AttRCNN structure. Then these deep semantic features were added to the statistical linguistic features and were fed into traditional regression algorithms. This is how finally the real-valued scores of Big Five personality traits were predicted.

Tatiana Litvinova et at., [5] have worked on Authorship Profiling. Authorship Profiling is a term that means analysing a user's text and disclosing information about them. This paper aims to find the probability of self-destructive behaviour of an individual through their text. Specifically Russian language text was used for this research. For the implementation purposes, a mathematical model was designed to predict the probability of self-destructive behaviour of a person which is calculated on the basis of calculations of set of correlations between scores on the Freiburg Personality Inventory scales and text variables (average sentence length, lexical diversity etc.).

Basant Agarwal [7] has given an overall review of the basic methods that are available for the personality detection from the social network texts of users and also he tried to highlight all the main and useful datasets that are publically available for doing the same.

PROPOSED METHODOLGY

3.1 ARCHITECTURE OF PROPOSED METHODOLGY

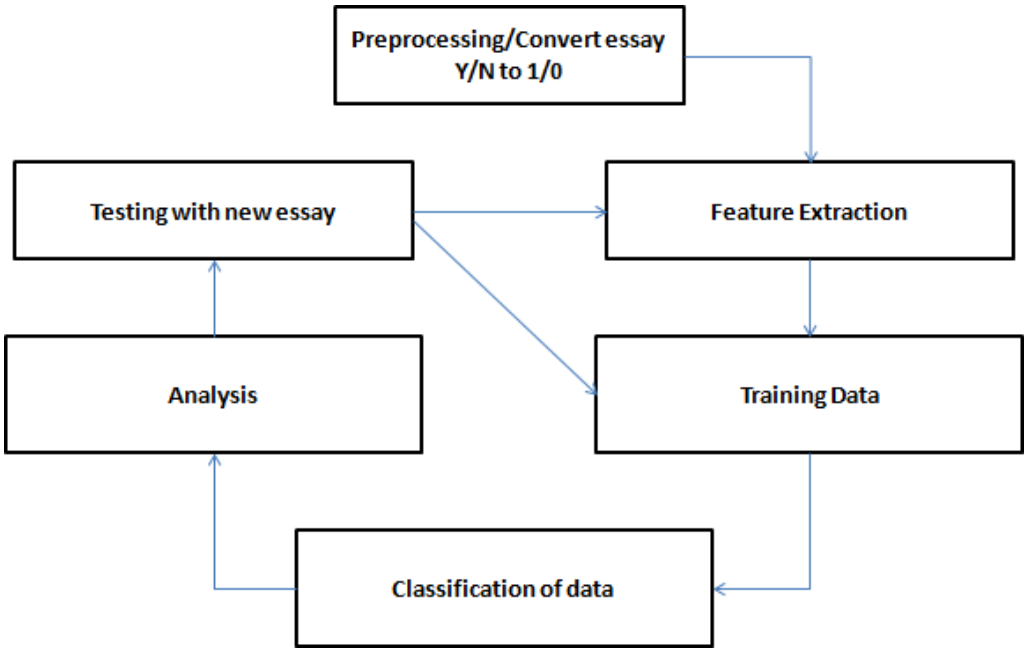


Figure 1 : proposed methodology

Pre-processing /Convert Y/N to 1/0

Pre-processing a text work is one of the very important jobs that is done for any **NLP** application. There are some standard steps that go along with most of the applications, whereas sometimes we need to do some customized pre-processing.

We do not always get general stop-words. The corpus contains some unnecessary repetitive words that are of no use in the analysis. So, those words act as noise. This decision of adding corpus specific repetitive threshold is usually decided by analyzing the corpus by finding percentage occurrence of each word. Always removing language specific stop words is not recommended because there might be a case where are they are useful like grammar correction where the system has to appropriately add articles in a sentence. Now here *a, an, the,* are not supposed to be in the stop list. Removing punctuations can pose serious issues. In general, boundary punctuations can be removed without any issues but same doesn't hold for the cases where punctuations occur within a word. Such cases don't work well with tokenizers.

Feature Extraction

Feature extraction is a process which reduces the dimensions of a dataset. In this process, some better manageable groups (features) are generated from the initial given set of variables which are then used for the purpose of processing. The completeness and accuracy of the original data set remains intact.

We need to reduce the input data to limited set of features for which feature vector is an alternate name, when the input data is in large volumes and is also expected to be repetitive. Feature selection is a process of determining subset of the initial features. The desired task can be performed by using the selected features because they hold all the necessary and important information from the input data.

Training of data

It is the most important stage of any machine learning project. Algorithms are nothing without data as they utilize the training data they're given for finding relationships, developing understanding, taking decisions, etc. The quality of the training data is a major reason behind the good or bad performance of the algorithm.

The success of the data project depends equally on both the quanlity and quantity of the training data as well as on the algorithms performances. Now, even if a large amount of well-structured data is stored, the labelling might not be appropriate for training the model.

In other words, we need labelled data for training. Or there might be a need of more data for better functioning of algorithms.

It is obvious, if we are trying to make a good model, good training data is the basic requirement.

Classification of data

Categorizing and sorting data into numerous forms , types or any other separate class is called data classification. We perform data classification to classify and separate data according to the requirements of data set for various personal objectives or businesses. For sorting data in a repository or dataset, there are many criteria and methods available in data classification.

Analysis

Analysis is that stage of the project where the final analytical results of the projects are derived. There are multiple ways of doing the analysis like graph analysis, pie-chart analysis etc. the major analysis tool is developing comparative graphs. Basically this stage helps in doing comparison between different techniques and drawing conclusion on the basis of performance parameters. Also checking accuracy of the techniques applied constitutes a major portion of this phase

3.2 DATASETS

3.2.1 Essays.CSV

Essays is a dataset containing a set-of-awareness texts (about 2468, one for each person), labelled with personality classes. It is made by analyzing people such that the person of this dataset can take help of psychologists because each and every person has various approaches towards a situation. Texts had been also produced by students who took the Big 5 test. Then the scores were computed by Mairesse and converted scores to numerical classes by authors with a median split. Output document is used widely in use particularly in the field of personality detection.

3.2.2 NRC_Emotion.TXT

This dataset is used to obtain charged emotionally words. Lexicon contains 14,182 words with 10 attributes: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, and trust. (http://saifmohammad.com/WebPages/NRCEmotion-Lexicon.htm).

We considered a word is charged emotionally if it had at least one of these features mentioned above; there are 6,468 words in the lexicon.

3.3 MODEL USED - BIG FIVE MODEL

- This model is the well worked or analyzed metrics of personality domain in recent times is the “Big Five” model for personality. It’s being evaluated by extracting and predicting certain patterns of texts repeatedly such that to arrive to a strong conclusion about a human, patterns plays an important role because one time or the other people would definitely shows their authentic nature’s pattern which requires time and space. This model is being widely used personality traits structure. The human personality is computed as a list of five values with respect to bipolar traits. This is model is popular among the language and computer science researchers.
  Personality is formally described in terms of the Big Five Personality traits, which are the following binary (yes/no) values:
- **OPN (Openness):** Artistic, imaginative, curious and intelligent. Those individuals who get high scores in this category tend to be sophisticated and artistic and appreciate different ideas, views and experiences.
- **CON (Conscientiousness):** Efficient, organized, responsible and persevering. Individuals who are conscientious are extremely reliable and most probably high achievers, planners and hard workers.
- **EXT (Extraversion):** Energetic, assertive, active and outgoing. People scoring high in this area are supposed to be energetic and friendly, extroverts who get inspired from their social situations.
- **AGR (Agreeableness):** Compassionate, helpful, cooperative and nurturing. High scorers in agreeableness are peaceful, optimistic people.
- **NEU (Neuroticism):** Anxious, self-pitying, tense, insecure and sensitive. Such individuals are generally moody and tense.

IMPLEMENTATION

4.1 ALGORITHMS IMPLEMENTED

4.1.1 Random Forest

- A Random Forest is a collection or ensemble of a large number of simple tree predictors. Each tree produces a result when a set of predictor values is given to it. It can be implemented for both the problems i.e., classification and regression. For regression problems, the response of each tree is an estimate of the dependent variable given by the predictors. And for classification problems, a set of independent predictor values are associated with one of the categories present in the dependent variable by its response which takes the form of a class membership.

4.1.2 Linear Support Vector Machine

- Support Vector Machines work on the concept of decision planes which define decision boundaries. A set of objects having different class memberships are distinguished by a decision plane. A schematic example is shown in the illustration below in which the objects belong either to class GREEN or RED. A boundary is defined by the separating line which separates the GREEN and RED objects by having all the GREEN objects on the right side and all the RED objects on the left side. Any new object (white circle) falling to the right is labelled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

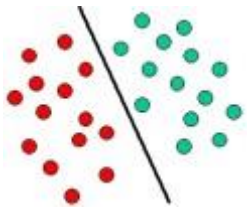


Figure 2: Example of linear classifier 1

4.1.3 K-Nearest Neighbour

- A very basic but important classification algorithm is K-Nearest Neighbour. It lies in the supervised learning area and has applications in pattern recognition, data mining and intrusion detection.

- As it is non-parametric it is widely used in real-life scenarios, meaning, it does not make any elementary assumptions about the distribution of data while other algorithms such as GMM assumes a Gaussian distribution of the given data.

  Some training data is given to us priory, which classifies coordinates into groups identified by an attribute.

- In K-NN, all computation is postponed and until classification and the function is only approximated locally. It is a type of instance-based learning, or lazy learning. The K-NN algorithm is one of the modest machine learning algorithms.

- In K-NN, no external training step is necessary and also it is not needed.

- The sensitivity to the local structure of the data is a very unique and peculiar feature of K-NN. Always remember k-means and k-NN are two totally different machine learning techniques.

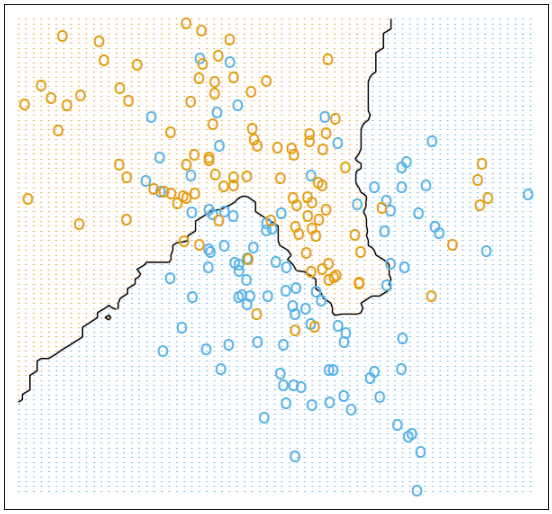


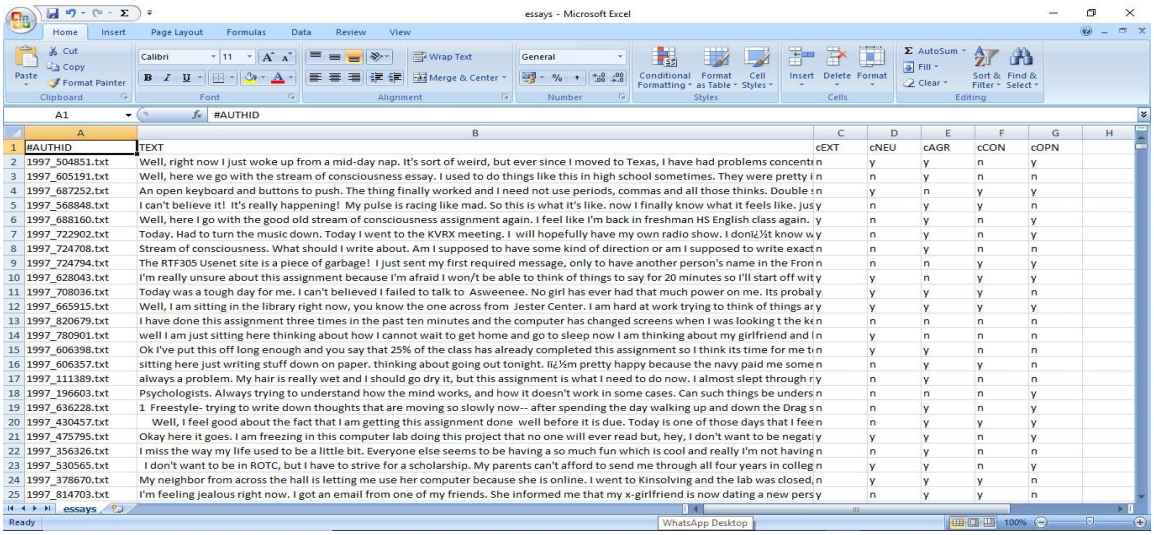
Figure 3 : KNN Algorithm

4.2 PHASES OF IMPLEMENTATION

4.2.1 Preprocessing/Convert Y/N to 1/0

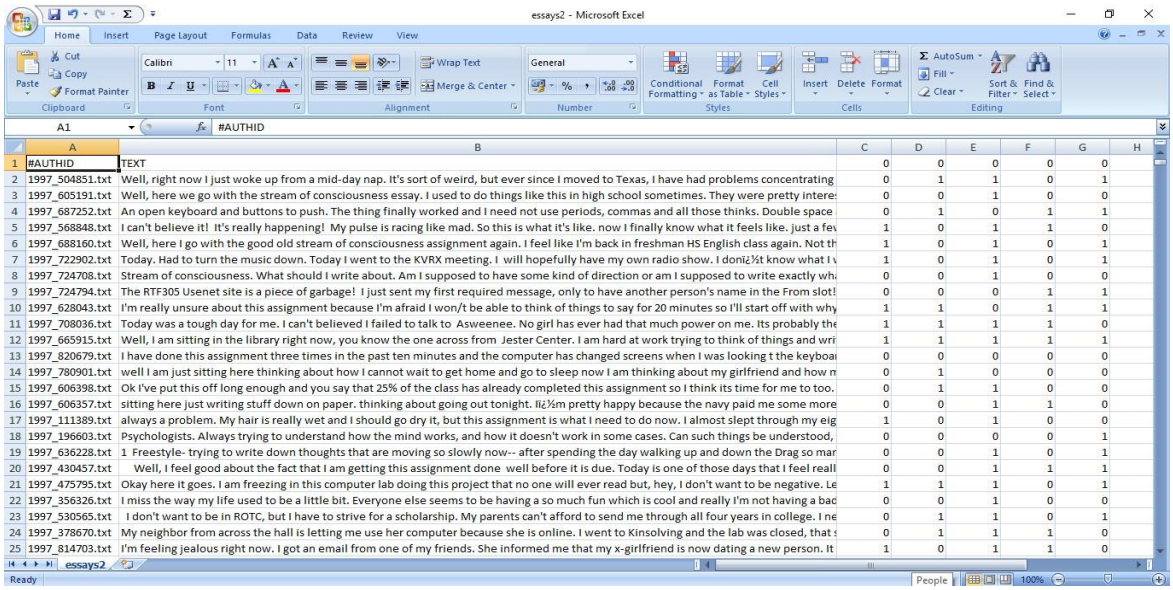
This is phase is required to clean data and also we needed to convert Y/N to 1/0 because in this project we are doing classification and classification is a process which is all based upon mathematical expressions and it doesn't support strings**.**

- Sentence splitting to words using delimiters.
- Data cleaning which has no emotionally charged words.
- Assuming that the sentence at least contains one emotionally charged word.
- NRC Emotion dataset is used to set weights according to the words used in a sentence



#AUTHID	TEXT	cEXT	cNEU	cAGR	cCON	cOPN
1997_504851.txt	Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever since I moved to Texas, I have had problems concentr	n	y	y	n	y
1997_605191.txt	Well, here we go with the stream of consciousness essay. I used to do things like this in high school sometimes. They were pretty i	n	n	y	n	n
1997_687252.txt	An open keyboard and buttons to push. The thing finally worked and I need not use periods, commas and all those thinks. Double s	n	y	n	y	y
1997_568848.txt	I can't believe it! It's really happening! My pulse is racing like mad. So this is what it's like. now I finally know what it feels like. jus	y	n	y	y	n
1997_688160.txt	Well, here I go with the good old stream of consciousness assignment again. I feel like I'm back in freshman HS English class again.	y	n	y	n	y
1997_722902.txt	Today. Had to turn the music down. Today I went to the KVRX meeting. I will hopefully have my own radio show. I donï¿½t know w	y	n	y	n	y
1997_724708.txt	Stream of consciousness. What should I write about. Am I supposed to have some kind of direction or am I supposed to write exact	n	n	y	n	y
1997_724794.txt	The RTF305 Usenet site is a piece of garbage! I just sent my first required message, only to have another person's name in the Fron	n	n	n	y	y
1997_628043.txt	I'm really unsure about this assignment because I'm afraid I won/t be able to think of things to say for 20 minutes so I'll start off wit	y	n	y	y	y
1997_708036.txt	Today was a tough day for me. I can't believed I failed to talk to Asweenee. No girl has ever had that much power on me. Its probal	y	y	y	y	n
1997_665915.txt	Well, I am sitting in the library right now, you know the one across from Jester Center. I am hard at work trying to think of things ar	y	y	y	y	y
1997_820679.txt	I have done this assignment three times in the past ten minutes and the computer has changed screens when I was looking t the ke	n	n	n	n	n
1997_780901.txt	well I am just sitting here thinking about how I cannot wait to get home and go to sleep now I am thinking about my girlfriend and I	n	y	n	n	n
1997_606398.txt	Ok I've put this off long enough and you say that 25% of the class has already completed this assignment so I think its time for me ti	n	y	n	n	n
1997_606357.txt	sitting here just writing stuff down on paper. thinking about going out tonight. Iï¿½m pretty happy because the navy paid me some	n	n	y	y	n
1997_111389.txt	always a problem. My hair is really wet and I should go dry it, but this assignment is what I need to do now. I almost slept through r	y	n	y	n	n
1997_196603.txt	Psychologists. Always trying to understand how the mind works, and how it doesn't work in some cases. Can such things be unders	n	n	n	n	y
1997_636228.txt	1 Freestyle- trying to write down thoughts that are moving so slowly now-- after spending the day walking up and down the Drag s	n	n	y	n	y
1997_430457.txt	Well, I feel good about the fact that I am getting this assignment done well before it is due. Today is one of those days that I fee	n	n	y	y	y
1997_475795.txt	Okay here it goes. I am freezing in this computer lab doing this project that no one will ever read but, hey, I don't want to be negati	y	y	n	n	y
1997_356326.txt	I miss the way my life used to be a little bit. Everyone else seems to be having a so much fun which is cool and really I'm not having	n	n	y	n	n
1997_530565.txt	I don't want to be in ROTC, but I have to strive for a scholarship. My parents can't afford to send me through all four years in colleg	n	y	y	n	y
1997_378670.txt	My neighbor from across the hall is letting me use her computer because she is online. I went to Kinsolving and the lab was closed,	n	y	y	y	n
1997_814703.txt	I'm feeling jealous right now. I got an email from one of my friends. She informed me that my x-girlfriend is now dating a new pers	y	n	y	y	n

Figure 4 Essay.csv



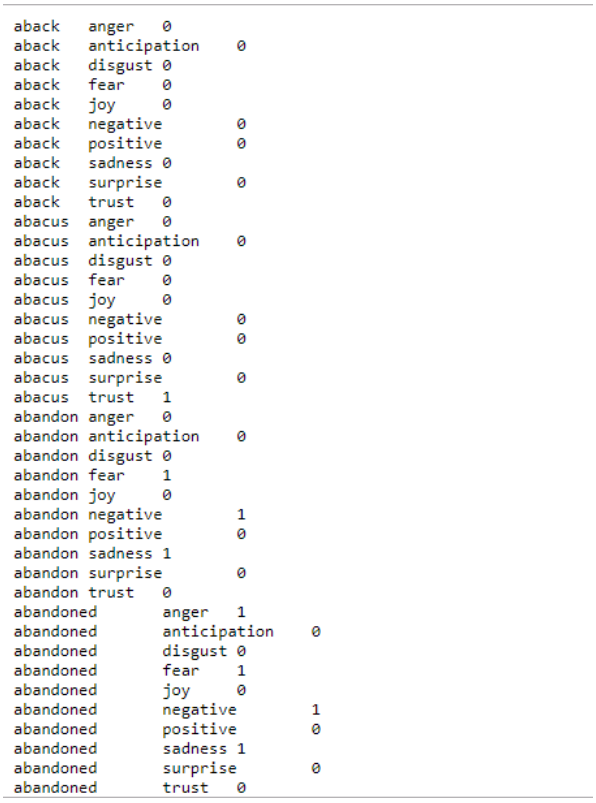
#AUTHID	TEXT					
1997_504851.txt	Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever since I moved to Texas, I have had problems concentrating	0	0	0	0	0
1997_605191.txt	Well, here we go with the stream of consciousness essay. I used to do things like this in high school sometimes. They were pretty interes	0	1	1	0	1
1997_687252.txt	An open keyboard and buttons to push. The thing finally worked and I need not use periods, commas and all those thinks. Double space	0	0	1	0	0
1997_568848.txt	I can't believe it! It's really happening! My pulse is racing like mad. So this is what it's like. now I finally know what it feels like. just a fe	0	1	0	1	1
1997_688160.txt	Well, here I go with the good old stream of consciousness assignment again. I feel like I'm back in freshman HS English class again. Not th	1	0	1	1	0
1997_722902.txt	Today. Had to turn the music down. Today I went to the KVRX meeting. I will hopefully have my own radio show. I donï¿½t know what I v	1	0	1	0	1
1997_724708.txt	Stream of consciousness. What should I write about. Am I supposed to have some kind of direction or am I supposed to write exactly wha	0	0	1	0	0
1997_724794.txt	The RTF305 Usenet site is a piece of garbage! I just sent my first required message, only to have another person's name in the From slot!	0	0	0	1	1
1997_628043.txt	I'm really unsure about this assignment because I'm afraid I won/t be able to think of things to say for 20 minutes so I'll start off with why	1	1	0	1	1
1997_708036.txt	Today was a tough day for me. I can't believed I failed to talk to Asweenee. No girl has ever had that much power on me. Its probably the	1	1	1	1	0
1997_665915.txt	Well, I am sitting in the library right now, you know the one across from Jester Center. I am hard at work trying to think of things and wri	1	1	1	1	1
1997_820679.txt	I have done this assignment three times in the past ten minutes and the computer has changed screens when I was looking t the keyboa	0	0	0	0	0
1997_780901.txt	well I am just sitting here thinking about how I cannot wait to get home and go to sleep now I am thinking about my girlfriend and how n	0	1	0	0	0
1997_606398.txt	Ok I've put this off long enough and you say that 25% of the class has already completed this assignment so I think its time for me to too.	0	1	1	0	0
1997_606357.txt	sitting here just writing stuff down on paper. thinking about going out tonight. Iï¿½m pretty happy because the navy paid me some more	0	0	1	1	0
1997_111389.txt	always a problem. My hair is really wet and I should go dry it, but this assignment is what I need to do now. I almost slept through my eig	1	0	1	0	0
1997_196603.txt	Psychologists. Always trying to understand how the mind works, and how it doesn't work in some cases. Can such things be understood,	0	0	0	0	1
1997_636228.txt	1 Freestyle- trying to write down thoughts that are moving so slowly now-- after spending the day walking up and down the Drag so mar	0	0	1	0	1
1997_430457.txt	Well, I feel good about the fact that I am getting this assignment done well before it is due. Today is one of those days that I feel reall	0	0	1	1	1
1997_475795.txt	Okay here it goes. I am freezing in this computer lab doing this project that no one will ever read but, hey, I don't want to be negative. Le	1	1	1	0	1
1997_356326.txt	I miss the way my life used to be a little bit. Everyone else seems to be having a so much fun which is cool and really I'm not having a bad	0	0	1	0	0
1997_530565.txt	I don't want to be in ROTC, but I have to strive for a scholarship. My parents can't afford to send me through all four years in college. I ne	0	1	1	0	1
1997_378670.txt	My neighbor from across the hall is letting me use her computer because she is online. I went to Kinsolving and the lab was closed, that s	0	1	1	1	0
1997_814703.txt	I'm feeling jealous right now. I got an email from one of my friends. She informed me that my x-girlfriend is now dating a new person. It	1	0	1	1	0

Figure 5 : Essay.csv after 1st phase

4.2.2 Feature Extraction

Ten features include:

- **Anger**
- **Anticipation**
- **Joy**
- **Positive**
- **Negative**
- **Disgust**
- **Surprise**
- **Fear**
- **Sadness.**
- **Trust**



```
aback   anger   0
aback   anticipation    0
aback   disgust 0
aback   fear    0
aback   joy     0
aback   negative        0
aback   positive        0
aback   sadness 0
aback   surprise        0
aback   trust   0
abacus  anger   0
abacus  anticipation    0
abacus  disgust 0
abacus  fear    0
abacus  joy     0
abacus  negative        0
abacus  positive        0
abacus  sadness 0
abacus  surprise        0
abacus  trust   1
abandon anger   0
abandon anticipation    0
abandon disgust 0
abandon fear    1
abandon joy     0
abandon negative        1
abandon positive        0
abandon sadness 1
abandon surprise        0
abandon trust   0
abandoned       anger   1
abandoned       anticipation    0
abandoned       disgust 0
abandoned       fear    1
abandoned       joy     0
abandoned       negative        1
abandoned       positive        0
abandoned       sadness 1
abandoned       surprise        0
abandoned       trust   0
```

Figure 6: NRC_Emotion.txt

- We will convert the above nrc.txt dataset into category wise ex: anger, anticipation etc. in a csv file along with the words (charged value) in the form of list (better.csv) as shown below .

4.2.3 Train data

- First it will fetch data from a file (essay2.csv). Then it will analyse words from fetched data and it will map the words with output of feature extraction module and based on the mapping it will assign the value to 10 features iteratively.

Example :

24,16,10,4,1,31,5,8,7,13

28,16,15,11,5,31,8,14,11,18

18,13,25,8,9,32,8,3,5,15

21,14,13,7,9,17,9,9,9,14

13,8,24,11,6,25,5,7,12,16

19,20,14,6,5,46,4,12,5,19

25,27,13,2,7,38,4,7,9,25

- Likewise 2468 essays will be analysed and vectors are generated corresponding to every essays such that a file (trainv1.csv) is obtained containing all the numeric vectors. So that a new input essay can be classified according to the prior experience/training.
- Now as per (essay2.csv) every essay had been charged with five big traits such that the new essay can be mapped to one of the big five traits on the basis of the essay.
- A new file (train_essay2.csv) containing vectors (list length of 15) is generated.

Example :

'anticipation', 'joy', 'negative', 'sadness', 'disgust', 'positive', 'anger', 'surprise', 'fear', 'trust', 'score', 'label'

- The list will contain 10 features and 5 big traits according to the training analysis.

Example :

24,16,10,4,1,31,5,8,7,13,0,1,1,0,1

28,16,15,11,5,31,8,14,11,18,0,0,1,0,0

18,13,25,8,9,32,8,3,5,15,0,1,0,1,1

21,14,13,7,9,17,9,9,9,14,1,0,1,1,0

13,8,24,11,6,25,5,7,12,16,1,0,1,0,1

19,20,14,6,5,46,4,12,5,19,1,0,1,0,1

25,27,13,2,7,38,4,7,9,25,0,0,1,0,0

10,11,14,6,7,19,7,2,6,7,0,0,0,1,1

- This step is important because the most of the time is spent to build the heart of the object because if training fails accuracy decreases, and the future prediction will be incorrect.
- This step also provides an input to the next classification phase so that labelled supervised learning can be implemented effectively.

If we want to check our own essay

- Essay is added in single_essay.csv in double quotes (“// string “) so that execution could be easy as the whole essay is treated as a string.
- We would execute trainBuild_single.py module to train the new essay on the basis of previously trained dataset.
- After the successful execution of trainBuild_single.py, we would run a module to classify each of the 10 features with their number of counts in the input file and in which of the big 5 trait category does it below.

4.2.4 Classification of data

- In the classification first we will split the vectors which were obtained in the previous phase where there was a list of 15 elements vector length.
- For classification purpose we would install scikit (sklearn) by executing the below command.
- We would import sci-kit sklearn to get access to various supervised classification functions:-
    1. svm.SVC()
    2. ensemble.RandomForestClassifier()
    3. neighbours.KNearestNeighbour()
    4. neural_network.MLPClassifier()

RESULT SCREENSHOTS

Steps Involved

- Run **convert_csv_y_n_to_0_1_essays.py** to convert initial **essays.csv's** 5 trait factor (i.e., in Y/N) to 1/0 known as embedding to enhance accurate predictions which is based on mathematical operations.
- In this step, we will be converting Y/N to 1/0 of every essays 5 big trait.

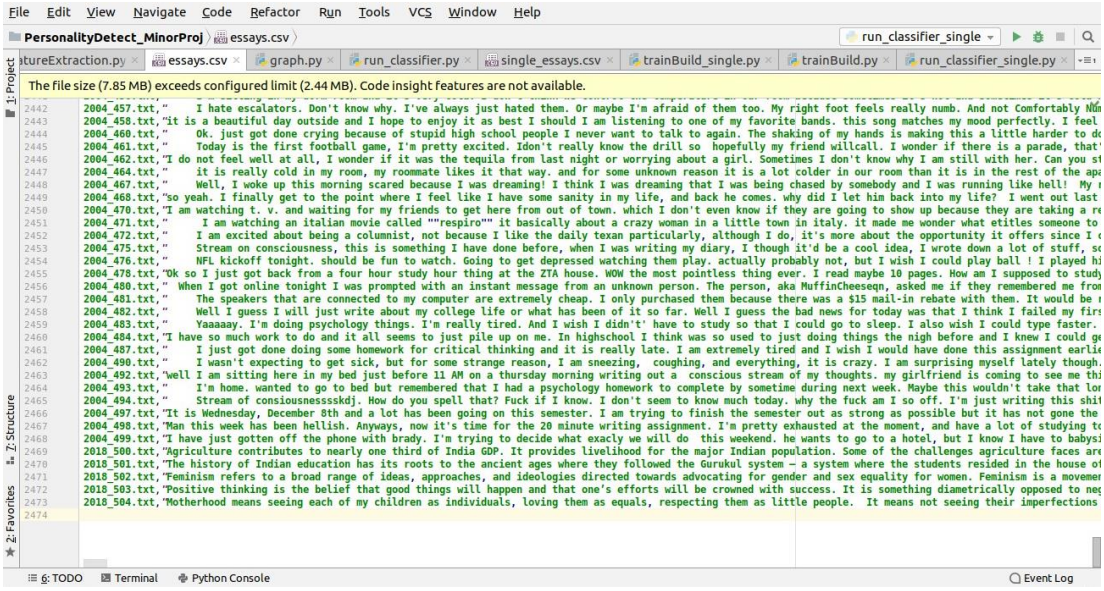


Figure 7: Essays.csv(1)

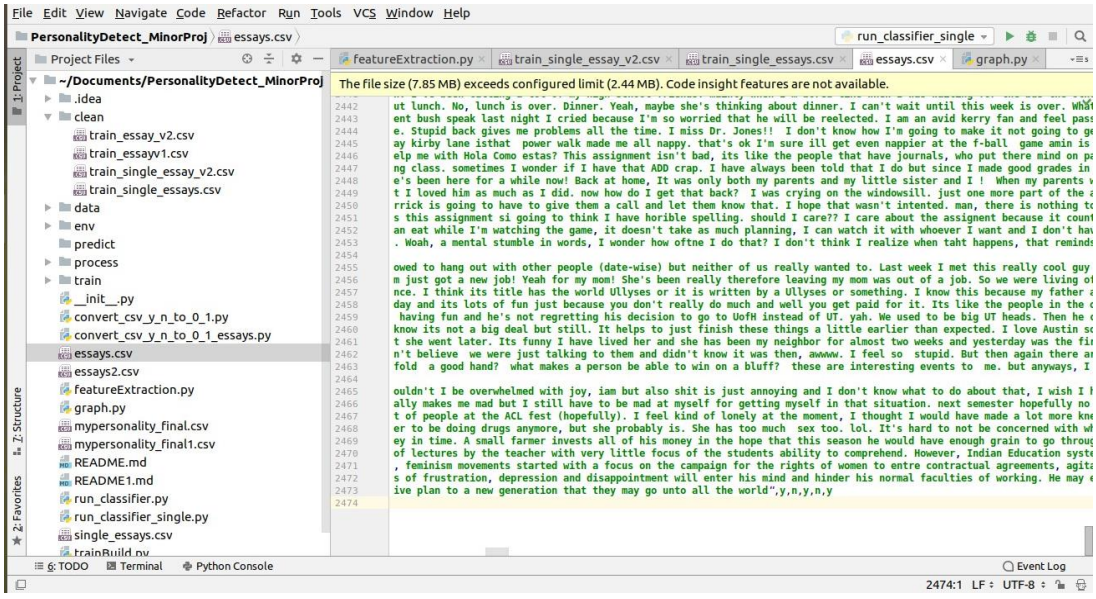


Figure 8: Essays.csv(2)

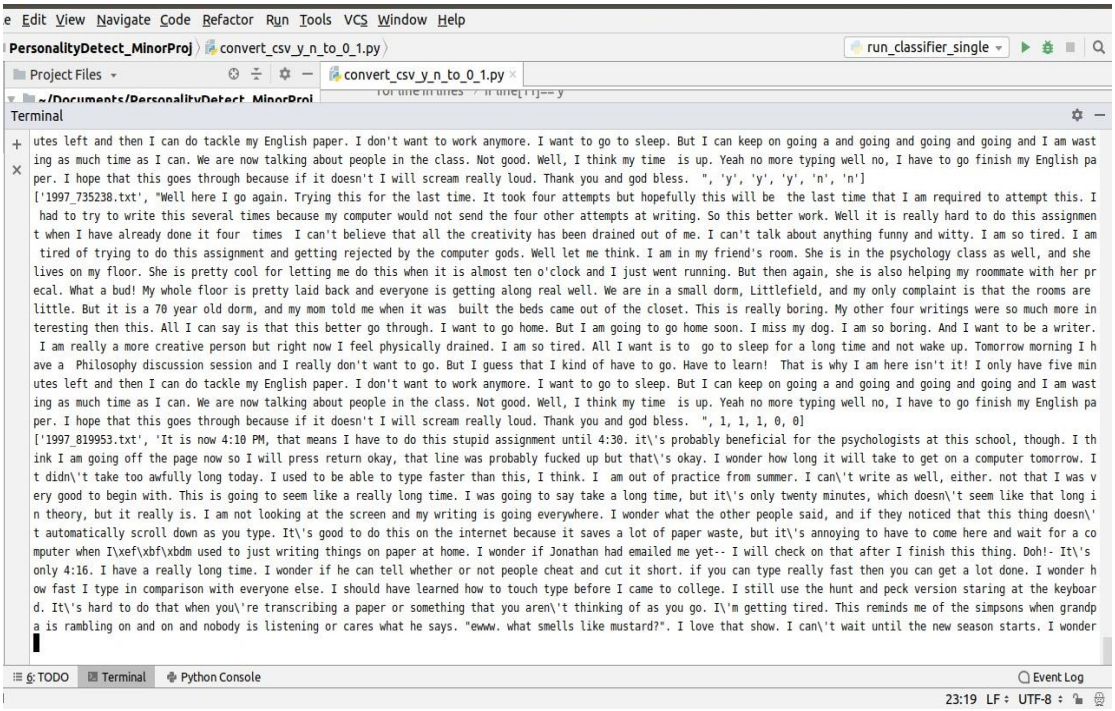


Figure 9: convert_csv_y_n_to_0_1.py terminal running

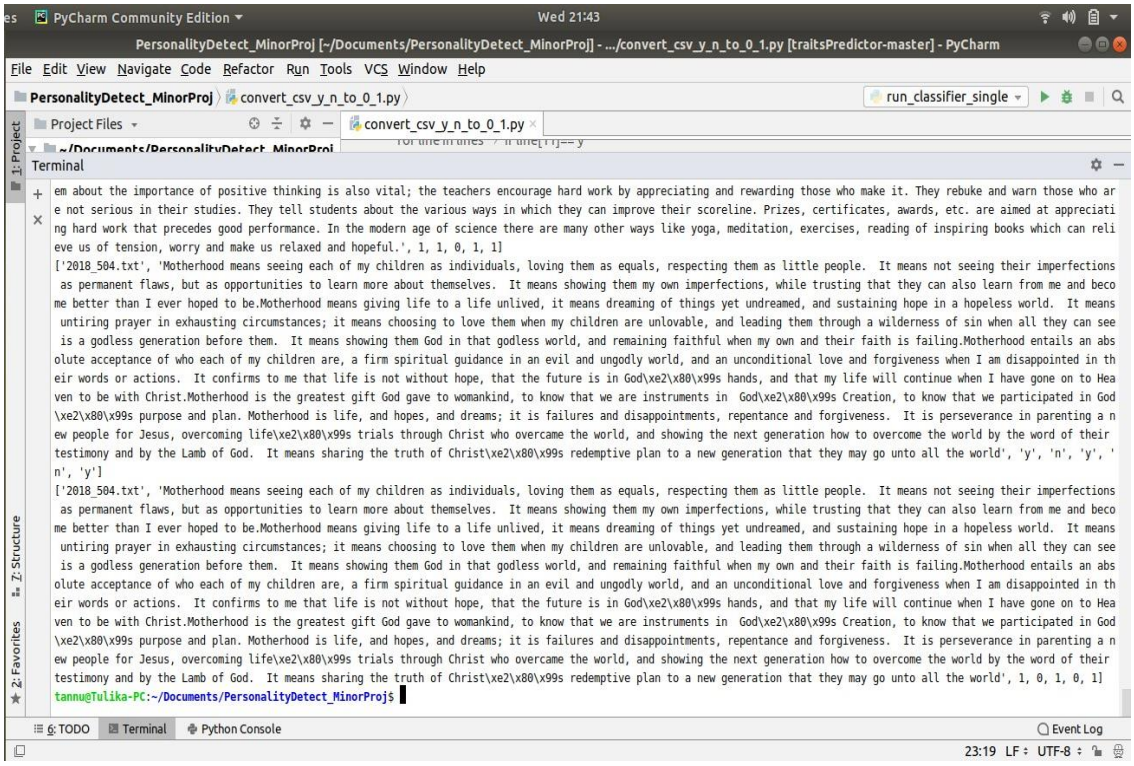


Figure 10:convert_csv_y_n_to_0_1.py executed

- Next step is to extract 10 features with the help of NRC.txt dataset
- Arranging every words of the nrc.txt dataset in the list of vectors format accompanied by the charged values of those words.
- This step is important which facilitates the training of the dataset efficiently.

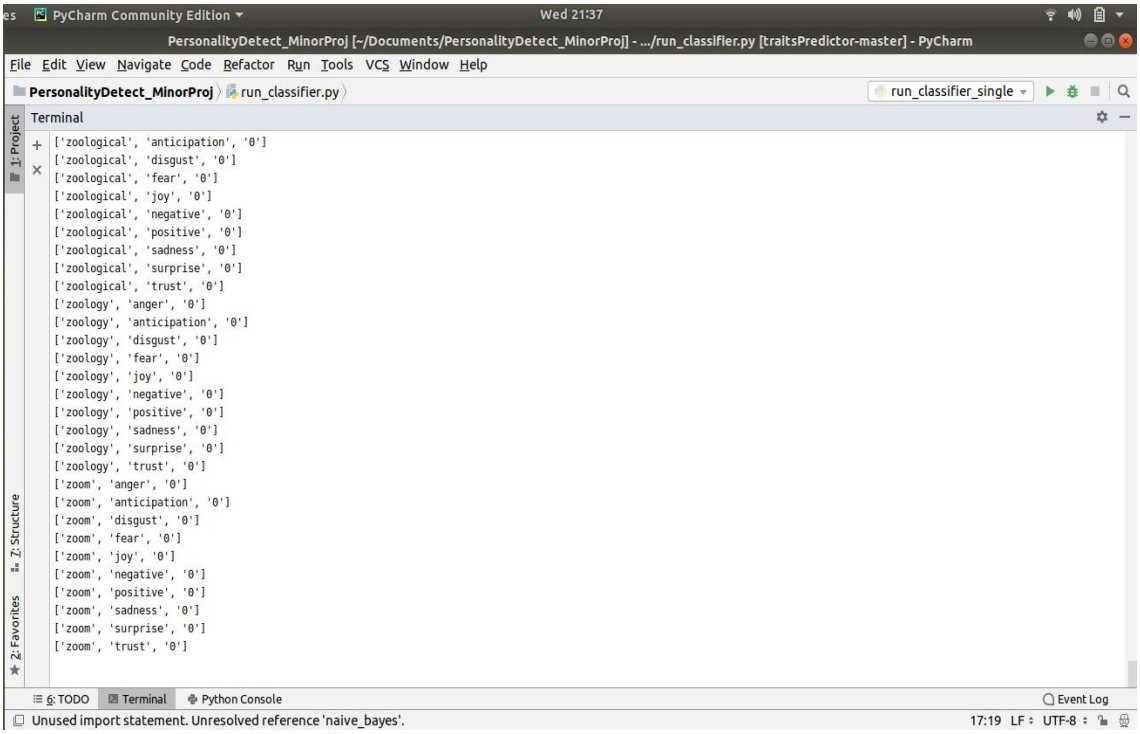


Figure 11: FeatureExtraction.py executed all words tillZ alphabet

- Better.csv file is obtained containing all words with its charged vector list which would be used in training.

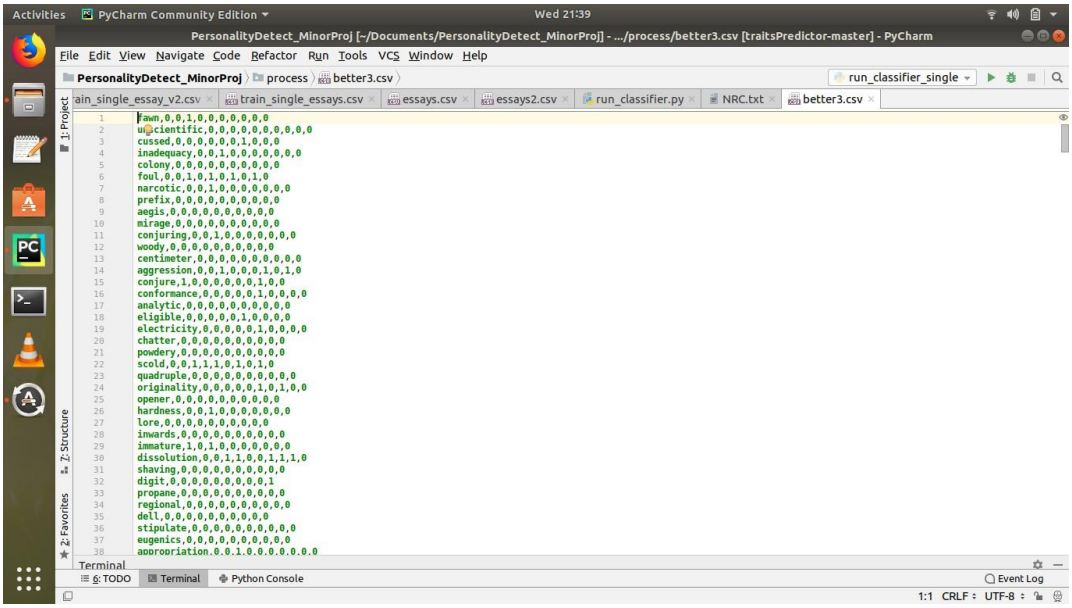


Figure 12: Better.csv file

- Next and the most important step is to train data based on the essay dataset.
- Accurate prediction depends upon the quality training done by the program accompanied by efficient dataset.
- trainBuild.py takes better.csv as input to train effectively.
- It generates train_essayv1.csv which is favourable when new essay or new document is added such that the next essay to be predicted can be judged accordingly.
- train_essayv1.csv contains the new file including which has been added recently such that a separate csv file is maintained every time a new essay has been added in essay dataset. The new train_essayv1.csv is generated only if dataset got altered or updated otherwise it remains the same on the basis of which prediction takes place.
- In simple terms, train_essayv1.csv contains "N" documents such that it could efficiently judge "N+1"th essay easily.

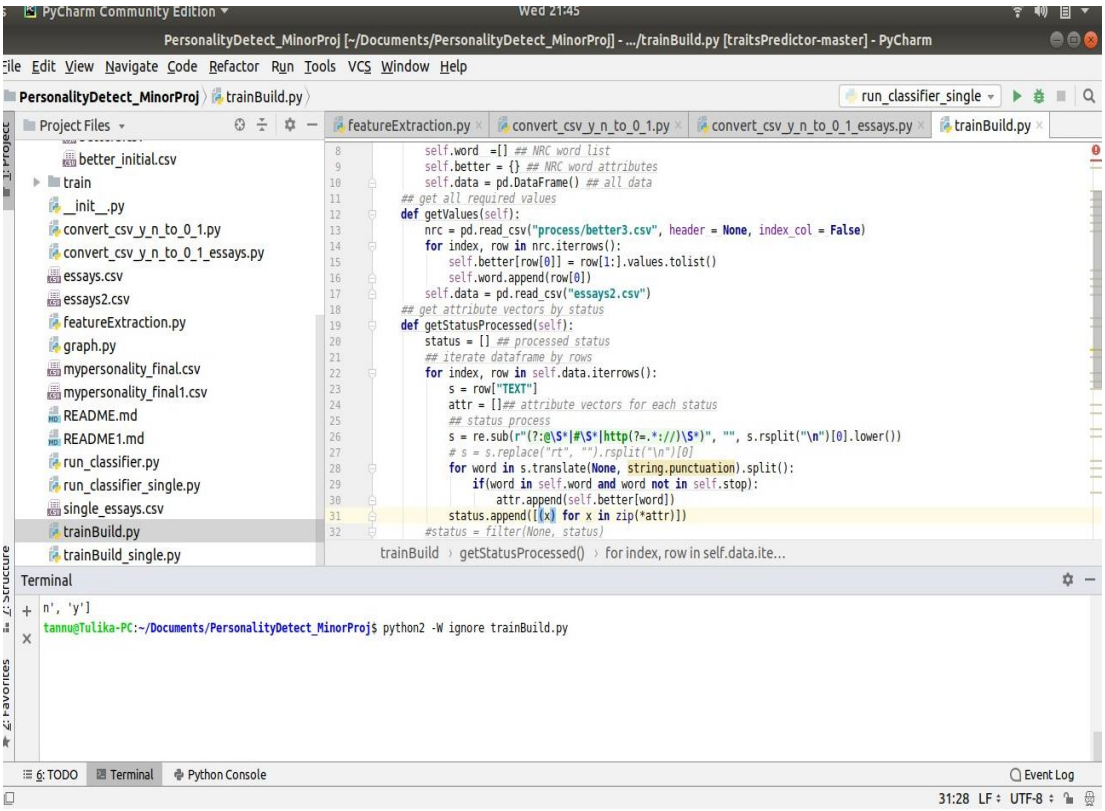


Figure 13: Run trainBuild.py in terminal

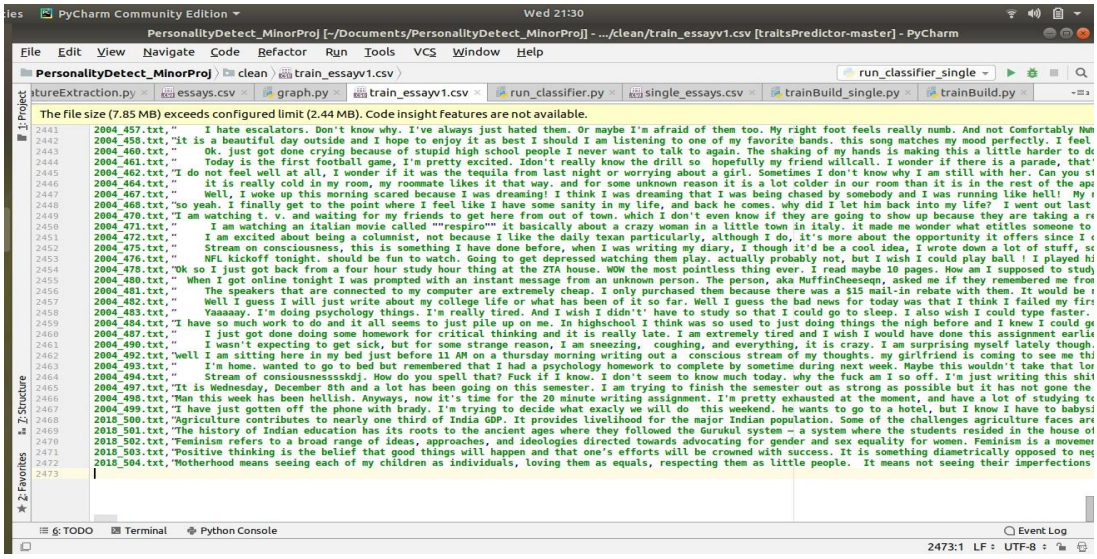


Figure 14: Train_essayv1.csv file (1)

- TrainBuild.py also generates train_essayv2.csv file such that it contains "N" document vector containing predictions based on 10 features and 5 big trait separately.
- This file only contains every essay's corresponding 10+5 features value vector and the predicted 5 big traits such that every list size is 15.
- This is helpful which would be easy to classify based on two parameters i.e., features and labels.
- Train_essayv2.csv is produced with the help of feature extraction better.csv file.

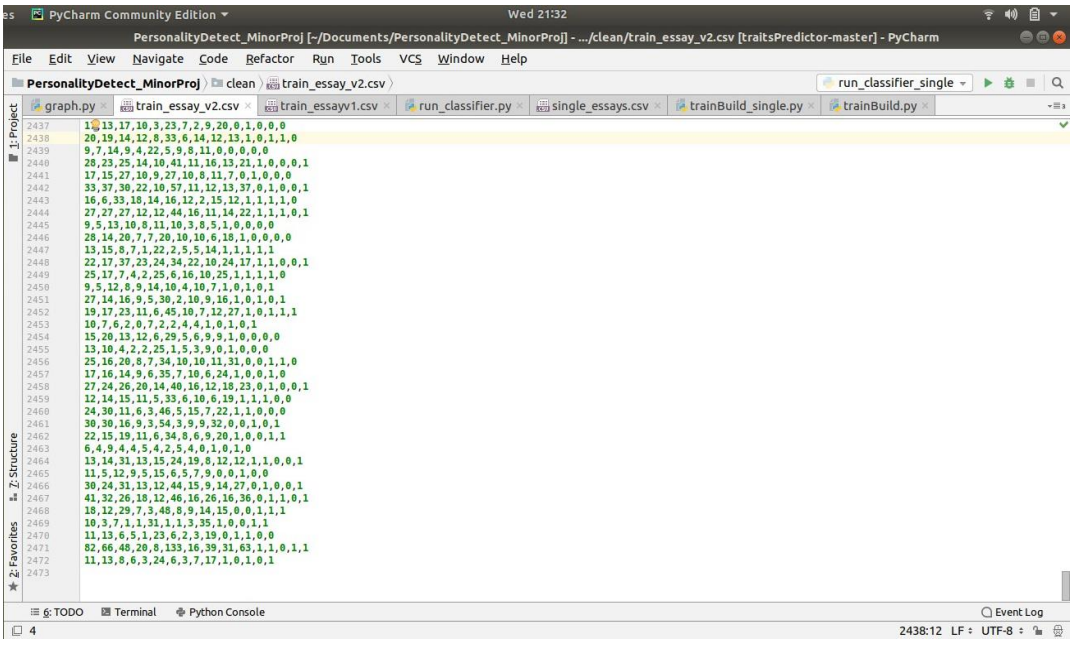


Figure 15: Train_essayv2.csv (features and labels)

- Next step is to classify based on train_data and test_data.
- Run_classifier.py splits the previous output (train_essayv2.csv) data into features and labels.
- We would be using four algorithms namely: - Linear SVM, Random Forest & KNN.
- Functions are imported from sklearn module.

Now, testing with our own new essay and to predict it trait.

- Adding a new essay in single_essays.csv inside (“ “).

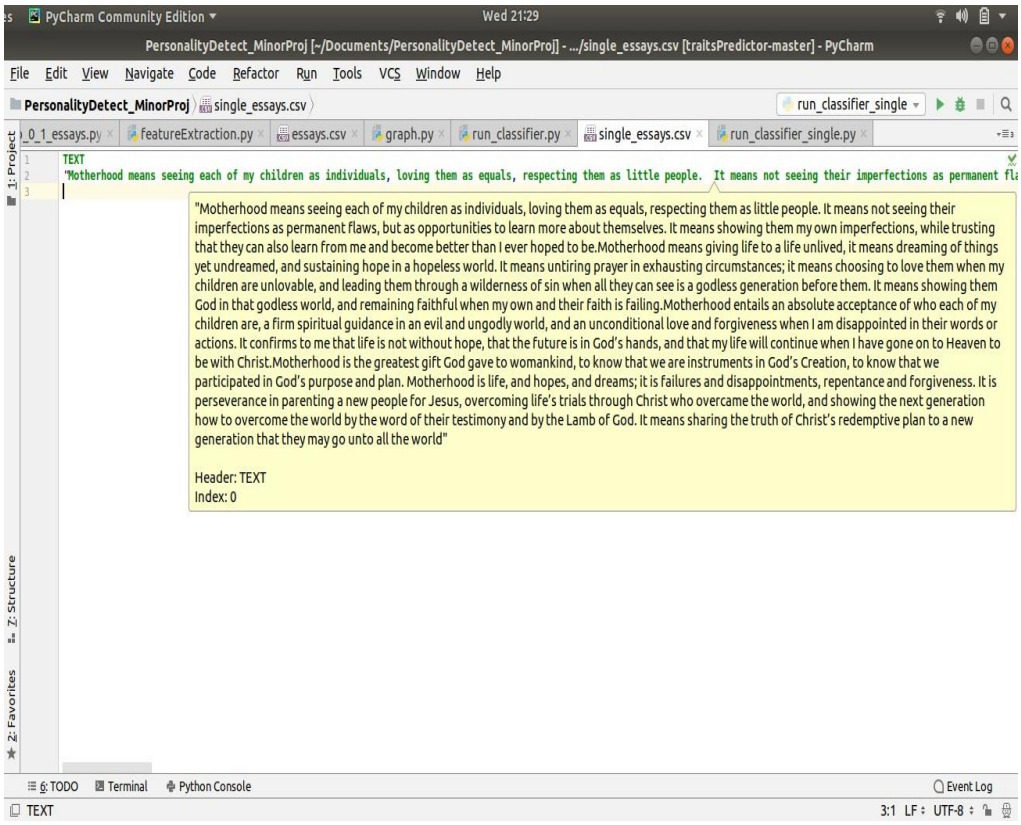


Figure 16: New essay prediction

- Based on the predictions did previously by train_essayv1.csv and train_essayv2.csv is used in this step to judge new essay or based on the experience new essay is predicted.
- We would run trainBuild_single.py to train new essay based on train_essayv1.csv and train_essayv2.csv which generates train_essayv1_single.csv and train_essayv2_single.csv.

- train_essayv2_single.csv contains 10 features predictions based on previous experiences.

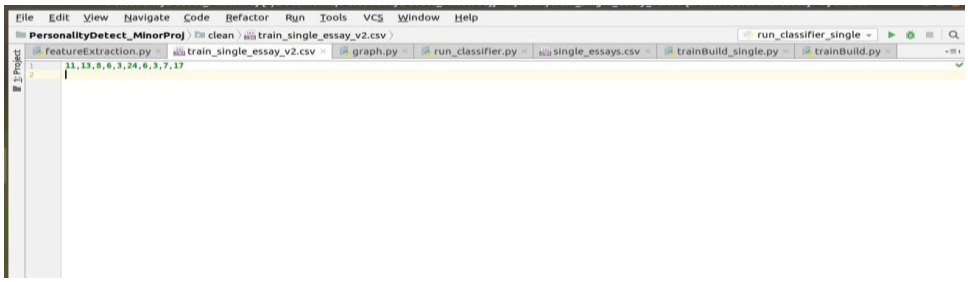


Figure 17: Train_essayv2_single.csv

- We would execute run_classifier_single.py to display output containing 10 features and judging on what trait does the essay belongs to using various algorithms.
- Displaying Output.

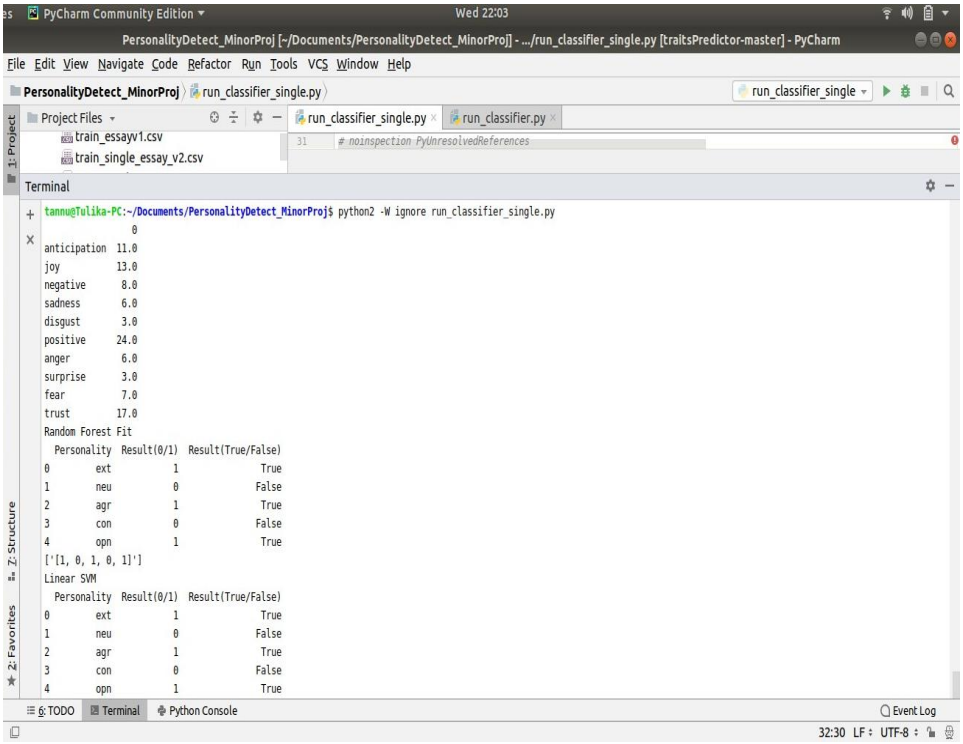


Figure 18: Output (1)

ANALYSIS

Random-Forest analysis

- Sqrt and Log2 max_feature are approximately performs the same.
- After considering 10 iterations it is found that log2 lags in some cases compared to sqrt feature.
- 9/10 predictions is found to be accurate in both the cases.
- **Conclusion :- Sqrt ~= Log2**

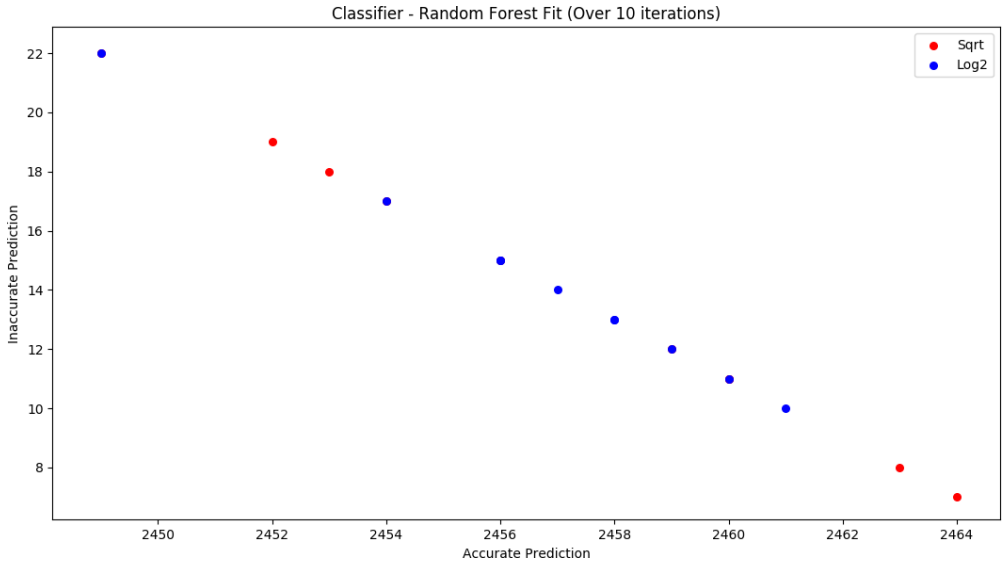


Figure 19: RF Analysis

K-Nearest Neighbour analysis

- Ball-tree performs the best as compared to all the four algorithms implemented.
- In Contrast, KD-Tree performs the worst out of four.
- Auto and Brute algorithm were approximately considered to be same but in some cases brute outperformed auto algorithm.
- **Conclusion :- Ball-Tree > Brute~=Auto > KD-Tree**

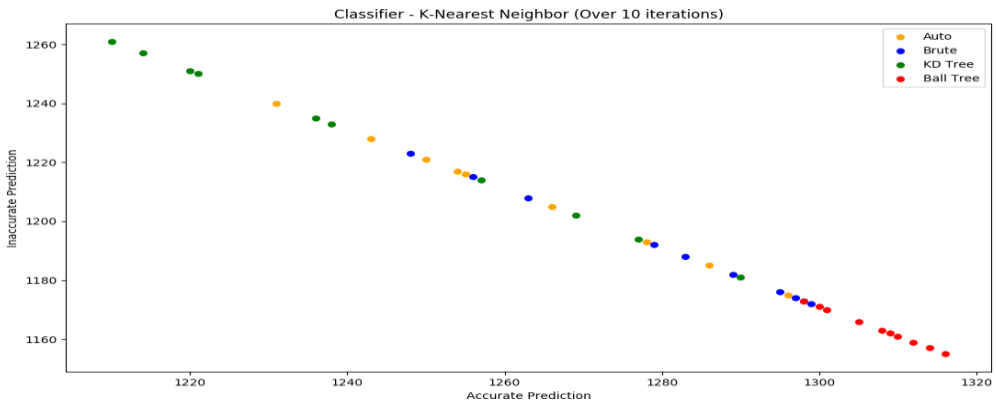


Figure 20: KNN Analysis

Support Vector Machine analysis

- Since the project is based upon supervised learning because datasets are primarily considered to be labelled, which restricts the usage of other variations of SVM
- **Conclusion: - SVM predicted 8.5/10 essays to be correct.**

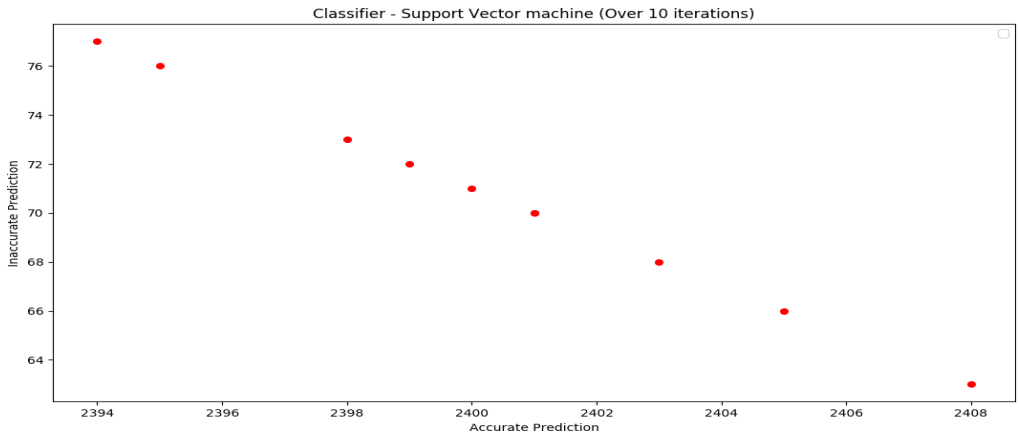


Figure 21: SVM Analysis

Conclusion Graph

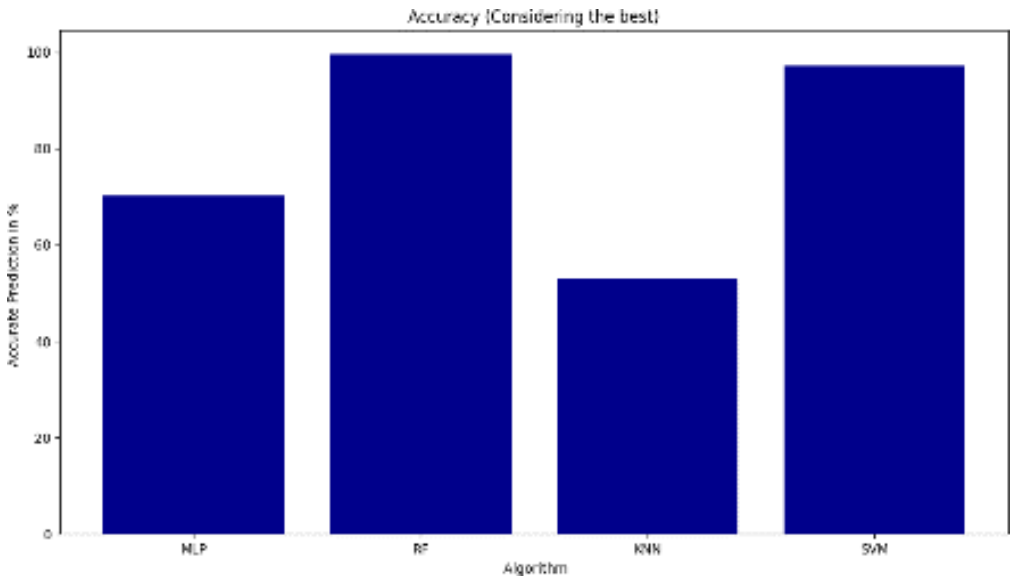


Figure 22: Conclusion graph

CONCLUSION

Everything in this world has some positive as well as negative attributes, how to overcome those negative attributes is the real challenge.

This project successfully predicts the nature of human being or any text being provided by human on the basis of five big traits. It's extremely helpful in future as we may need program instructions to decide human's nature especially in the field of psychology, recruitment process etc.

There may be chances that the human may fake the document by writing pretty things which may be uncovered by the program may lead to incorrect prediction as the document is unauthentic.

As far as the project is concerned, the project works pretty well as far as the document is authentic from the human point of view.

Random forest and SVM success rate is above 95% as compared to MLP and KNN which has a success rate of 74% and 52% respectively.

FUTURE WORK

- Extracting features from emoticons [😖, 😉, 😐 ,☹].
- Adding a new trait to detect negativity, depression, pessimism.
- Editing NRC.txt with charges to add certain abbreviations [ROFL, LOL, DND, ASAP TC, TTYL etc.] to facilitate better understanding of human text and derive the emotions.
- Making attractive Graphical User Interface for the Windows version and developing android application for mobile usage.
- Rather than using text on a digital device for personality detection, real handwriting to be used for input, the system will scan the image and consequently personality type will be detected.

REFERENCES

[1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "*Deep Learning-Based Document Modelling for Personality Detection from Text*" in *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017.

[2] B. Liu "*Sentiment Analysis and Opinion Mining, ser. Synthesis Lectures on Human Language*", Technologies. Morgan & Claypool Publishers, 2012.

[3] "*Variations of Support Vector Machine classification Technique*", Bhavsar, Hetal & Ganatra, Amit, 2013.

[4] F. Mairesse et al., "*Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*" J. Artificial Intelligence Research, vol. 30, 2007, pp. 457–500.

[5] Litvinova, Tatiana & Seredin, P & Litvinova, Olga & Zagorovskaya, Olga. (2016). "*Profiling a set of personality traits of text author: What our words reveal about us. Research in Language*."

[6] X. Sun, B. Liu, J. Cao, J. Luo and X. Shen, "*Who Am I? Personality Detection Based on Deep Learning for Texts*" 2018 *IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018, pp. 1-6.

[7] Basant Agarwal *"Personality detection from text: A review"* 2014.