Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020)
IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

Machine intelligence based personality prediction using social profile data

Rohit GV, K Rakesh Bharadwaj, Hemanth R, Bariti Pruthvi, and Manoj Kumar MV

Department of Information Science and Engineering,
Nitte Meenakshi Institute of Technology, Bengaluru, Karnataka, India,560064.
gvrohit6@gmail.com, rakesh1998krb1@gmail.com, hemanth04051998@gmail.com, rapperpruthvi111@gmail.com, manojmv24@gmail.com

Abstract—Social network usage is growing exponentially every day. Different information are typically exchanged via social media platforms such as Facebook. User knowledge and what they have conveyed through changes in status are useful for learning about the behavior and human personality assessment. This work aims at setting up a framework that can predict the individual's personality based on Facebook user details. In order to analyze the individual's personality, big five model is used. The aim of this research is to predict the personality of user by using the status information present in their social media profile. Based on the analysis result, the user's personality is further classified into one of the categories present in the OCEAN model. The accuracy of personality prediction achieved by using Random Forest Classifier is 64.25%. The mean squared error is achieved using random forest regressor is 5.25.

Keywords— Personality prediction, Big Five Model, Random Forest, TFIDF Vectorizer.

I. INTRODUCTION

Facebook has 1.9 Billion users and nearly 800 million people spend an hour of their day using it. These users usually share their views and ideas via updating status, posts or comments [1]. After the advent of online social networking websites like Facebook, Instagram, and people have increasingly used these to express their desires, thoughts and views without being examined publicly.

These people, who are online on social networking websites make user-generated content as a highly productive source, which can be expressed in distinct ways like text, pictures, clips, and interaction between people. This material has influenced various forms of applications; including personality assessment. The details available in online user profiles on social media is meant to reflect the users' true identities. The traits of the users' personality are predicted by obtaining elements using the content which remains accessible. Research related to this topic has been using a growing analytical approach in order to analyze the digital impressions left on social networking websites by users. In the past few years, numerous studies have started to take advantage of the vast information available on social media for automated user temperament analysis [2].

Most of the studies addressed visual characteristics derived from photographs taken from Flicker, Instagram, Quora and Facebook. While Facebook is mostly used to share pictures and video clippings more broadly, the method presented in this paper focuses on linguistic dimension of users, which is their status that are posted. The aim of this paper is to create a framework that can predict user personality accurately based on their Facebook status text.

The main objectives of this paper are:

- To fetch data from the Facebook profiles of the users using automated browser and that data is stored in MongoDB.
- Analyze their personality using the machine learning models like Random Forest Regressor and Classifier.
- To provide a summary based on their profiles using big five model.

The contents in the upcoming sections are organized as follows, section II briefs OCEAN – Big Five Model of personality, Section III discusses the most significant literature related to Personality Prediction System, Section IV briefs the framework used for personality prediction in this paper, Section V gives the methodology, results of the proposed method is presented in section IV. This paper concludes with the future direction and a brief conclusion.

II. OCEAN – BIG FIVE MODEL OF PERSONALITY

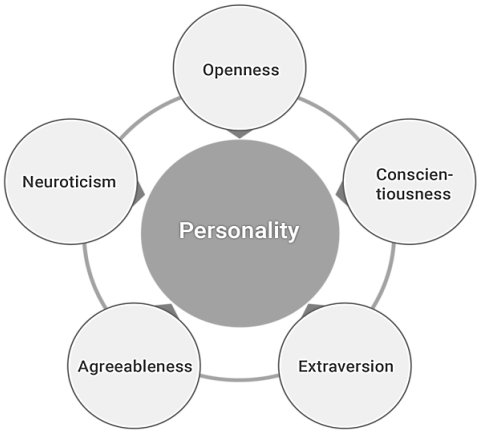

Fig 1: Big Five Model

The OCEAN model states that the personality of a person can be grouped into five traits. This is also called as psychological traits theory. These five traits are Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism.[3]

- ***Openness*** is the scale of how open a person is to try new things and is curious about everything.
- ***Conscientiousness*** is the scale of how disciplined and organized a person is in his life.
- ***Extraversion*** is the scale of how energetic and solitary a person is.
- ***Agreeableness*** is the scale of how Compassionate and challenging a person is in his life.
- ***Neuroticism*** is the scale of how sensitive minded and resilient a person in his life.

This paper adopts these five traits, based on the analysis result, a number ranging from 1 to 5 or 1 to 100 is assigned to each trait in the big 5 model. This number for each trait will signify the personality of the user.

III. Literature Survey

Numerous works have begun in the past few years to make use of the huge amount of data available on social networking sites for temperament analysis of a user. Previously similar research was done to retrieve personality details of individuals using Facebook. The model employed in the analysis is the Big Five model [4]. This study aims to incorporate multiple architectures of deep learning to see the difference through the accuracy outcome by applying a comprehensive analysis approach. The findings were successful in achieving preceding comparable testing accuracy with an overall precision of 74.17. This classification is focused on conventional machine learning and deep learning methods like gradient boosting, logistic regression, naïve bayes, and support vector machine (SVM) [6]. Another work in this area considers predicting social media personality features using text semantics [7]. This work offers a method to infer personality traits based on a semantic text assessment. Various user text depictions are proposed in combination with many semantics-based measures to analyze user's personality using their Facebook status updates. In this model, the Vector Space Model is used for automatic personality evaluation.
Another work in Facebook status-based personality prediction was done in two methods, they are Linguistic Inquiry and open vocabulary Differential Language Analysis and Word Count (LIWC) features [8]. The research was also conducted using Facebook, which described characteristics using unigrams and bag-of-words approach. Another research was carried out to build a method of personality prediction system using Machine reading comprehension (MRC) and LIWC as features on Twitter.
The scope of this paper is to study human personality and predict if a person has any kind of mental health issues. This research can be used in getting the personality of the candidate during the hiring process, and in the field of psychology to recognize mental health issues in people.

IV. Framework

The framework followed in the paper is shown in figure 1, it aims at determining the personality scores of the user by evaluating their online social fingerprints and to create a web-based framework for personality prediction.
The function of this framework involves several modules such as Web scraper, Selenium, Random Forest Classifier and Random Forest Regressor.

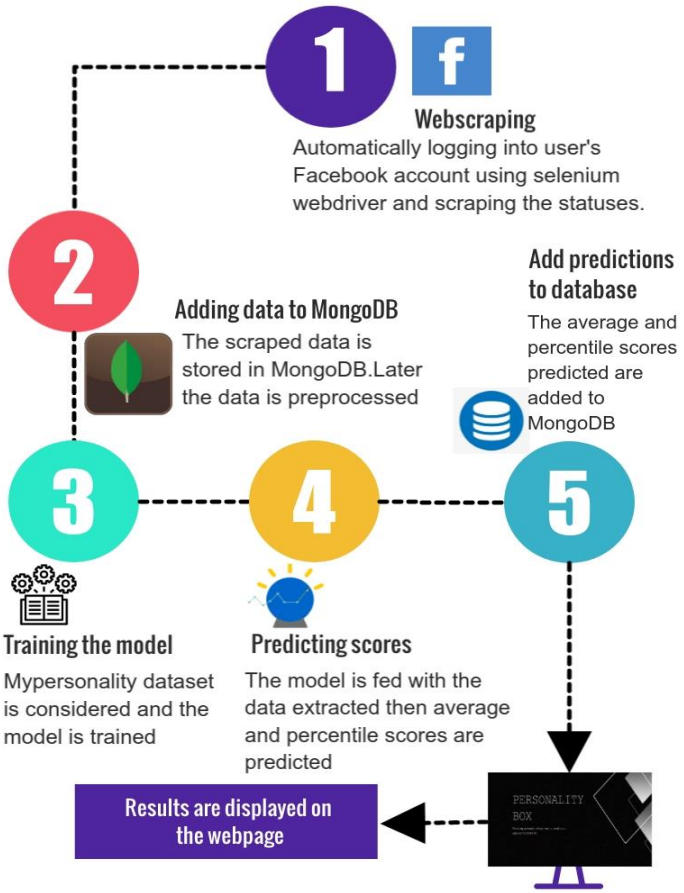

Fig 2: Methodology of proposed scheme.

A. Preprocessing

The status are automatically scraped using selenium automation tool, which in turn uses Facebook graph Application Program Interface to collect the data from user profiles. All the information which is in the form of English words undergo the pre-processing phase for further proceedings. Pre-processing phase includes eliminating links, icons, usernames, whitespaces, lowering case, trailing, and eliminating words from stop. TFIDF (Term Frequency-Inverse Document Frequency) [9] vectorizer is used in order to convert the status into vectors as well as to remove unwanted data for prediction (briefed in equation 1).

$$tfidf_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \text{ - (1)}$$

Where,
$tfidf_{i,j}$= tf-idf weight for token in document j.
$tf_{i,j}$ = Number of occurrences of token i in document j.
df_i = Number of documents that contain token i.
N =Total number of documents.

B. Ensemble Technique:

This framework employs ensemble technique for classification and prediction. In Machine Learning, ensemble methods are an approach that has been effective in improving system accuracy by integrating the predictions of various component classifiers. In addition to approaches based on the development of different classifiers from different sub-sets of data or different sub-sets of features (bagging, random forests) or methods based on the combination of poor learners (boosting), it is also possible to create meta-learners (sometimes called stacked learning) who learn to make predictions based on input features[10].

C. Training and Prediction

In this process, some popular machine learning algorithms like Random Forest Classifier and Random Forest Regressor are used.

- **Random Forest classifier:** It is a tree-based ensemble learning algorithm. It is a set of decision trees from the training subset chosen at random. It further sums up the votes of several decision trees to determine the final class of test items [11].
- **Random Forest Regressor:**
  A random forest regressor is a meta determiner which fits a few categorizing decision trees on several sub-samples of the dataset and uses averaging to improve the accuracy of prediction and over-fitting control.

During this work, many machine learning algorithms are studied to find the ideal classifier. Therefore, Random Forest algorithm is chosen as a classification algorithm based on the decision trees. Data is mapped into an n-dimensional space to construct the decision tree, where each dimension stands for one like category. Using both random forest classifier and random forest regressor prediction of the Big 5 traits i.e., OCEAN is successful. It considers Mean squared error of Random Forest Regressor, which was 5.25, and accuracy obtained was 64.24% as shown in the fig 3. The formula for calculating MSE and accuracy is given in equation 2 and 3.

$$MSE = \frac{1}{n}\sum_{i=1}(y_i - \tilde{y}_i)^2 \text{ - (2)}$$

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN} \text{ - (3)}$$

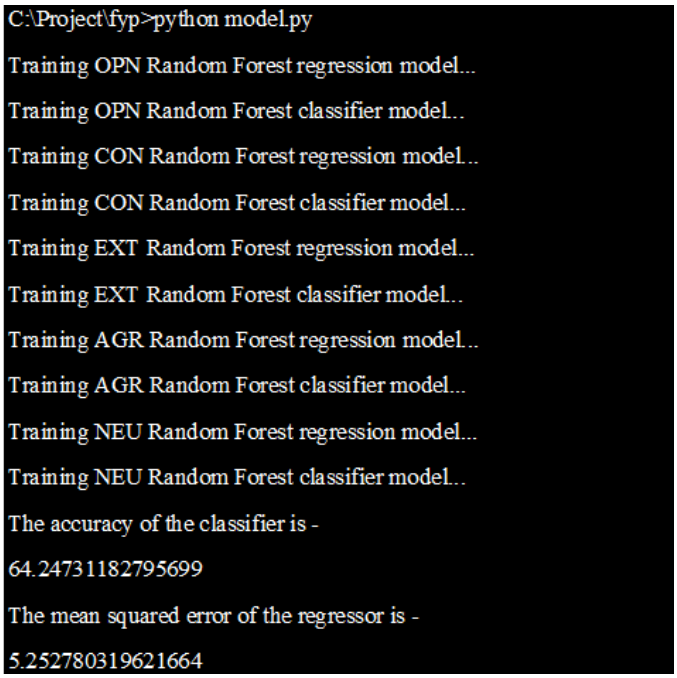

Fig 3: Accuracy and mean squared error

D. Why Random Forest

The advantages of Random forest include:

1. It can handle binary features, categorical features and numerical features.
2. It is great in handling high dimensional data and unbalanced data.
3. Prediction speed is considerably faster than the training speed.

The algorithm then attempts to discover a boundary to the decision and divides the data set into two non-overlapping partitions. This continued until a fully split remaining group could result in buckets of very different categories. In this environment of regression algorithms, prediction of user's Big 5 personality traits using various learning algorithms is done.
The Big five personality traits are Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. Shortly it is termed as OCEAN.

A. UI Details

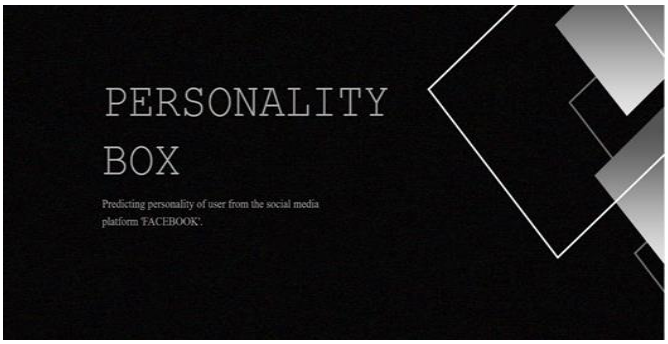

Fig 4: Screenshot of web application

Personality prediction system uses (Hyper Text Pre-Processor) PHP as its frontend structure. PHP is an open source scripting language that is commonly used for webpage creation. It uses php because it supports handling large amount of database and it is platform independent. The UI shows a single individual's personality statistics. As mentioned, after acquiring the profile details from the user, it scrapes the data from their status, store the data in a database using MongoDB. After this the linking of stored data to UI is done using Rest API7. The application will showcase the results to the user the predicted values along with their Name and Profile picture. These predicted values will be in the form of percentiles and averages.UI is developed in such a manner that it is easier to use and understand.

V. RESULTS AND DISCUSSION

This section briefs the results of the experiments carried out for predicting the personality of the user based on his social media profile. The user's profile is evaluated to obtain OCEAN values, these values represent the user's personality.

	sEXT	sNEU	sAGR	sCON	sOPN
count	9917.000000	9917.000000	9917.000000	9917.000000	9917.000000
mean	3.354760	2.609453	3.616643	3.474201	4.130386
std	0.857578	0.760248	0.682485	0.737215	0.585672
min	1.330000	1.250000	1.650000	1.450000	2.250000
25%	2.710000	2.000000	3.140000	3.000000	3.750000
50%	3.400000	2.600000	3.650000	3.400000	4.250000
75%	4.000000	3.050000	4.150000	4.000000	4.550000
max	5.000000	4.750000	5.000000	5.000000	5.000000

Table 1: Dataset summary statistics

The IDE (Integrated Development Environment) used is the visual studio. In this analysis Selenium is used for web scraping purposes. The language used for coding is python. MongoDB is the database that is used to store the status collected by scraping user profiles.

A random forest regressor and a random forest classifier are the models used. The models used are trained on a dataset taken from my Personality dataset. In this study PHP is used along with the CSS to build the web application.

This work uses 'MyPersonality' dataset which is a famous Facebook app created in 2007 which enabled people to take some real psychometric evaluation tests and check the results immediately. As well as test data, about 40% of respondents also chose to share information from their Facebook profile, resulting in one of the largest repositories in history for social science research. The application was active until 2012, and during this time it collected data from more than 6 million volunteers. Respondents were coming from different age groups, backgrounds and cultures.

They were extremely motivated to respond genuinely and cautiously, because the sole reward they are going to get is feedback on their results 'MyPersonality' dataset used for this study contained about 5000 positive and 5000 negative status related to each characteristic in the Big five model [2]. Training and testing of the proposed model is carried out on my personality dataset. Table 1 is the sample data representing the key characteristics (summary statistics) of the dataset.

The Big 5 model was derived based on common descriptive adjectives through factor analysis of the questions.This research brought forth five distinct personality traits.

Table 1 gives statistical information related to the dataset like mean, standard deviation etc. In this picture scores of openness, neuroticism, agreeableness, and conscientiousness are denoted by sEXT, sNEU, sAGR, sCON, sOPN respectively. Another dataset that is the second one used is the status updated by the Facebook users which are gathered manually. This is btained by scraping status of users on the facebook website.The status obtained from the social networking sites are used to analyze the characteristics of the user considering the scores obtained by using the model.

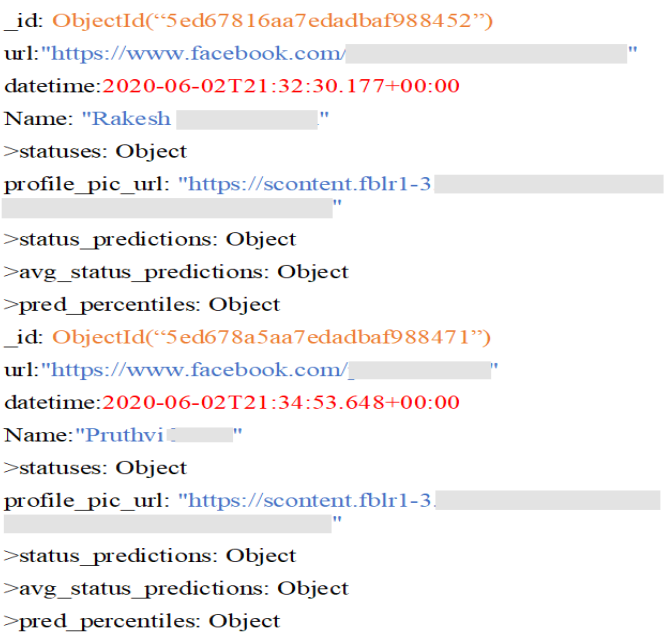

Fig 5: Database in MongoDB.

Figure 5 shows the Mongo-Database where the scraped status of the user are stored for pre-processing where unwanted data is removed and at the end the predicted results are stored.

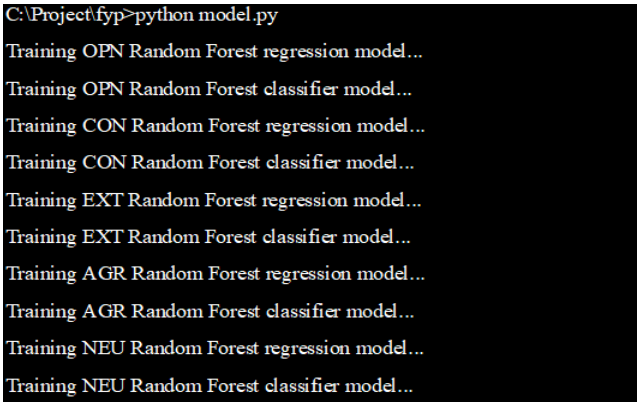

Fig 6: Training of models

After the preprocessing stage, using random forest regression and random forest classification the prediction models are built and fit transform method is used to train the model. Here for training purpose, the dataset is divided into two parts i.e., Training set and test set. The training set is used in training process, test set is the status which are extracted. About 85% records from the dataset are taken as training set, and the rest as test set.

```
C:\Project\fyp>python predict.py
Inserting average personality score of :Rakesh Bharadwaj K
Inserting average personality score of :Pruthvi Bariti
Inserting average personality score of :Hemanth Ramesh
Inserting average personality score of :Rohit Gv
Calculating percentiles for the following person: Rakesh Bharadwaj K
Calculating percentiles for the following person: Pruthvi Bariti
Calculating percentiles for the following person: Hemanth Ramesh
Calculating percentiles for the following person: Rohit Gv
```

Fig 7: Predicting the personality of user on the scale of 1 to 5 using random forest regressor and classifier.

Name	O	C	E	A	N
Hemanth	4.12	3.48	3.33	3.42	2.7
Pruthvi	4.18	3.5	3.33	3.49	2.64
Rakesh	4.08	3.58	3.43	3.54	2.62
Rohit	4.1	3.47	3.45	3.48	2.65

Table 2: predicted scores based on the status scraped.

The given values in table are predicted personality values of the user from a scale of 1-5, 1 being low and 5 being high. Each column in the table represents one of the Big 5 Traits.
For each status personality traits (OCEAN) are calculated using the test set and these scores are then being stored onto the MongoDB. Calculated values for each scraped status is shown in table 2. It summarizes the calculated values of OCEAN for five user's profiles. The process of calculating the personality values – i.e., predicting the personality using the built model is summarized on given test profiles is shown in table 2.

Name	O	C	E	A	N
Hemanth	40.69	41.67	35.29	45.1	58.82
Pruthvi	60.78	48.04	36.27	55.88	44.61
Rakesh	37.75	72.55	62.25	59.8	39.22
Rohit	37.75	39.71	63.73	55.39	47.06

Table 3: The percentile OCEAN values
Table 3 gives the percentile normalized predicted values (out of 100) of personality of the user, 1 being low and 100 being high. Each column in the table represents one of the Big 5 Traits
Screenshot of the web application developed for validating the method of personality prediction is given in figure 6. This screenshot refers to the UI which we developed using PHP, HTML, and CSS in order to show the scores predicted to the user.

VI. FUTURE POSSIBLE RESEARCH PROBLEMS

Some of the possible research direction in the realm of user personality prediction using the user's social profile data rare security, fake data, and language barrier.
Securing the user's data, i.e., scrapped user's status and the predicted results is challenging research problem, further, the other challenge is to recognize the fake user's profile and fake user status is highly motivating problem and challenge to address. The ultimate problem in predicting the user's personality is to address the language barrier. Social media users belonging to different geo locations and speaking various languages and related dialect pose a greater challenge while predicting the user's personality.

VII. CONCLUSION

User content generated on social networks provides an efficient way for predicting the traits of a user. The paper suggests an approach using Traditional Machine Learning models and Ensemble Learning Method. We focused on the feasibility of modelling users Big five traits based on the data extracted from Facebook profiles of users. Further improvements can be made to this prediction's by using further accurate data collection to increase the accuracy. This allows a relatively accurate evaluation of personality characteristics of an individual and can be used in a wide variety of applications. The insights we gained while doing our research can now be combined with the findings of other researchers to create a learning ensemble that can predict a very accurate approximation of a social media user's personality.

VIII. REFERENCES

[1] F. Celli, E. Bruni, and B. Lepri. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In Proc. ACM Multimedia -MM '14', pages 1101--04. ACM, 2014.

[2] Personality Prediction System from Facebook Users Author links open overlay panel TommyTanderaHendro, Derwin Suhartono ,RiniWongso, Yen Lina, Prasetio ,Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9 Kemanggisan, Jakarta 11480, Indonesia.

[3] R. McCrae and O. P. John. An Introduction to the Five-Factor Model and its Applications. J. Pers., 60(2):175 -- 215, 1992

[4] Soto, Christopher J., Anna Kronauer, and Josephine K. Liang. "Five-Factor Model of Personality." *The Encyclopedia of Adulthood and Aging* (2015): 1-5.

[5] Wijaya A, Febrianto N, Prasetia I, Suhartono D. Sistem Prediksi Kepribadian "The Big Five Traits" Dari Data Twitter. Jakarta: Bina Nusantara University, School of Computer Science; 2016.

[6] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining facebook data for predictive personality modeling," in Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA, 2013.

[7] Hassanein, M., Hussein, W., Rady, S. and Gharib, T.F., 2018, December. Predicting personality traits from social media using text semantics. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)* (pp. 184-189). IEEE.

[8] Sumner, Chris, et al. "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets." *2012 11th international conference on machine learning and applications*. Vol. 2. IEEE, 2012.

[9] Qaiser, Shahzad, and Ramsha Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents." *International Journal of Computer Applications* 181.1 (2018): 25-29.

[10] Kunte, Aditi, and Suja Panicker. "Personality Prediction of Social Network Users Using Ensemble and XGBoost." *Progress in Computing, Analytics and Networking*. Springer, Singapore, 2020. 133-140.

[11] Aydin, Berkay, et al. "Automatic personality prediction from audiovisual data using random forest regression." *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016.

[12] Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. In National Academy of Sciences; 2015. p. 1036-1040.

[13] Ryan, T., & Xenos, S. (2011). Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. Computers in Human Behavior, 27(5), 1658–1664.

[14] Kosinski M, Matz S, Gosling S, Popov V, Stillwell D. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist. 2015 Feb; 70(6): p. 543..

[15] Singh, Bhawna and Singhal, Swasti, Automated Personality Classification Using Data Mining Techniques (May 16, 2020). Available at SSRN: https://ssrn.com/abstract=3602540.

[16] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, *6*, 61959-61969.

[17] Bai, Shuotian, Tingshao Zhu, and Li Cheng. "Big-five personality prediction based on user behaviors at social network sites." *arXiv preprint arXiv:1204.4809* (2012).

[18] Ortigosa, Alvaro, Rosa M. Carro, and José Ignacio Quiroga. "Predicting user personality by mining social interactions in Facebook." *Journal of computer and System Sciences* 80.1 (2014): 57-71.

[19] Imran, Azhar, Muhammad Faiyaz, and Faheem Akhtar. "An enhanced approach for quantitative prediction of personality in facebook posts." *International Journal of Education and Management Engineering (IJEME)* 8.2 (2018): 8-19

[20] Mugunthan, S. R. (2019),"Security And Privacy Preserving Of Sensor Data Localization Based On Internet Of Things",Journal of ISMAC, 1(02) 81-9.

[21] Pandian, A. P. (2019),"Artificial Intelligence Application In Smart Warehousing Environment For Automated Logistics", Journal of Artificial Intelligence, 1(02), 63-72.