International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 6, September-October 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

User Personality Prediction on Facebook Social Media using Machine Learning

Poonam L Patil[1], Dr. S. R. Jadhao[2]

[2]Assistant Professor,
[1,2]Department of Computer Engineering, R.H. Sapat College of Engineering,
Management Studies and Research Savitribai Phule Pune University, Nashik, Maharashtra, India

ABSTRACT

In recent years, Social network use is increasingly build-up. The 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 used 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 media, big five model, machine learning, personality prediction, feature analysis, social network*

How to cite this paper: Poonam L Patil | Dr. S. R. Jadhao "User Personality Prediction on Facebook Social Media using Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-6, October 2020, pp.503-509, URL: www.ijtsrd.com/papers/ijtsrd33414.pdf

IJTSRD33414

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]. The Fig 1.shows that the number of monthly social media users in India from Jan 2020- June 2020. 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. 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].

A personality prediction model based on texts extracted from social media that can be useful in several areas, including marketing intelligence and social psychology, due to the high volume of information generated and the exposure level. The recognition of personality traits helps to find out the mutual conduct and may provide a subjective view to text mining in social media, such as: sentiment analysis, text clustering, and recommendation systems.

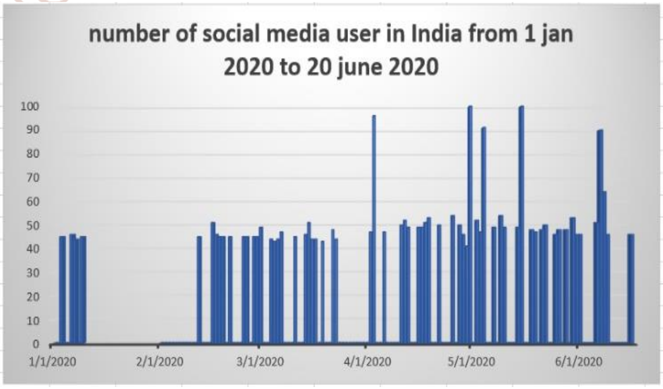

Figure. 1 number of social media users in India from 1 January 2020 to 20 June 2020 (using google trends)

First, we select the most favorable feature for each personality major and portable forecast person personality. Next, we proposed XGboosting method for predicting the personality of social media user. we propose the method one category of social network Feature, we analyzed the interrelationship between each of the personality traits.. In the social network Feature some classes of anatomical network properties such number of friends on social media 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.

A. Big Five Mode

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, Extraversion, Agreeableness, Neuroticism.

- **Openness**: It is a general gratefulness for art, emotions, imagination and variation of knowledge.
- **Conscientiousness:** It is tendency to display self-discipline, try for achievements. The average level of conscientiousness moves up among the young to adults and then dismiss among the older to adults.
- **Extra-version:** It is distinguishing by spread of activities, energy creation from external means.
- **Agreeableness:** It is Trait reflects the personality in which the people are more cooperative. It reflects the help-fullness personality.
- **Neuroticism:** It is negative traits of personality express the sad emotions like depression, or moody.

TABLE1. OVERVIEW OF BIG FIVE MODEL

Personality Traits	Some characteristics
Openness(O)	Openness measures the curiosity, tolerance, imagination, creativity, tolerance, political liberalism, and appreciation for culture. People scoring high on Openness such as switch, new and unusual ideas, appreciate and have a good sense of aesthetics
Conscientiousness(C)	Conscientiousness measures desire for a coordinated approach to life in comparison to a spontaneous one. People scoring high on Conscientiousness are more likely to be well organized, reliable, and consistent. people scoring high on conscientiousness pursue long-term goals, enjoy planning and seek achievements.
Extraversion(E)	Extroversion measures a tendency to seek stimulation in the external world,to express the positive emotions and give the company of others. People scoring high on Extroversion tend to be more outgoing, friendly, and socially active. They are usually energetic and talkative.
Agreeableness(A)	Agreeableness relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. People scoring high on Agreeableness tend to trust others and adapt to their needs. They also related to the short bound by social trust and conventions
Neuroticism(N)	Neuroticism measures the tendency to experience mood swings andemotions such as guilt, anger, anxiety, and depression.

The remaining of this paper is arrange as follows. We discussed the literature survey related to personality predication. we describe the System Architecture and System Overview. we describe the System Analysis. Next, Result and Discussion. Finally, we conclude our paper.

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, they 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 personality related 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 classified the personality traits of Twitter using Naïve Bayesian prediction model.

Argaman et al [5] They classified the personality traits namely neuroticism and Extraversion using Lingui sting 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 Naïve 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 user- written text is used as input for predicting the personality then match them to find the partner on online dating sites.

SoujanyaPoria et al. [7] They proposed a new approach for personality detection which is based on incorporating the sentiment, affective and common-sense knowledge from the text using resources. In their approach, they combined common sense knowledge-based features with phycho-linguistic features and frequency-based features and later the features were employed in supervised classifiers. Further, they developed five support vector machine models for five personality traits. They designed five Social Media Optimization (SMO) based supervised classifiers for five personality traits. Their experimental results show that the use of common-sense knowledge with perceptual and sentiment information with psycho-linguistic features and frequency-based analysis at lexical level that upgrades the accuracy of the current frameworks.

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 personality based on Twitter information in Bahasa Indonesia. The system uses the 329 users of Twitter social media to predict the personality. They use the XGboost classification model.

SYSTEM OVERWIEW

A. Problem statement

To effectively evaluate the performance of XGBoost using machine learning for the accuracy of personality prediction of user on social media.

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: Analyze 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.

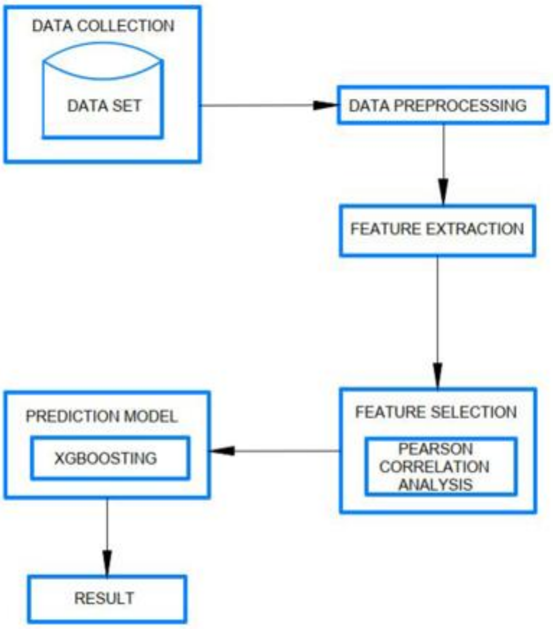

Figure 2 System overview

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.

SNA (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 Transitivity.

Network size: This introduce the number of social friends on social media [17].

Betweenness: Refers the number of parks between pair of individuals 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 [18].

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 on 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 [19] [20].

3. Feature Selection

For constructing classification model the feature connection is important. feature selection is preprocessing step for a machine learning. feature selection is use for dimensionality reduction and can effectively find the both irrelevant and redundant features. There are two methods to major the correlation between two random variables.

Based on Classified Liner correlation.

Based on information theory [21].

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 [21]. For the pair of variable(X,Y),the linear Correlation coefficient formula of r(X,Y) is given by:

$$r(X,Y) = \frac{\sum_{i}(X_i - \overline{X}_i)(Y_i - \overline{Y}_i)}{\sqrt{\sum_{i}(X_i - \overline{X}_i)^2}\sqrt{\sum_{i}(Y_i - \overline{Y}_i)^2}} \qquad (3)$$

Where

$$\overline{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (4)$$

And

$$\overline{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (5)$$

These 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 variables are independent on each other it takes the value as Zero [22][23].

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 interfaces 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.

XGboosting algorithm working

1. First model F0 is defined to predict the target variable Y.
2. fit model to the residual h1(x) =Y-F0.
3. Create a new model using the h1(x) and F0 to give the F1, it is boosted version of F0.
4. The mean squared error of F1 will be lower than the F0.
5. To Improve the performance of F1 model, then residuals of F1 and create a new model F2. F2(x)<-F1(x)+h2(x)
6. This can be done for the 'm' iterations.

RESULT AND ANALYSIS

Table 2 shows that the word "love" shows the positive emotions to words the extraverted personality traits. Table shows the Probability score of all personality traits and category of the personality. Table 3 shows the personality score of the "python" word. The Python word gives the openness type of personality trait. Based on the correlation results for the social network features, we found that extraversion represents the highest correlated trait.

TABLE 2 SHOWS THE RESULT OF THE TEXT ANALYSIS OF THE "LOVE" WORD

Text analysis	Personality traits	Prediction Score	Prediction Probability score	Trait Category
"Love"	Openness(O)	0.8692	4.319	True
	Conscientiousness(C)	0.4149	3.662	False
	Extraversion(E)	0.4452	3.472	True
	Agreeableness(A)	0.5590	3.776	True
	Neuroticism(N)	0.5982	2.627	True

TABLE 3 SHOWS THE RESULT OF THE TEXT ANALYSIS OF THE "PYTHON" WORD

Text analysis	Personality traits	Prediction Score	Prediction Probability score	Trait Category
"Python"	Openness(O)	0.664	4.123	True
	Conscientiousness(C)	0.441	3.518	False
	Extraversion(E)	0.351	3.330	False
	Agreeableness(A)	0.424	3.393	False
	Neuroticism(N)	0.461	2.808	False

Table 4 shows the personality analyzer which indicate that the all profiles are collected from the social media and calculate the personality score.

TABLE 4 SHOWS THE RESULT OF THE SOCIAL MEDIA FRIENDS PERSONALITY PREDICTION SCORE

Users	Prediction Score					Prediction Probability Score				
	O	C	E	A	N	O	C	E	A	N
User1	0.98	0.19	0.21	0.67	0.17	4.23	3.14	3.29	3.49	2.71
User2	0.64	0.71	0.23	0.6	0.50	3.67	3.49	3.23	3.68	3.01
User3	0.99	0.10	0.14	0.12	0.05	4.20	3.19	3.38	3.04	2.53
User4	0.71	0.39	0.37	0.51	0.38	4.16	3.43	3.37	3.49	2.70
User5	0.66	0.44	0.35	0.42	0.47	4.12	3.51	3.21	3.39	2.82

The Personality Prediction score of the all five personality traits are shown in the figure 3. The result shows that the openness personality traits give the highest personality score after prediction.

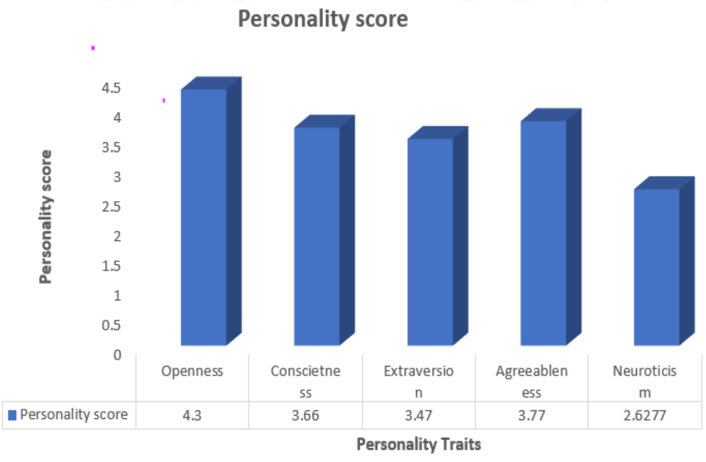

Figure 3 Shows the personality score of all Five personality traits prediction

We compare the result of accuracy with the existing machine learning algorithm. Figure 4 shows that the personality prediction result analysis. Finally, for achieving maximum prediction accuracy, Xgboosting model is used for the maximum accuracy.

TABLE 4 SHOWS THE ACCURACY OF ALL PERSONALITY TRAIT IN PROPOSED AND EXISTING SYSTEM

Performance Measures	Personality Traits	Proposed System % accuracy	Existing system % accuracy
Accuracy	Openness(O)	78.66	68
	Conscientiousness(C)	64.5	61.6
	Extraversion(E)	76	44
	Agreeableness(A)	60.8	54.8
	Neuroticism(N)	64	64.8

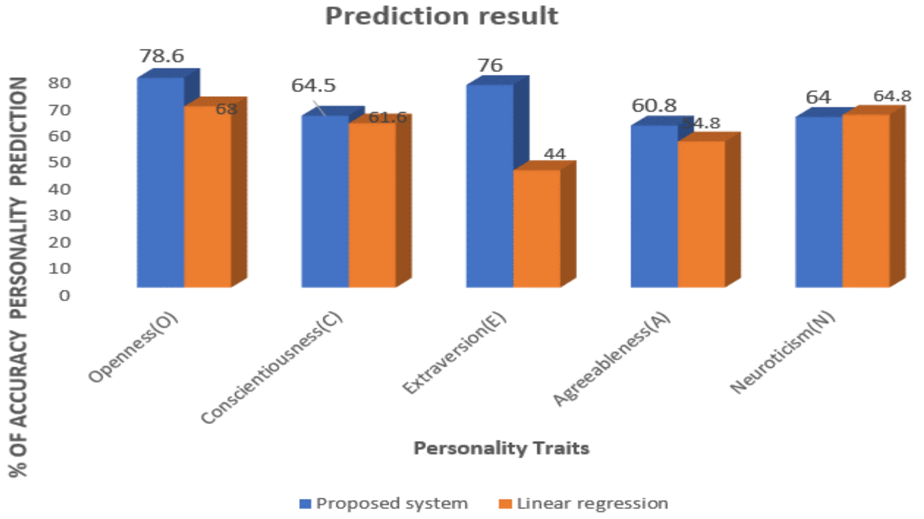

Figure 4 Performance analysis

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 on the all personality traits.

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 our HOD Dr. D. V. Patil to Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies Research, Nashik-5 for giving required equipment's, web get and Research Books. I also Thank to Teaching and Non-teaching staff of GESRHSCOE who helped me.

References

[1] Islam Md. Rafiqul, Kabir Ashad, Ulhaq Anwar, wang Hua,, "Depression Detection from Social network data using Machine Learning techniques," by springer Nature Switzerland AG 2018.

[2] Tandera T, Hendro, Suhartono D, Wongso R, Yen Lina Praseti, "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] SoujanyaPoria, AlexandarGelbukh, 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] V. 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-occurencefeaure", IOSR Journal of computer Engineering, 2014, pp 34-38.

[17] J. E. Lonnqvist, J. V. Itkonen, M. Verkasalo, and P. Poutvaara, “The five factor model of personality and degree and transitivity of Facebook social networks,” J. Res. Person., vol. 50, pp. 98–101, Jun. 2014

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Cramer, Fundamental, “Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows,” Evanston, IL, USA: Routledge, 2003.

[23] Hall M A, “Correlation-Based Feature Selection for Machine Learning,” 1999.