Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

# *A Study of Data Mining Methods for Prediction of Personality Traits*

Helly. N. Desai
Computer Engineering Department
Sarvajanik College of Engineering and Technology
Surat, India
hellydesai2010@gmail.com

Prof. Rakesh Patel
Computer Engineering Department
Sarvajanik College of Engineering and Technology
Surat, India
rakeshpatel.ce@gmail.com

**ABSTRACT**

Personality is an important psychological feature in humans. Personality reflects what human is, what human likes and how humans behave in certain situations. Nowadays usage of social media has increased tremendously. People update their thoughts, opinions and activities on social media which can be analysed to discover the personality of the social media user. The goal of this paper is to survey the prediction of personality traits more accurately by combining user's demographics features, likes and activities. Facebook is a widely used social networking site among all the other social networking sites. In this paper, various supervised, unsupervised and semi-supervised data mining techniques for personality traits prediction based on different features are discussed.

***Keyword: Big Five, Digital footprints, OCEAN, data mining***

## 1. INTRODUCTION

Personality is a human construct that explains human behaviour in different situations. It reflects human behaviour and important aspects like happiness, emotions and sadness. Personality affects people's life choices. During the social interaction, an individual interacts with the unknown. In this situation, the personality of that individual plays a vital role. Automatic identification of personality traits can be helpful in various ways like an online book and job recommendation system, personalized search engine, mental health diagnosis, etc.

The main purpose of this paper is to survey techniques that have been used for identifying the personality traits task.

The personality of a user can be predicted using the Big Five Model or Five-Factor Model, which consists of

1. Openness to Experience: People who are excited to try and learn new experience
2. Conscientiousness: People who are disciplined and want things in a certain way
3. Extraversion: People who are extrovert and easily socialize
4. Agreeableness: People who are kind and helpful
5. Neuroticism: People who tend to have a negative opinion

Social media have a huge impact on people's lives. The usage of social networking sites is increasing tremendously. Many people use social media for their day to day life. They post statuses, post pictures and videos, share content and their opinions. By judging their online aspects, the personality of users can be predicted. Facebook is the most popular SNS (Social Networking Site) among all. Over 241 million people use Facebook in India. That is a very huge number of users. This makes identifying the personality traits an easy task. Researchers have a wide range of data on Facebook to analyse. Facebook has various users' features called Digital Footprints. Digital Footprint includes the user's demographics, activity, likes, pictures, etc. Demographic features include the user's name, age, gender, hometown, posts, etc. Activity features include events that have been created by the user, events attended by the user, links shared by the user, check-ins by the user. Lastly likes means posts that the user has liked.

### Personality Identification Using Facebook

According to [1], The relation between features of Facebook and The Big Five traits known as OCEAN are,

**Openness to Experience:** People high in Openness to Experience are willing to try new communication methods. People tend to have a wide variety of interests on Facebook. They are willing to try new

things than old ones. They use new features of Facebook out of curiosity.

**Conscientiousness:** People high in Conscientiousness are very organized, hard-working and honest. They spend less time on messaging tools. They tend to have limited Facebook activities.

groups. They tend to use Facebook to state their opinions, to share information.

**Agreeableness:** People who are high in Agreeableness, they spend less time online. They tend to have a greater number of Facebook friends.

**Neuroticism**: People of high neuroticism tend to overthink about what information is shared. "Wall" is

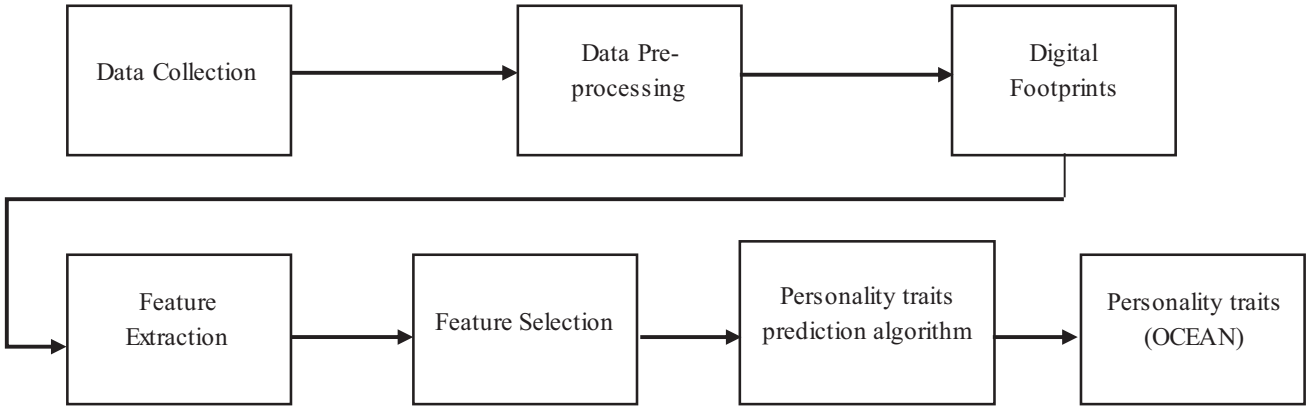


Figure 1:A general personality prediction framework

their favourite Facebook feature. They spend a great amount of time on Facebook to avoid real-life problems.

indicates the general personality prediction framework.

**Extraversion:** People who score high in Extraversion tend to have membership in various

**The general architecture of Personality prediction**

A general personality prediction framework, shown in Fig 1, consists of following phases like:

**Data collection:** In this phase, data is collected from the Facebook graph API. An API dataset that is publicly available for research. These data are in the form of demographics, activities and likes.

**Data Pre-processing:** In the data pre-processing phase, extracted data are pre-processed. hashtags, URLs, mentions, phases are ignored. Text unification and tokenization are performed on data. After pre-processing, important sentences are taken into consideration for feature extraction.

**Feature Extraction:** In the third phase, features are extracted from sentences. Features like linguistic features are semantic features that are helpful.

**Feature Selection:** In this phase, features that help predict personality traits are selected. It further goes for the training phase.

**Personality traits prediction:** Researchers used different algorithms and feature set for the personality prediction task. The output of personality is the big five traits known as OCEAN. Fig 1

The rest of the paper is organized as follows: Section II discusses the work related to predicting personality traits. Section III discusses machine learning-based personality prediction techniques. In section IV, personality traits prediction based on important attributes such as approaches used, algorithms is discussed. Section V concludes the paper.

## 2. RELATED WORK

For personality traits prediction, various machine learning and deep learning methods are used. The correct selection of digital footprints helps predict personality traits easily. Personality traits depend on a feature like demographics, activities and likes. It is shown in Fig 2. Features like demographics, activity and likes reflect users' behaviour on the social media platform. By using that the personality traits can be predicted. Demographic features include users' id, first_name, last_name, gender, age, hometown. Demographic features reveal basic information about the user [10]. The activity includes events created by the user, events attended by the user, check-ins, posts, no. of friends, no. of pages a user follows. The activity feature reveals users' interest in fields [10]. Users' preferences on different domains like a job

interview, political domain, movies, sports etc. Likes denote the posts that users have liked.

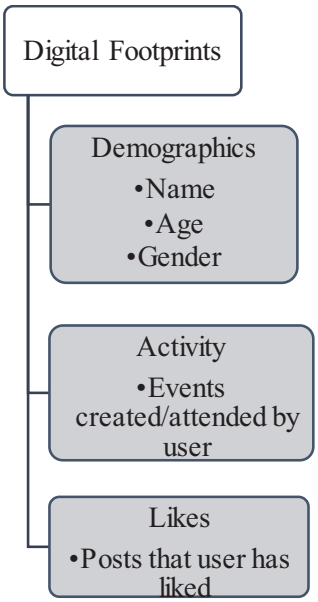


Figure 2: User's Digital Footprints of Facebook platform.

## 3. MACHINE LEARNING BASED TECHNIQUES

Machine learning is an ability of that model that can learn from experience. In machine learning, the model learns itself so no explicit programming is needed. The model learns from data and patterns in that data. Various techniques have been proposed in past to identify personality traits on types of data like labeled data (for example, supervised learning), unlabelled data (for example, unsupervised learning), and partially labeled data (for example, semi-supervised learning). These are described below.

**Supervised learning**

Supervised learning is the task when a model is getting trained on a labeled dataset. In this type of learning, training and validation dataset has labels. By learning, the model develops logic on its own. The various algorithms generate a function that maps inputs to desired outputs. In supervised learning algorithms, the classes are pre-determined. These classes are defined by humans and created infinite sets. The task of the algorithm is to classify data into given classes [13]. Figure 3 shows a flow diagram of supervised learning.

Many authors have used supervised learning for personality traits prediction. The author in [2] has used supervised learning in their analysis for predicting personality traits. Firstly, the data is gathered from the Facebook Graph API. Before classifying personality traits, several pre-processing methods are utilized. It involves the removal of URLs, hashtags, tags, emojis, etc. They proposed an algorithm to filter out the outliers present in data. For that, they have used z-score value and Mahalanobis distance measure. Z-score value helps to identify outliers. Thereafter outliers are discarded. After outlier elimination, different classification algorithms are applied like KNN and SVM. It is examined that KNN gives a better outcome than SVM. The author can use more machine learning algorithms for personality traits prediction.

The author in [5] has introduced two deep learning structure for personality traits prediction. Multi-modal deep learning techniques use word-embedding and acoustic prosodic features. The authors used features from LLD (Low-Level Descriptor), LIWC (Linguistic Inquiry and Word Count), DAL (Dictionary of Affect in Language). After that, there are two approaches. The first MLP (Multi-Layer Perceptron) approach merges all features set into a single input vector. Then input vector is given as an input to DNN (Deep Neural Network). The second MLP approach feeds each feature sets separately to

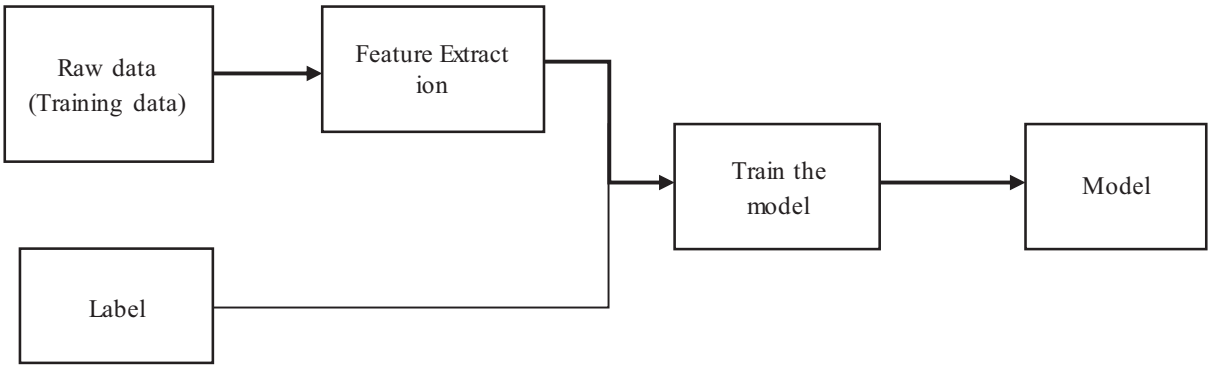


Figure 3:Flow diagram of supervised learning

DNN. The results explained that among all ML methods LDA performs best. The dataset used in this paper can be expanded. It should contain more English words. The other deep learning model can be used for the process.

**Unsupervised Learning**

Unsupervised learning uses unlabelled data to find a hidden pattern from the dataset. In this type of model training, there is no validation parameter to find potential output. In unsupervised learning, the machine simply receives a set of inputs. The machine does not obtain any labeled data or targeted output or any rewards. The goal of unsupervised learning is to develop a framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be seen as finding patterns in the data above and beyond what would be considered pure unstructured noise [14].

Results explained that SNA with XGB (eXtreme Gradient Boosting) performs better than all three combined together.

**Semi-supervised learning**

Semi-supervised learning is somewhere between unsupervised and supervised learning. In fact, most semi-supervised learning strategies are based on extending either unsupervised or supervised learning to include additional information typical of the other learning paradigm. The study of semi-supervised learning is influenced by two factors: its practical value in making better computer algorithms, and its theoretical value in understanding learning in machines [17].

The author in [4] has proposed a new structure called AttRCNN. AttRCNN is short of Attention Recurrent-Convolutional Neural Network. The process uses unsupervised learning for data training and testing. After data collection, in data pre-processing text tokenization and text unification are performed. After the pre-processing phase, two strategies are applied,

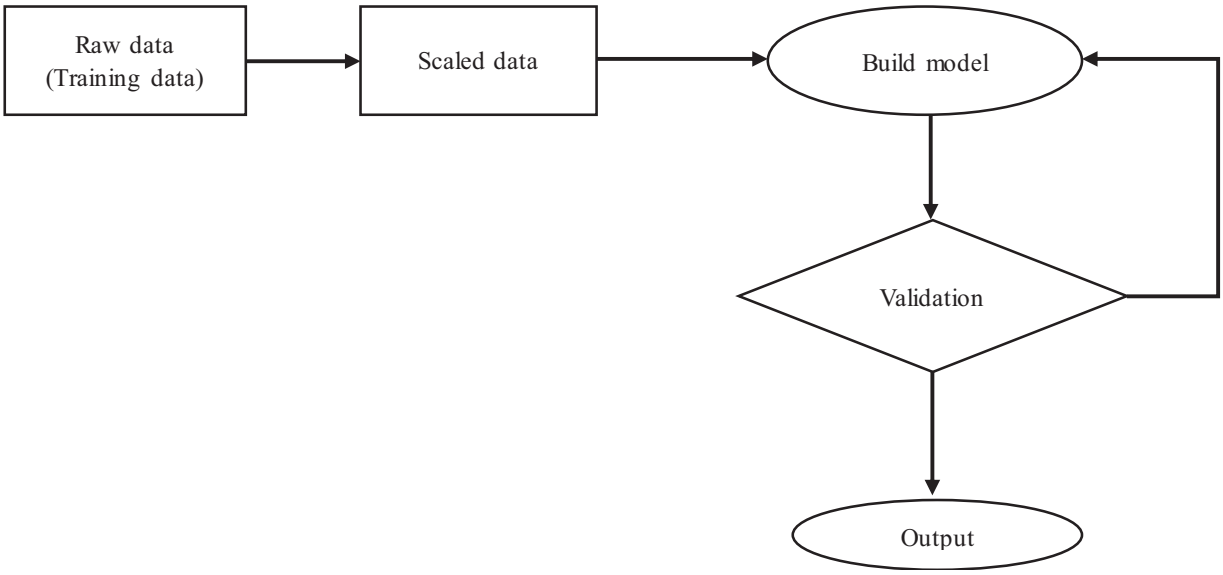


Figure 4:Flow diagram of unsupervised learning

The authors in [6] used unsupervised learning for personality traits prediction. The method predicts personality based on users' social behaviour and their language usage habits on the Facebook platform. For that LIWC (Linguistic Inquiry and Word Count), SNA (Social Network Analysis) and SPLICE (Structured Programming for Linguistic Cue Extraction). In data pre-processing; removal of URLs, symbols, names, spaces, and lower cases. Unsupervised learning on word vector and supervised learning on deep semantic features. They proposed AttRCNN structure in which sentences are given as an input. The sentence Vectorization technique is performed by the model. Thereafter, the model gives the output as personality scores. The limitation of this technique was the author of this paper have only used text posts and no other social media features.

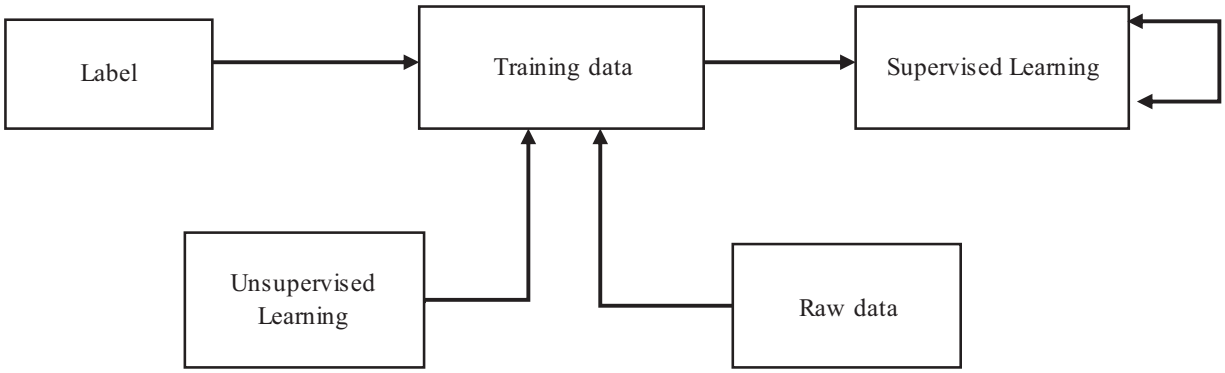


Figure 5:Flow diagram of Semi-supervised learning

The author in [3] has used a semi-supervised learning approach for personality traits prediction. For model training, co-training [13] strategy is used. PMC (Pseudo Multi-view Co-training) is an effective semi-supervised learning algorithm that is used for model training. Both labelled and unlabelled data trained parallel. The output was in the form of a personality score. Both classifiers train on different views. The linguistic features of the data support to

find the correlation between LIWC tags and personality traits. The linguistic features were extracted from LIWC and n-gram.

## 4. ANALYSIS OF EXISTING PERSONALITY TRAITS PREDICTION TECHNIQUES

Referring after various papers, it is concluded that personality traits prediction can be done using supervised, unsupervised and semi-supervised learning. In the table, it is summarized the approaches and the algorithm used for personality traits prediction.

Table 1 shows all the comparison between various techniques used in personality traits prediction.

## 5. DISCUSSION

With the analysis of different techniques that are used for personality traits prediction, it is shown that using only the user's post is not reliable. Different types of Digital Footprints can be used for personality traits prediction.

The personality traits can be correlated with a user's online posts. For example, people with high openness score tend to use positive words such as "nice", "perfect" etc. People with high Neuroticism trait prefers to stay online longer than normal.

With the data acquired from the dataset and the algorithm used, it is clear that all the algorithm works in a different way.

It is clear that to use different machine learning algorithms several aspects are needed to be considered. The first aspect is a dataset. Predicting user's personality from a social media platform is a very long process. For that, the dataset must be large. Secondly, the features that are considered for personality traits prediction. All the features must have some role in personality traits prediction process. Thirdly, the algorithm used in personality traits prediction process must be firm.

more Digital footprints help predict the personality traits better.

Table 1: Comparison of various personality prediction method

| Sr No. | Paper | Year of | Approach used | Digital Footprints | Brief Description |
|---|---|---|---|---|---|
| 1 | [1] | 2009 | Unsupervised Learning | Users' post, hometown (demographic, activity) | The authors of this paper have composed Big Five personality traits and the user's Facebook behavior. They have stated all online user's behavior for each personality traits. |
| 2 | [2] | 2018 | Supervised Learning | Users' post (demographic) | The goal of this paper was to develop an appropriate algorithm with some degree of tolerance to outliers. Maximum Margin Criterion and SVM algorithm used for personality traits prediction. |
| 3 | [3] | 2019 | Semi-supervised Learning | Users' status, time (demographic) | In this paper, the personality traits prediction framework using semi-supervised learning is introduced. Semi-supervised learning takes advantage of unlabeled data and helps to improve accuracy. |
| 4 | [4] | 2018 | Unsupervised Learning | Users' post (demographic) | The authors of this paper have introduced AttRCNN structure for personality traits prediction. They have tested that AttRCNN gives better accuracy. |
| 5 | [5] | 2018 | Supervised Learning | Users' Answer (Questionnaire) | In this study, the authors of this paper have introduced two deep learning structures for personality traits prediction. They have used multi-layer perceptron (MLP) and LSTM. |
| 6 | [6] | 2018 | Unsupervised Learning | Users' post, No. of friends, interests, age, gender (like, activity) | The authors of this paper have compared four machine learning models. They performed a correlation between feature sets and personality traits. They have acquired good accuracy. |

## 6. CONCLUSION

Due to the rapid growth of social media platforms, the volume of users' Facebook posts and activity grows. Users post their views and opinions on any topic; this helps to predict a particular user's personality. Predicting personality traits helps in various ways like Job or product recommendation, political views, mental health diagnoses, etc. In this paper, different techniques have been reviewed for predicting personality traits based on Supervised, Unsupervised and Semi-supervised learning. This paper has discussed various features that influence personality traits prediction. Personality traits prediction process rely on several Social media features called Digital Footprints. Combining two or

## REFERENCES

[1] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with Facebook use," Computers in Human Behavior, vol. 25, no. 2, pp. 578–586, 2009.

[2] Feng Zhong, X., Ze Guo, S., Gao, L., Shan, H. and Xue, D. (2018). "A General Personality Prediction Framework Based on Facebook Profiles." ACM.

[3] H. Zheng and C. Wu, "Predicting Personality Using Facebook Status Based on Semi-supervised Learning,"

Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC 19, 2019.

[4] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, and J. Sun, "Deep learning-based personality recognition from text posts of online social networks," Applied Intelligence, vol. 48, no. 11, pp. 4232–4246, May 2018.

[5] G. An and R. Levitan, "Lexical and Acoustic Deep Learning Model for Personality Recognition," Interspeech 2018, Feb. 2018.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality Predictions Based on User Behaviour on the Facebook Social Media Platform," IEEE Access, vol. 6, pp. 61959–61969, 2018.

[7] J. Yu and K. Markov, "Deep learning-based personality recognition from Facebook status updates," 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017.

[8] R. R. Mccrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications," Journal of Personality, vol. 60, no. 2, pp. 175–215, 1992.

[9] Tripathi, A. (2010). Personality Prediction with Social Behaviour by Analysing Social Media Data- A Survey. ACM.

[10] D. Azucar, D. Marengo, and M. Settanni, "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis," Personality and Individual Differences, vol. 124, pp. 150–159, 2018.

[11] W. Yiny, K. Kanny, M. Yuz, and H. Schijutze, "Comparative Study of CNN and RNN for Natural Language Processing," arXiv:1702.01923v1, Feb. 2017.

[12] S. Kleanthous, C. Herodotou, G. Samaras, and P. Germanakos, "Detecting Personality Traces in Users' Social Activity," Springer.

[13] V. Nasteski, "An overview of the supervised machine learning methods," Horizons.b, vol. 4, pp. 51–62, 2017.

[14] Z. Ghahramani, "Unsupervised Learning," SpringerReference.

[15] https://en.proft.me/2015/12/24/types-machine-learning-algorithms

[16] G. Bajrami, M. O. Derawi, and P. Bours, "Towards an automatic gait recognition system using activity recognition (wearable based)," 2011 Third International Workshop on Security and Communication Networks (IWSCN), 2011.

[17] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3, no. 1, pp. 1–130, 2009.

[18] S. Uhlmann, S. Kiranyaz, and M. Gabbouj, "Semi-Supervised Learning for Ill-Posed Polarimetric SAR Classification," Remote Sensing, vol. 6, no. 6, pp. 4801–4830, 2014.