International Journal of Current Engineering and Technology
©2021 **INPRESSCO**®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

User Personality Prediction on Social Media using Machine Learning

Poonam L. Patil and S. R. Jadhao

Department of Computer Engineering Gokhale Education Society's R.H.Sapat College of EngineeringManagement Studies and Research, Nashik-5

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

various statistics are split widely through social media Such as Facebook, Twitter. Data about the person and what they communicate through the status updates are important for research in human personality. This paper intends to scrutinize the forecasting of personality traits of Facebook users bases on machine learning and part of the Big five model this experiment uses my personality data set of Facebook users are used for linguistic factors respective to personality correlation. We use the Data Prepossessing concept of data mining after that feature Extraction.Next we will work on feature selection. The Personality Prediction system built in the XGboosting classification model.

Keywords: *Social network, Big five Model, Feature Analysis, Personality Prediction.*

Introduction

Now a Days from social media like Facebook, Twitter , Reddit Have become most trendy The propagation's of internet and intelligence technology, exclusively the online social network have revitalized how users communicate with other electronically , the social media application such as Facebook, Twitter , Instagram ,Reddit not only introduce the written and multimedia contain but also grant to circulate their feelings, Moods, emotions online [1] Figure 1 Shows no of users monthly in India per year . Personality is the characteristic the way of thinking, feeling behaving. The distinct personality is associated to the structure of various social relations and co-Operations behavior on status profiles. Our research predicts the personality based on user's social behavior and their language used for posting the status on social media platform although the Facebook is the presently longer used to share photos, Video status.

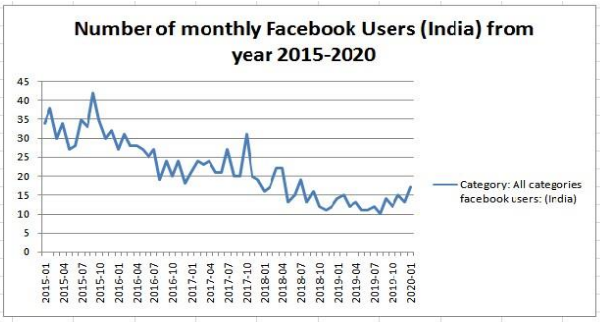


Fig. 1. Number of Facebook users monthly in India between 2015-2020

This accepts used to predicate personality there for the goal of this research is to build the prediction system that can automatically predict the user personality based on their activity on the social media [2]. First, we select the most favorable feature for each personality major and profitable forecast person personality Next, we propose the method one category of social network feature, we analyze the interrelationship between each of the feature set personality traits. In the social network feature some classes of anatomical network properties such as Networksize, Betweenness, Transitivity, Density,Brokerage majors as well as their connections with personality traits. we investigate the feature with larger co-relation with Personality Traits. Finally we proposed machine learning algorithm for predication of personality by using Boosting. The data collected by means of social media platform my personality project data set used for feature category like SNA, LIWC, SPLICE.

A. Big Five Model

Big five model is mostly used recently to measure the personality. It describes the human personality structure. It diminishes the greater number of personal objectives five most personality traits. That model the composition OCEAN [12][13].The Table 1 Shows the five facts defining each of factor with their characteristics [14] such as Openness, Conscientiousness, Extra-version, Agreeableness, Neuroticism.

- Openness: It is a general gratefulness for art, emotions, imagination and variation of knowledge.

- Conscientiousness: It is tendency to display selfdiscipline, try for achievements. The average level of conscientiousness move up among the young to adults and then dismiss among the older to adults.
- Extra-version: It is distinguish by spread of activities, energy creation from external means.
- Agreeableness: It is Trait reflects the personaliy in which the people are more cooperative.It reflects the helpfullness personality.
- Neuroticism:It is negative traits of personality it express the sad emotions like depression,or moody.

Table I Overview of Big Five Model Personality Traits

Personality Traits	Some Characteristics
Openness(O)	Curious, open to new ideas, creative, intellectual, original, Artistic, curious, imaginative, curious, intelligent, and imaginative
Conscientiousness(C)	Efficient, organized, responsible, organized, and persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners.
Extraversion(E)	Outgoing, talkative, Express positive emotion, satisfied, Friendly, Energetic, active, assertive, outgoing, amicable, assertive, Friendly and energetic, extroverts draw inspiration from social situations.
Agreeableness(A)	Affable, tolerant, sensitive, trusting, kind, and warm truthful, helpful, nurturing, optimistic, concerned, trusting of others, cooperative, compassionate, nurturing.
Neuroticism(N)	Anxious, Irritable, temperamental, moody, angry, upset tense, self-pitying, insecure, sensitive. Neurotics are experiencing negative emotions.

The remaining of this paper is arrange as follows Section II we discussed the literature survey related to personality predication in Section III we describe the System Architecture and System Overview in Section IV we describe the System Analysis Section V represent Result and Discussion. we conclude our paper in Section VI.

Review of Literature

There are number of research paper on personality predication on social media Personality predication subjects divided in to two methods: Computational Fundamentals Social network analysis.

Tandera et al.[2] They used the two datasets, one from mypersonaliy facebook dataset and other is manually composed. They Predict the Personality of the person by using the Big Five Model. Using the Support vector machine the achieved the topmost Prediction accuracy of 70.40%.

Pennebaker Key et al. [3] Introduce work related to personality extraction from the text they examine the words in different factors such as diaries, college assignments and social psychological manuscripts to observe the personalityrelated features with linguistic library. The Result shows that Agreeableness personality trades tents to use more text the neuroticism used more negative/sad words.

Ana CES lim et al.[4] wrote a pioneering work dedicated to personality prediction into a multi-label classification problem. In that, They process more than one personality trait. They were classied the personality traits of Twitter using Naıve Bayesian prediction model..

Argaman et al [5] They classified the personality traits namely neuroticism and Extraversion using Linguisting feature. They observed that neuroticism is correspondence to the functional lexical feature and the extraversion trait result in less observed.

In N.M.A listeria et al [6] They introduced the Naive Bayesian method for the classification of personality traits. In that, the Naive Bayesian method consists of two phases such as the Learning phase and classification phase. The userwritten text is used as input for predicting the personality then match them to find the partner on online dating sites.

Soujanya Poria et al [7] The Author proposes incorporating the sentiments, common sense knowledge and affective from text using resources. They mixed the common-sense knowledge-based feature with psycho-linguistic features and frequency-based features and then employed with a supervised classifier. Next, they developed five SVM models for five personality traits. There result enhance the accuracy of the existing framework by using the psycho-linguistic feature and frequency-based analysis at lexical level.

Go beak et al [8] predicted the personality of 279 Facebook users. In which the find the word count as Linguistic feature and friend count as SNA feature.

Sibel Adult et al [9] Predict the personality of user from Facebook Data and text from Twitter.They introduced the number of measures related to number of social media. They analyzed these features based on textual analysis of message send by another user. The aim of our study examines the all personality traits from the structure of social network analysis to the personality interaction using my personality project dataset [10] as well as Facebook API.

Ong e al.[11] Predict the personaliy based on Twitter informaio in Bahasa Indonesia.The system uses the 329 users of Twitter social media to predict the personlaity.They uses the XGboost classification model.

System Architecture

A. Problem Statement

To design the Xgboosting algorithm for increasing the accuracy of personality prediction of user on social media.To Predict the personality of user on social media using machine learning approach.

B. *System Architecture*

Personality Prediction model consist of following terms

1) Data Preprocessing

All the data goes through preprocessing stages before it processed. Preprocessing perform steps. Consist of removing URLs, Symbols, Names, Stemming, removing stop word and lower cases. In our work we use python and machine learning. Data before using the machine learning perform the data preprocessing.

2) Feature Extraction

A User's behavior on social media is offered by current behavior of another user. In many applications available –for explaining such behavior happen and expand [15]. –In our work all data from Dataset classified in two parts:

- Text Feature Extraction: Analyse the content of social media status texts uses the dictionaries.
- Social network behavior analysis: Which consist of Network size, Transitivity, Brokerage, Density this information denotes the users network behavior on Facebook Before the text data feed in machine learning for extorting the feature the raw text status represented in the following forms: –Bag of words representation [16]: In this presentation every sentence is the different set of words in which we don't consider a grammar here repetition of word together in feature future classification .SPLICE (Structure programming for linguistic cue extraction) It is latest dictionary in this dictionary mostly used for personality – predication task [17]. We used it to extract 80 linguistic features which are related to positive or negative.

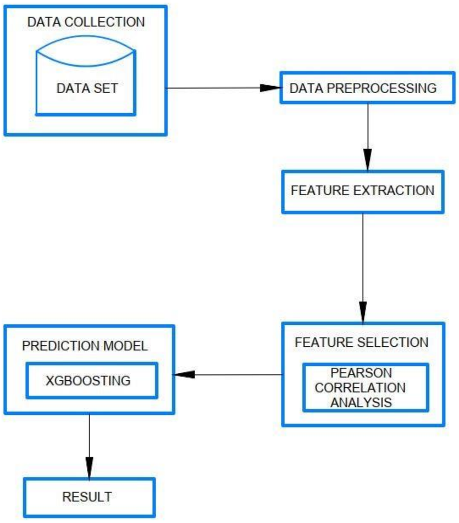


Fig. 2. Overview of System

NA (Social Network Analysis) It is method of collecting and examining data from social network such as Facebook, tweeter and Instagram. In our study we used feature related to social network of user with personality trades such as network size betweenness, density, brokerage and transivity .

- Network size: this introduce the no of social friends on social sides [18].
- Betweenness: reefers the number of parks between pair of individual those are not connected to each other directly through the density it indicates the potential connection on network that are actual connections more density network induces more dissipation between persons in information flow [19]
- Brokerage :it is a state or situation in which person connects another people and the unconnected action or fills the gap in social structure.
- Transitivity: it is based of friend of my friend is also my friend in which single are directly connected to each other one of them is only accessible with other individual represent them frequency of interaction between the network nodes [20] [21].

C. LDA algorithm

LDA(Linear Discriminant Analysis)Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the preprocessing step for pattern-classification and machine learning applications.The Feature Extraction is done by using the LDA algorithm. Five general steps for performing the a Linear discriminant anal- sis: – Calculate the d-dimensional mean vectors for the different classes from the personality dataset. – Compute the scatter matrices. Within the class scatter matrix *Sw* is computed by the following equation.

$$S_w = \sum_{i=1}^{c} S_i \quad (1)$$

where

$$S_i = \sum_{x=1}^{n} (x - m_i)(x - m_i)^T \quad (2)$$

where m_i is Mean Vector.

Compute the eigenvectors (e1,e2,...,e_d) and corresponding eigenvalues (1,2,...,d) for the scatter matrices.

Compute the eigenvectors (e1,e2,...,ed) and corresponding eigenvalues (1,2,...,d) for the scatter matrices.

Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a dk dimensional matrix W_w.

Use this d * k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: Y=X * W (where X is a n * d dimensional matrix representing the n samples, and Y are the transformed n * k dimensional samples in the new subspace). 3) Feature Selection

For constructing classification model the feature connection is important. feature selection is preprocessing step for a machine learning. feature selection is used for dimensionality reduction and can effectively find the both irrelevant and redundant features. There are two methods to major the correlation between two random variable i) Based on Classified Liner correlation. ii) Based on information theory [22]. To Measure the stability of linear correlation between two variable and criticize the important features for personality traits prediction we use the Pearson Correlation coefficient. It is a part of the linear correlation between two variables X and Y[22].For the pair of variable(X,Y),the linear Correlation coefficient formula of r(X,Y) is given by:

$$r(X,Y)=\frac{\sum_i (X_i-\overline{X}_i)(Y_i-\overline{Y}_i)}{\sqrt{\sum_i (X_i-\overline{X}_i)^2}\sqrt{\sum_i (Y_i-\overline{Y}_i)^2}}$$

$$\overline{X}_i=\frac{1}{n}\sum_{i=1}^{n}X_i$$

where

$$\overline{Y}_i=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

are the mean of the X and Y Variable and n is sample size.The value of correlation coefficient in between 1 and 1.If X and Y Variable are completely parallel then r(X,Y) takes the Value 1 as Positive Correlation or for Negative Correlation it takes the value as -1.If both variable are independent on each other it takes the value as Zero[23][24].

4) Prediction Model

After finding the correlation between the features we apply the classifier for predicting the personality traits. We use the XGBoost model as classifier.XGBoost is an algorithm. Also, it has recently been dominating applied machine learning. XGBoost is a gradient boosted decision trees implementation.It is a type of software library in Python.There are number of inerfaces available to access this model such as Python interface along with integrated model in scikit-learn.we use the Python interfaces for XGBoost model. The Algorithm for Building the XGBoosting Model Perform the following Steps:

- Load All the Python Libraries Here We load all the libraries of python such as XGBoost,Readr,Stringr.
- Next part is to load all collected Dataset (Here We Use the mypersonality Dataset of Facebook User)
  - First Load the label Of Train Data.
  - Next Combine the Training and Testing Data .
- Perform Data Cleaning
  - Here All the Feature are Categorized in various format and Perform The Data Prepossessing.
  - Splitting of Training and Testing Data.
- Tune and Run he Model.
- Predicting score test set.

System Analysis

A. Mathematical Model

The Mahematical Model Of User Personality Prediction Method is:

Set Of Inputs

D:Set Of Mypersonality Facebook Dataset.

Functions DP:Data-Prepossessing FE:Feature Extraction: LDA Algorithm

FS:Feature Selection

PM:Prediction Model

Output

PT:Personality Traits.

B. Implementation Details

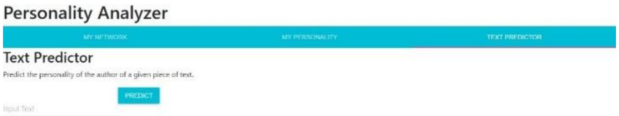
Hardware Requirement

1) CPU Speed:2GHz
2) RAM:Minimum 3GB
3) Hard Disk:Minimum 100 GB
4) Input Devices and Mouse,Keyboard,Monitor

Software Requirement

1) Operating System:Windows 7
2) Programming Language:Python
3) Database:Mongodb

Result And Discussion



Personality Analyzer

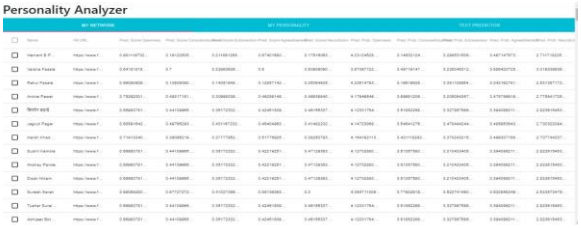
Text Predictor

Predict the personality of the author of a given piece of text.

Fig 3. Shows the image to enter text to predict the personality of user.

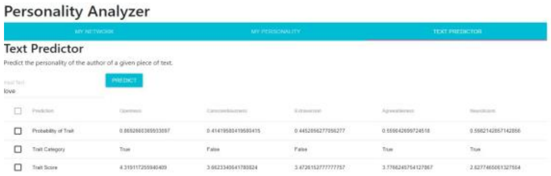
Fig 4. The Feature Selection of the user personality

Fig. 3. image of text predictiontrait. Fig 5. shows the text prediction on openness personality traits.



Personality Analyzer

Fig. 4. Personality Classification Analyser



Personality Analyzer

Text Predictor

Predict the personality of the author of a given piece of text.

love

□	Prediction	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
□	Probability of Trait	0.8932660365933097	0.41419583419580415	0.4452956277056277	0.559042699724518	0.5982142857142856
□	Trait Category	True	False	False	True	True
□	Trait Score	4.315117255940409	3.6623340641780824	3.472615277777757	3.7768245754127867	2.6277465061327564

Fig. 5. image of Openness Personaliy Prediction

Conclusion

Social network analysis has increased largely in recent times. To extract the personality of any person on the social networking websites is very useful for many applications in various domains like including job success, attractiveness, and happiness. Personality detection from social media is to extract the feature from there updates and the behavior attribute of a person from the written text on social media. This Prediction Model help to predict the personality of user from social media. Xgboosting prediction model outperforms than the other prediction models.

Acknowledgment

I have huge delight in presenting the project "User Personality prediction on Social Media Using Machine Learning" under the guidance of Dr. S. R. Jadhao and PG Coordinator Prof.A. S. Vaidya. I gratefully thanks to Gokhale Education Society's R.H.Sapat College f Engineering,Management Studies Research, Nashik-5 for giving required equipment's,web get and Research Books. I also Thank to Teaching and Nonteaching staff of GESRHSCOE who helped me.

References

[1]. Md.Rafiqul Islam,Ashad Kabir,Hua Wang,Anwaar Ulhaq,"Depression Detection From Social network data using machine Learning techniques,".by springer Nature Switzerland AG 2018.

[2]. T. Tandera et al,"Personality prediction system from Facebook users,"Procedia Comput. Sci.,vol. 116, pp. 604–611, Dec. 2017.

[3]. J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015.

[4]. Lima, Ana CES, and Leandro N. De Castro.,"Multi-label Semi-supervised Classification Applied to Personality Prediction in Tweets," Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI and CBIC), 2013 BRICS Congress on. IEEE, 2013.

[5]. S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, "Lexical predictors of personality type," Tech. Rep., 2005.

[6].] N.M.A Lestari,I.K.G.D .Putra, A.A.K.A. Cahyawan,"Personality types classification for Indonesian text I partners searching website using Na¨ıve Bayes Methods",International Journal of software and Informatics Issue,2013.

[7]. Soujanya Poria , Alexandar Gelbukh, Basant Agarwal,Erik Cambria , and Newton Howard, "Common Sense Knowledge Based Personality. Recognition from Text ",Springer-Verlag Berlin Heidelberg pp. 484–496,2013.

[8]. J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in Proc. Extended Abstr. Hum. Factors Comput. Syst. (CHI), 2011, pp. 253–262.

[9]. S. Adali and J. Golbeck, "Predicting personality with social behavior," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2012, pp. 302–309.

[10]. M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges,ethical considerations, and practical guidelines," Amer. Psychol., vol. 70,no. 6, pp. 543–556, 2015.

[11]. Ong et al., "Personality prediction based on twitter information in Bahasa Indonesia," in Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS), 2017, pp. 367–372.

[12]. L. R. Goldberg, "The structure of phenotypic personality traits," Amer.Psychol., vol. 48, no. 1, pp. 26–34, 1993.

[13]. E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," J. Pers., vol. 60, no. 2, pp. 225–251, 1992.

[14]. O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," Handbook of Personality: Theory and Research, vol. 2. 1999, pp. 102–138

[15]. T. P. Michalak, T. Rahwan, and M. Wooldridge,"Strategic social network analysis," in Proc. AAAI, 2017, pp. 4841–4845.

[16]. Soumya George K,Shibily Joseph,"Text Classification by Augmenting Bag Of Words(BOW) Representation with Co-occurence feaure",IOSR Journal of computer Engineering ,2014,pp 34-38.

[17]. K. Moffitt, J. Giboney, E. Ehrhardt, J. Burgoon, and J. Nunamaker, "Structured programming for linguistic cue extraction (SPLICE)," in Proc. HICSS Rapid Screening Technol., Deception Detection Credibility Assessment Symp., 2012, pp. 103–108.

[18]. J.E. Lonnqvist, J. V. Itkonen, M. Verkasalo, and P. Poutvaara, "The¨ fivefactor model of personality and degree and transitivity of Facebook social networks," J. Res. Person., vol. 50, pp. 98–101, Jun. 2014

[19]. H. Lin and L. Qiu, "Sharing emotion on Facebook: Network size, density,and individual motivation," in Proc. Extended Abstr. Hum. Factors Comput. Syst. (CHI), 2012, pp. 2573–2578.

[20]. M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 68, no. 3, p. 036122, 2003.

[21]. M. Aghagolzadeh, I. Barjasteh, and H. Radha, "Transitivity matrix of social network graphs," in Proc. IEEE Stat. Signal Process. Workshop (SSP), Aug. 2012, pp. 145–148.

[22]. L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in Proc. 20th Int. Conf. Mach. Learn. (ICML), 2003, pp. 856–863.

[23]. D. Cramer, Fundamental, "Statistics for Social Research: Step-byStep Calculations and Computer Techniques Using SPSS for Windows," Evanston,IL, USA: Routledge, 2003.

[24]. M. A. Hall, "Correlation-Based Feature Selection for Machine Learning,"1999.