# Personality Recognition using Social Media Data

**Medhavini Rao[1], Pooja Jayant Kanchugar[2], Pooja R[3], Prakshitha M N[4], Anitha R[5]**

*[1,2,3,4]Eight Semester, Dept. of CSE, The National Institute of Engineering, Mysore*
*[5]Professor, Dept. of CSE, The National Institute of Engineering, Mysore*

---

**Abstract -** *With the emerging technology, social media has also increased to a great extent. Social media is one of the platform for opinions and thoughts of an individual. People share their personal information on social media account. Social media mainly predict the personality of a person. Personality mainly accounts for individual differences in people. To reliably, validly, and efficiently recognize an individual's personality is a worthwhile goal; however, the traditional ways of personality assessment through self-report inventories or interviews conducted by psychologists are costly and not so useful and less commercial. This project proposes a method of Big Five personality recognition from Facebook data. It helps in the analysis of Facebook data and train the model using large datasets. Random Forest Regression is used to build the model for getting high accuracy. Firstly, we collected raw data by accessing Facebook directly through chrome extension, then the friend list will be accessed to review their status and posts. In the further step, the status and posts of each one in the friend list will be stored in MongoDB database as a dictionary. Then the data will be fitted to our model to train the model and then the model is tested for some more data. At last we visualize the results by plotting graphs.*

**Key Words: Machine learning, Facebook, Personality Recognition, Big Five, Random Forest algorithm.**

## 1. INTRODUCTION

Online social networks nowadays has become popular in many ways for users to show their identity and connect and share information with one other through social networks. Facebook is the most popular social network. Personality describes the uniqueness of a person. One can understand others behavior on the basis of day-to-day observation of an individual's underlying personality traits. Many researches in psychology suggested that the behavior and preferences can be explained to a great extent by underlying psychological constructs i.e. personality traits. Facebook profiles of users is the most important source of information. The study of personality can benefit Social Network analysis, Deception detection, Authorship attribution, Sentiment analysis. Personal information on social networking platforms render it possible to analyze the user's behaviors and infer their personality traits on the web. It has become one of the most ubiquitous means of communication today.

The most widely accepted model of personality is Big Five or five factor model mainly consist of Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. Openness is nothing but the measure of creativity, curiosity, imagination etc. Conscientiousness is how efficient, organized, responsible, and persevering a person is. Extroversion is to seek stimulation in external world and to express positive emotions. Agreeableness is maintaining positive social relations, friendly and cooperative. Neuroticism is tendency to experience mood swings and emotion such as guilt, anger, depression. We basically use big five approach because it has better performance and undergoes detailed analysis of an individual.

The technique used in our proposed system is the machine learning approach based on social media activities. The earliest efforts mainly focused on Personality Recognition from written text which may be the handwriting of that person. The first pioneering work was done by Shlomo about 10 years ago , which extracted word categories and relative frequency of function words from 2236 written essays of 1200 students as the input of Support Vector Machines (SVM) to discriminate between students at the opposite extremes of Extraversion and Neuroticism. The same data set and approach were used by Mairesse to recognize individuals in the upper and lower half of the observed scores for all Big Five traits. They studied the effectiveness of different sets of textual features extracted from psychologically oriented text analysis tools (e.g. LIWC[2]) or psycholinguistic dictionary (e.g. MRC[3]), and found that the openness trait yields the best accuracy using SVM.

In a separate study by Lee and Nass, interaction effects between user factors, and media factors on feelings of social presence were investigated.

In a study by Saati it was found that extroverts tended to interact faster with the user interface than introverts. T. Correa concludes that extraversion and openness to experience are positively related to use of social applications on internet, emotional stability was negatively associated.

## 2. OBJECTIVES

The main objective of this entire system is to analyze the personality of an individual based on social media activities. The study of personality is of central importance in psychology, and personality recognition also benefits many other applications such as Social network analysis, Recommendation systems, Deception detection, Authorship attribution, Sentiment analysis/opinion mining and so on. With the development of various social media, modern computer science has become the real potential for advancing that endeavor. The rich digital traces and self-disclosed personal information on social networking platforms render it possible to analyze the user's behaviors and infer their personality traits on the web, which has attracted much attention from different disciplines.

Many data scientists use many different kinds of machine learning algorithms to discover patterns in big data. There are two groups to make predictions: supervised and unsupervised learning. Supervised learning is where you have input variables 'a' and an output variable 'b' and you use an algorithm to learn the mapping function from the input to the output b = f(a) . Unsupervised learning is where you only have input data 'a' and no corresponding output variables. Supervised are grouped into regression and classification. Regression is nothing but when the output variable is a real or continuous value such as "salary" or "weight". Examples are Predicting age of a person etc. Classification is nothing but when the output variable is a category such as "red" or "blue" or "disease" and "no disease". Examples are Predicting house price based on area.

However in our project we have used Random forest algorithm for getting high accuracy. Random forest can used for both classification and the regression kind of problems. Random forest algorithm is a supervised classification algorithm. As the number of trees in the forest increases, accuracy also increases. Random Forest is one of the easy-to-use algorithm, because its default hyper parameters often produce a good prediction result. The number of hyper parameters is also not that high and they are easy to understand. Random forest algorithm mainly used in Banking, Medicine, Stock Market, E-commerce.

## 3. EXISTING SYSTEM

Personality is considered as an important criteria in explaining human behaviors. Recognizing one's personality has many advantages. It can be used in Deception detection, Recommendation systems and Sentimental analysis. The existing methods consists of various interviews or counseling by experts which is not economical. There are various other methods which can be used for personality recognition. Using Social media data, personality can be recognized as a system implemented in [4]. As proposed in, personality can be recognized using data of social media like Facebook, Twitter, Instagram etc.

In a China, a blog called Sina Weibo is used. The main drawback of existing system is scalability. This system has considered 113 features and 994 entries using Big Five personality (OCEAN model).

### 3.1 Drawbacks of existing systems

- The traditional methods like counseling and personal interviews by experts are expensive. The number of data entries used is less and it is static. It is developed on a Chinese blog which many people aren't aware of and less number of people are using it.

## 4. PROPOSED SYSTEM

The proposed system proposes a methodology of Big five personality recognition from Facebook environments as described in [3] with machine learning approach of supervised learning called Random forest. Nowadays the interest of scientific fields is shifting towards personality recognition that helps in automated analysis of a person during the hiring process, providing the shopping recommendations for online customers and so on .The system fetches the Facebook status text of the friend list and performs regression methodology on it so that it can get the percentile report of each OCEAN model trait of each individual in the friends list .This is visualized using a radar plot that gives a clear understanding of the personality of that person.

### 4.1 Advantages

- Many a times, the manual interview process for hiring process may not provide a clear understanding of his personality and also since most of the people today hold a social media account we can perform their five trait prediction and hire them based on the requirement. This system can also be implemented for the online shopping recommendation systems that help in determining the personality of a person and then recommending them with the suitable products based on their interests and preferences. This can also be applied to matrimonial sites for getting the personality of an individual's interest. Apart from this we can get the OCEAN traits report in the form of radar plots that helps in analyzing ourselves.

## 5. SYSTEM DESIGN

The diagram in Fig-1 shows the architectural system view of our proposed system that describes how the entire system works. Initially, we train our model with data that is an output file from a Facebook personality test application [2] that gives around 10,000 status texts with their corresponding trait values. Data that we require may not contain the actual value or may contain incorrect values. So the cleaning is required for such dataset. There can be redundant data as well as noisy data mixed with the dataset. This is cleaned manually by filling average value of the whole column after that we need to prepare data for training.

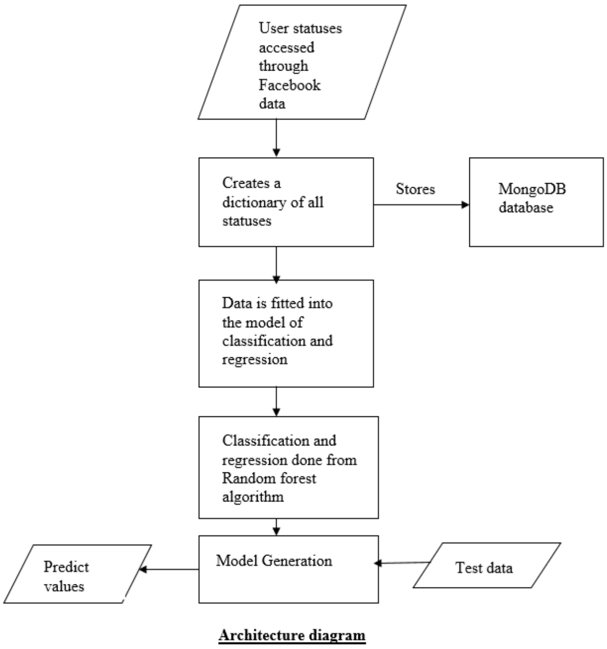
User statuses accessed through Facebook data
Creates a dictionary of all statuses
Stores
MongoDB database
Data is fitted into the model of classification and regression
Classification and regression done from Random forest algorithm
Model Generation
Predict values
Test data
Architecture diagram

**Fig-1 :** Architecture diagram

## 5.1 Dataset

The reason we have exclusively chosen Facebook because it gives concrete huge data set and almost everyone these days have a Facebook account. And also increased social media activities these days may bring both beneficial and harmful effects that needs to be carefully handled [1]. These are the reasons we have chosen Facebook as our dataset.

Data pre-processing involves renaming, rescaling, binning and so on. In our project, the raw data contains lot of irrelevant words which is not useful for analysis, such words are removed using TfidVectorizer [9](E.g.: is, for, the etc.). Only the features which are required are selected by omitting unwanted columns from the original dataset.

After this we train the model with the data which is then used to predict values. Once training is done now we need to collect more data for testing. For that we are going to access Facebook through chrome driver. We manually give email id and password credentials required to login to Facebook. Once we run our system our Facebook account is logged in and all the friend's status texts will be parsed one by one. From this we create friends dictionary containing all the status updates along with its date and time of post. This data is stored in MongoDB database.

Using the trained data, the model creates pickle files of all five traits that is helpful in predicting the values for the testing data through regression. The main algorithm used here is Random forest algorithm which is a supervised machine learning approach which can perform both regression and classification .It constructs decision trees and decides the final output based on output of each decision trees The input data is taken from Facebook and is fitted into the model . Then we predict the corresponding regression values for the status text through our model .All the five traits will be given an average regression value and is stored in the database. These values are finally converted to percentile values as we can easily visualize all the traits clearly.

And using these five trait's percentile values finally we create a radar plot that clearly visualizes the percentile of all traits. Hence our system finally outputs the radar plot report of each of our fiend that helps in predicting their personality. The radar plot which is obtained as output for one of the person is shown in fig-2.

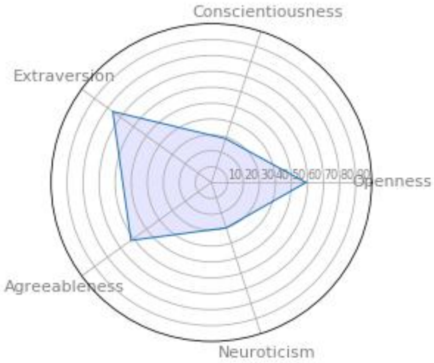

**Fig-2** Radar plot showing five personality trait percentile

## 6. CONCLUSION

A system like this clearly visualizes the personality of each one of the friend on Facebook with radar plots which is helpful in preventing many anti-social activities. Apart from this system can be implemented as a sub module in online shopping product recommendation systems, matrimonial sites and also in hiring candidates by recruiters.

## 7. FUTURE ENHANCEMENT

We can further extend our implementation of personality recognition to determine traits related to prepare for future care needs of older adults as described in [8].
Using Personal Traits for Brand Preference Prediction which is an important topic in marketing. This can also be a future enhancement of our system as described in [9].

## 8. REFERENCES

[1] Ashish K. Tripathi "Personality Prediction with Social Behavior by Analyzing Social Media Data" Survey Department of Applied Computer Science and Society University of Winnipeg 2008.

[2] Dejan Markovikj, Sonja Gievska, Michal Kosinski and David Stillwell "Mining Facebook Data for Predictive Personality Modeling" published in Association for the Advancement of Artificial Intelligence in 2013.

[3] Jianguo Yu and Konstantin Markov "Deep learning based personality recognition from Facebook status updates" Published in IEEE 8th International Conference on Awareness Science and Technology (iCAST) 2017.

[4] Marie-Francine Moens, Martine De Cock, Golnoosh Farnadi and Susana Zoghbi "Recognizing personality traits using Facebook status updates" published IEEE 8th International Conference on Awareness Science and technology 2017.

[5] Alireza Souri, Shafigheh Hosseinpour and Amir Mosoud Rahmani "Personality classification based on profiles of social networks' users and the five-factor model of personality" published in Human Centric Computing and Information Science 2018.

[6] A. G. C. Wright, ''Current directions in personality science and the potential for advances through computing,'' *IEEE Trans. Affect. Comput.* vol. 5, no. 3, pp. 292–296, Jul. 2014.

[7] J. Golbeck, C. Robles, and K. Turner, ''Predicting personality with social media,'' in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, Vancouver, BC, Canada, 2011, pp. 253–262.

[8] Silvia Sorensen and Paul R duberstin "How Are Personality Traits Related to Preparation for Future Care Needs in Older Adults?" Published in J Gerontol B Psychol Sci Soc Sci 2008.

[9] Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan "Using Personal Traits for Brand Preference Prediction" published in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing in 201