Using textual data for Personality Prediction:A Machine Learning Approach

Aditi V.Kunte
School of Computer Engineering
Dr.Vishwanath Karad MIT World Peace University
Pune,India
Aditikunte95@gmail.com

Suja Panicker
Assistant Professor
School of Computer Engineering
Dr.Vishwanath Karad MIT World Peace University
Pune,India
suja.panickar@mitwpu.edu.in

***Abstract*—Personality is an important parameter as it differentiates various individuals from one another. Personality prediction is an evergreen area of research. Predicting personality with the help of data through social media is a promising approach as this method does not require any questionnaires to be filled by users thus reducing time and increasing credibility.**

Thus having knowledge of personality is an interesting domain for researchers to work on. Predicting personality has many applications in real world. Use of social media is increasing day by day. Huge amount of textual data as well as images continue to explode to the web daily. Current work focuses on Linear Discriminate Analysis, Multinomial Naive Bayes and AdaBoost over Twitter standard dataset.

Keywords—Machine Learning, Big Five test, Social Media, Statistical analysis.

I. Introduction

Personality is a key aspect of human life. More specifically personality is a branch of psychological study. Personality is constituted of elements like person's thoughts, feelings, behavior which continuously keeps on changing over time. Prediction of personality is an area of study where person gets categorized in a class according to his/her personality. There are number of psychological tests that yield different type of personality classes. Popular tests include Big Five test [1, 2, and 3] which yield personality classification in five categories as openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. DISC is another test of psychology that classifies personality in categories as Dominance, Influence, Steadiness and Compliance [4]. MBTI psychological test has 16 categories of personality [5, 6]. All these traditional methods of personality prediction use questionnaire for personality prediction. Filling a lengthy questionnaire is time consuming and tedious job.

Advanced machine learning algorithm - AdaBoost has been used in [8] for prediction. Boosting algorithms are used to gain highest accuracy and high performance. Hence we have employed AdaBoost with other algorithms like LDA and Multinomial Naïve Bayes to analyze accuracies of algorithms

Social media is a promising approach for this task. It is easier to gather a dataset from social media like Twitter or Facebook and use it for prediction. Hence in this work we have used real-time Twitter dataset for predicting personality. Twitter provides API like Twitter streaming API [7] which is useful for building dataset.

This paper is structured as follows: Section 2 describes overview of literature survey, Section 3 gives detailed analysis of applied algorithms, Section 4 illustrates future scope and Section 6 presents the conclusion.

II. Literature survey

Various researchers have contributed to the task of personality prediction. Some of them have taken psychological tests into account for deciding labels of personality while others have used machine learning algorithms like Naïve Bayes[9] for prediction. Given below is brief overview of literature survey.

In [10] linear regression and support vector regression have been used to predict personality of Facebook users. MyPersonality dataset is used for this work. Results of the study shows that linear regression proves better option for prediction. In [11] suicide related posts are determined from Twitter posts using random forest, simple logistics and J48. Twitter streaming API is used to collect dataset and Martingale framework is used to predict results. This is a noteworthy approach. [12] Uses binary SVM for mental disorder detection. Dataset for this is generated using Amazon Mechanical Turk. Future work of this study can be in the domain like fetching multimedia contents for prediction.

In [13] CNN and DCNN are used to find microblog sentiments. Sina Weibo micro blogging site has been used to gather dataset. This study evaluates sentiment like positive, negative and neutral. In [14] Tree, SVM, Ensemble, kNN, and ANN are used for ECG authentication and gender recognition. CYBHi and ECG-ID datasets have been used for this study. This study has accurate results of about 98% for age prediction and 94% for gender prediction.

User's digital footprints have been extracted and analyzed in [15]. Several meta analyses have been done for prediction using digital footprints. Big five personality traits were used and this study is useful for recommending products or services to user according to his taste. [16] Aims in finding marijuana based posts on Facebook posted by people of different background and trying to analyze sentiment based on their replies on posts. For this task 15,000 posts and 14 million users' reaction from high times

978-1-7281-3651-6/19/$31.00 ©2019 IEEE

magazine Facebook page is collected and features such as post ID, number, and types of reactions and comments with associated user IDs, etc are recorded. NLP processing such as removing stop words, tokenization, stemming is applied to the extracted posts. Thus negative and positive sentiments according to word usage are determined [23].

[17] Analyzes different features of troll and authorize user on different twitter platforms. Sequential mining optimization, random forest and Naive Bayes algorithms are used for this study. Dataset for this study is prepared on the basis of users who are affected and not affected by trolling. Results of this study demonstrate that automatic classification is useful for the process of detection. [18] Focuses on finding negativity in personality using data from Facebook. Approaches used to execute this include LIWC software, Hogan Development Survey, Ordinary Least Square regression and Least Absolute Shrinkage and Selection Operator regression. Data of 51,712 Facebook users were collected to analyze and features are extracted according to 11 measures of HDS. Results of the study are recorded separately for two machine learning algorithms.

[19] Proposed prediction of emotions using deep multimodal architecture. Approaches used in this study include Deep Multimodal Long Term Memory (DMLTM) and Long Short Term Memory (LSTM). Seempad dataset is used in this study and emotional features as angry, surprised, disgusted, happy, neutral and EEG data (workload and engagement) are captured. Results of this study demonstrate DMLTM approach has highest accuracy, i.e. 69% over LSTM.

A. *Summary of literature survey*

- There has been extensive work in the area personality prediction still there are some areas that remained uncovered.
- Datasets used by researchers are not available publicly also most of the APIs mentioned in the work have been deprecated so those APIs cannot be used for dataset collection.
- There has been work on textual datasets or image datasets but concept of multimodal dataset have not been extensively studied.

III. Proposed System Architecture

System architecture for the current work is given in Fig.1.

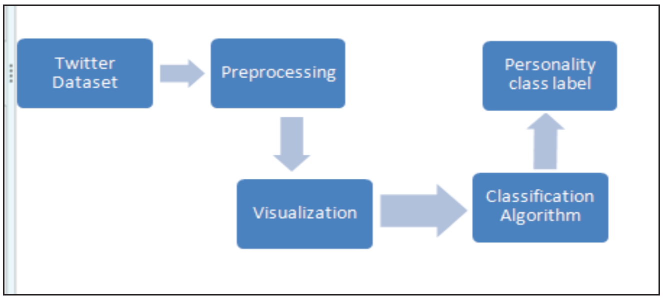


Fig. 1. System Architecture

As illustrated in Fig.1 [20], real time dataset is collected using Twitter's streaming API and stored in '.csv' file. This collected dataset is then pre-processed. Pre-processing involves processes like tokenization, stemming and removal of stop-words. Pre-processing is necessary as it avoids further complications in execution. Pre-processed data is then visualized using visualization library provided by python.

After all the initial settings, classification algorithms like AdaBoost, Multinomial Naïve Bayes and LDA have been applied to our dataset, result of which classify personality in any one of the five class labels provided by Big five test of psychology.

Entire execution flow of system is explained in next sections.

IV. Experimentation and Results

A. *Dataset Creation*

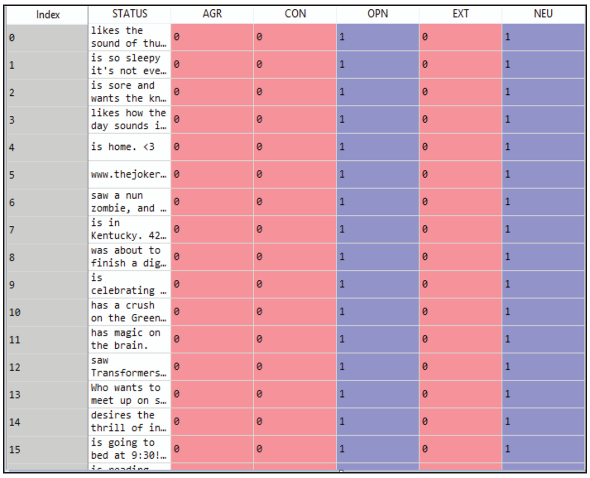
As given in Fig.1, dataset is created using Twitter streaming API. Twitter provides a Twitter streaming API which is used to collect real-time tweets from Twitter.

Following are the steps to fetch real-time data from twitter.

- In order to collect tweets from Twitter, one needs to have a Twitter account.
- After logging in to your Twitter account go to Twitter developer settings and follow the steps specified by the developers.
- After all the processing is completed, you will get 'Twitter APIKey' and 'APISecret' key which are needed to be used in python program to fetch tweets from Twitter.

Python provide library named 'Tweepy' which is important for all these processing.

Snapshot of collected dataset is given below in Fig.2.



Index	STATUS	AGR	CON	OPN	EXT	NEU
0	likes the sound of thu…	0	0	1	0	1
1	is so sleepy it's not eve…	0	0	1	0	1
2	is sore and wants the kn…	0	0	1	0	1
3	likes how the day sounds i…	0	0	1	0	1
4	is home. <3	0	0	1	0	1
5	www.thejoker…	0	0	1	0	1
6	saw a nun zombie, and …	0	0	1	0	1
7	is in Kentucky. 42…	0	0	1	0	1
8	was about to finish a dig…	0	0	1	0	1
9	is celebrating …	0	0	1	0	1
10	has a crush on the Green…	0	0	1	0	1
11	has magic on the brain.	0	0	1	0	1
12	saw Transformers…	0	0	1	0	1
13	Who wants to meet up on s…	0	0	1	0	1
14	desires the thrill of in…	0	0	1	0	1
15	is going to bed at 9:30!…	0	0	1	0	1

Fig. 2. Snapshot of Twitter Dataset

- Description of dataset

As given in Fig.2, first column is 'index' which is for indexing of tweets. Second column is 'Status' which contains text of the tweet. Rest of the columns used to decide class labels. The class labels are decided according to categories of personality given by Big Five psychological test [21].

Given below is the brief description of class labels.

TABLE I. FEATURES OF DATASET

Feature	Description
NEU	**Neuroticism** People of this category are often tensed and depressed.
EXT	**Extraversion** People of this category are energetic and they like the company of people.
OPN	**Openness** People of this category are creative and non-judgmental.
CON	**Conscientiousness** People of this category are focused to their work
AGR	**Agreeableness** People of this category are kind and good natured

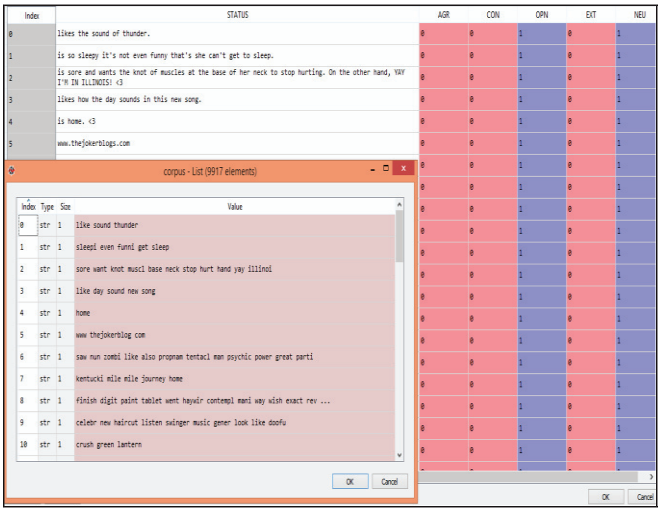
B. Pre-processing

As discussed in section II, data need to be preprocessed in order to achieve expected accuracy. Data mining provides the feature for data pre-processing which is required for cleaning the data, removing outliers and inconsistencies.

For the current work we have used and replaces through preprocessing steps like tokenization, stemming and removing of stop words.

Tokenization separates each word of a sentence, which is referred as 'tokens'. Stemming finds root word of the sentence and replace it with original word. For example, the word 'Leaves' will be converted to 'Leaf' in stemming. In the process of stop wards removal, entire stop words of the document like -is, am, are etc are removed.

Snapshot of preprocessed dataset is given below in Fig.3.



Index	STATUS	AGR	CON	OPN	EXT	NEU
0	likes the sound of thunder.	0	0	1	0	1
1	is so sleepy it's not even funny that's she can't get to sleep.	0	0	1	0	1
2	is sore and wants the knot of muscles at the base of her neck to stop hurting. On the other hand, YAY I'M IN ILLINOIS! <3	0	0	1	0	1
3	likes how the day sounds in this new song.	0	0	1	0	1
4	is home. <3	0	0	1	0	1
5	www.thejokerblogs.com	0	0	1	0	1

corpus - List (9917 elements)

Index	Type	Size	Value
0	str	1	like sound thunder
1	str	1	sleepi even funni get sleep
2	str	1	sore want knot muscl base neck stop hurt hand yay illinoi
3	str	1	like day sound new song
4	str	1	home
5	str	1	www thejokerblog com
6	str	1	saw nun zombi like also propnam tentacl man psychic power great parti
7	str	1	kentucki mile mile journey home
8	str	1	finish digit paint tablet went haywir contempl mani way wish exact rev ...
9	str	1	celebr new haircut listen swinger music gener look like doofu
10	str	1	crush green lantern

Fig. 3. Snapshot of pre-processed dataset

As given in Fig.3, the corpus window just below the main dataset shows preprocessed data.Following are the changes that occurred after preprocessing the data.

- First text of the dataset 'likes the sound of thunder' has been converted to 'like sound thunder' ,this indicate stopwords like 'the' and 'of' are removed in stopword removal process.
- 'is sore and wants the knot of muscles at the base of her neck to stop hurting. On the other hand, YAY I'M IN ILLINOIS! <3' has been converted to 'sore want knot muscl base neck stop hurt yay illinois' ,thus here all the capital letters has been converted to lower case and special symbol '<3 ,!' also have been eliminated.
- In 'www.thejokerblogs.com', preprocessing removes '.'and makes it like 'www thejokerblogs com'.

Thus, this is how preprocessing works in removing stop-words ,converting words to lowercase and removing of special characters.

C. Experimentation and Results

For the given dataset we are focusing on AdaBoost, Multinomial Naïve Bayes and LDA algorithms. Results of these are discussed below.

- Comparison of accuracies

Given below is the graphical representation of accuracies of AdaBoost, Multinomial Naïve Bayes and LDA algorithms.

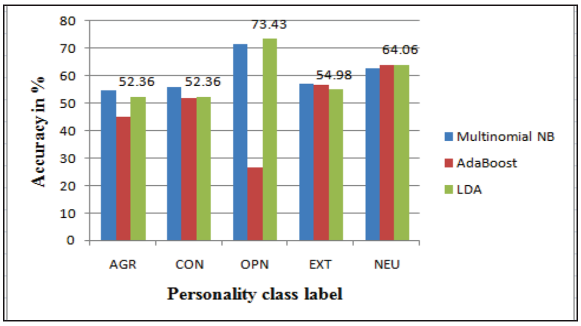


Fig. 4. Accuracies of applied algorithms

As given in Fig.4 accuracy is measured in percentage and it is plotted on Y-axis whereas five class labels of personality are plotted on X-axis. According to the results of the algorithms, it is found that Multinomial Naïve Bayes has highest accuracy of 73.43% for the feature 'OPN'. AdaBoost and LDA both have almost similar accuracies except for the feature 'OPN'. Thus from the graphical representation of accuracies it can be inferred that Multinomial Naive Bayes and LDA can be implemented further in order to increase the accuracy rates.

- Comparison of Precision

Given below is the graphical representation of precision of AdaBoost, Multinomial Naïve Bayes and LDA algorithms

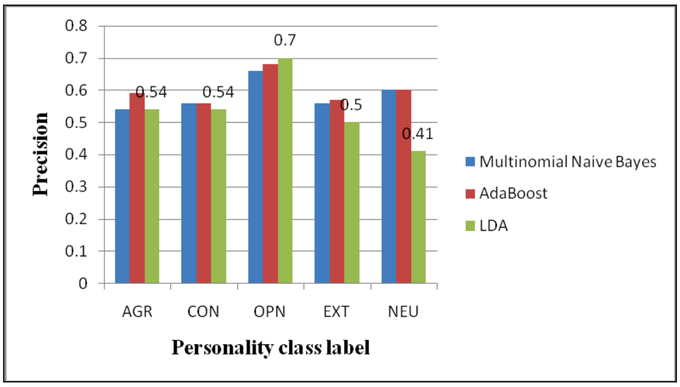


Fig. 5. Precision for Multinomial Naïve Bayes,AdaBoost and LDA

Precision is defined as fraction of appropriate instances among total instances [22]. In statistics, high precision refers to algorithm gives more accurate results than approximate results.

As shown in Fig.5, LDA has precision of 0.7 for feature 'OPN' which is highest among other categories. Multinomial Naïve Bayes has the second largest precision after LDA which is 0.69.

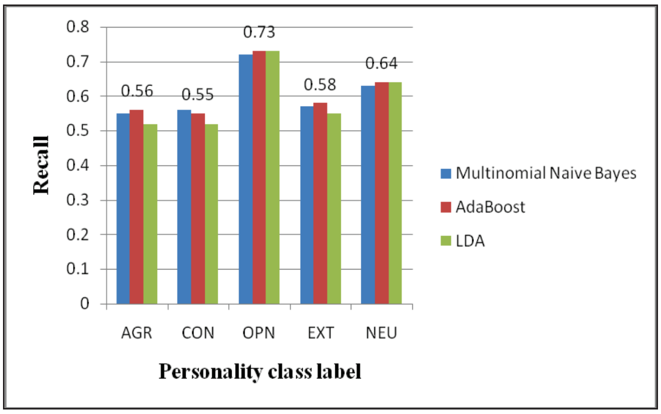


Fig. 6. Recall for Multinomial Naïve Bayes,AdaBoost and LDA

In information retrieval, recall is defined as fraction of relevant instances among retrieved instances [22].

As shown in Fig.6, AdaBoost and LDA have highest recall of 0.72 for category 'OPN'. Multinomial Naïve Bayes also gives significantly higher response 0.71 for same class. As compared to precision, all our class labels have significantly higher recall.

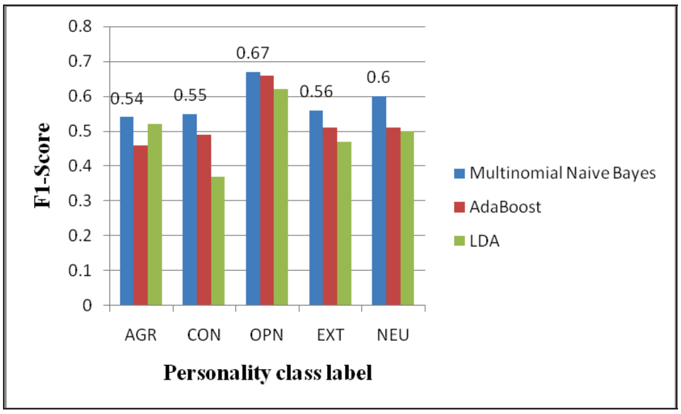


Fig. 7. F1-score for Multinomial Naïve Bayes,AdaBoost and LDA

In statistics, F1-score is used to measure accuracy of algorithm.F1-score consider both precision and recall to calculate accuracy. F1-score often used in Machine Learning for applications like information retrieval.

As plotted in Fig.7, Multinomial Naïve Bayes has highest F1-score of 0.72 for 'OPN'. Also it is clear from graphical representation that Multinomial Naïve Bayes has highest F1-score among other algorithms.

Thus from all the graphical representations it is clear that Multinomial Naïve Bayes outperforms LDA and AdaBoost in terms of accuracy, precision, recall and F1-score.

V. Future Scope

It is clear from literature overview that there has been extensive work happened in the domain of personality prediction still there are multiple areas needed to be focused. Personality is a broad domain and it comes under psychological studies.

In this work we have focused mainly on Big Five model of personality prediction which contributes to five categories of personality .There is a scope of research in combining multiple tests of personality to find most accurate class labels. Also more focus can be thrown on real-time data which can have significance with real world. Combining Machine Learning algorithms can be useful in improving accuracy. Lastly, there is a scope to work in multimodal approach of prediction in which different biomedical signals can be considered.

VI. Conclusion

Psychology is a broad domain of study and personality prediction is its integral part. Social media is a most promising platform to determine personality of a person. In the world full of competition, everyone is not able to present his/her views to the world. Hence social media like Twitter, Facebook or Instagram etc. proves most promising solution in this scenario. People express their opinion completely on such platforms without thinking whether they are right or wrong.

Hence personality prediction helps in such scenarios where it is easy to compute sentiment of a tweet or post posted by user by applying algorithms.

As discussed in our work we have applied Multinomial Naïve Bayes, AdaBoost and LDA to compare which algorithm has higher relevance. Thus, according to our results it is found that Multinomial Naïve Bayes has highest accuracy of 73.43, precision of 0.7, and recall of 0.71 and F1-score of 0.72. Future scope aims in improving accuracy of algorithm.

References

[1] Golnoosh Farnadi, Jie Tang, Martine De Cock, Marie-Francine Moens, "User Profiling through Deep Multimodal Fusion", WSDM'18, Marina Del Rey, CA, USA, February 5-9, 2018.

[2] https://openpsychometrics.org/tests/IPIP-BFFM/

[3] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio , "Personality Prediction System from Facebook Users",2nd International Conference on Computer Science and Computational Intelligence , Bali, Indonesia , 13-14 October 2017.

[4] https://www.discprofile.com/what-is-disc/overview/

[5] AnaCarolina. E.S.Lima,Leandro N.de Castro "Predicting Temperament from Twitter Data" , International Congress on Advanced Applied Informatics,2016.

[6] Louis Christy Lukito, Alva Erwin, James Purnama, and Wulan Danoekoesoemo, "Social Media User Personality Classification using Computational Linguistic", 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta,Indonesia.

[7] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html

[8] Eleanna Kafeza, Andreas Kanavos, Christos Makris, Pantelis Vikatos, "T-PICE Twitter Personality Based Influential Communities Extraction System", IEEE International Congress of Big Data ,2014

[9] Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, Michele Tomaiuolo, "A holistic system for troll detection on twitter", Computers in Human Behaviour, 27 March 2018.

[10] Jim Smith, Phil Legg, Milosmatovic, and Kristofer Kinsey, “Predicting Facebook –users’ Personality Based on Status and Linguistic Features Via Flexible Regression Analysis Techniques”, SAC’ 18 Proceedings of the 33rd Annual ACM Symposium on Applied Computing ,pp. 339-345, April 09 - 13, 2018 .

[11] M. Johnson Vioul_es, B. Moulahi,J. Az_e,S. Bringay, “Detection of suicide-related posts in Twitter data streams”, IBM Journal of Research and Development, vol. 62 , pp. 7:1 - 7:12 Jan.-Feb. 1 2018.

[12] Amir Hussain, Erik Cambria, “A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining”,IEEE Transactions on Knowledge and Data Engineering , vol. 30 , pp. 1212 – 1225, July 1 2018 .

[13] Fuhai Chen, Jinsong Su, Donglin Cao, Yue Gao, “Predicting Microblog Sentiments via Weakly Supervised Multi-Modal Deep Learning”, 2017,IEEE Transaction on Multimedia.

[14] Jose-Luis Cabra,Diego Mendez,Luis C. Trujillo, “Wide Machine Learning Algorithms Evaluation Applied to ECG Authentication and Gender Recognition”, ICBEA ’18, Amsterdam, Netherlands ,May 16–18, 2018.

[15] Danny Azucar, Davide Marengo, Michele Settanni, "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis", Personality and Individual Differences, Computers in Human Behavior, pp. 150–159, August 8, 2018.

[16] Tuan Tran , Dong Nguyeny, Anh Nguyeny, and Erik Golenz,"Sentiment Analysis of Emoji-based Reactions on Marijuana-Related Topical Posts on Facebook", IEEE International Conference on Communications (ICC), 2018.

[17] Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, Michele Tomaiuolo, “A holistic system for troll detection on twitter”, Computers in Human Behaviour, 27 March 2018.

[18] Reece Akhtara, Dave Winsboroughb,Uri Ortd,Abigail Johnsonb,Tomas Chamorro-Premuzic, “Detecting the dark side of personality using social media status updates”, Personality and Individual Differences ,pp.90–97,2018

[19] Ange Tato, Roger Nkambou, Claude Frasson, “Predicting Emotions from Multimodal Users 'Data”, UMAP ’18, Singapore, July 8–11, 2018.

[20] Suja Panickar, Aditi Kunte, “Personality Prediction using Social Media”, IEEE 5th I2CT conference, March 29, 2019.

[21] Samuel DGoslingPeter, JRentfrowWilliam, BSwannJr., “A very brief measure of the Big-Five personality Domains”, Journal of research in personality, Vol.37, Issue.6, pp.504-528, 2013.

[22] https://towardsdatascience.com/precision-vs-recall-386cf9f89488

[23] Tuan Tran, Dong Nguyen, Anh Nguyen, Erik Golen. "Sentiment Analysis of Marijuana Content via Facebook Emoji-Based Reactions", 2018 IEEE International Conference on Communications (ICC), 2018