Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 177–184

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method

Ninda Anggoro Utami[a], Warih Maharani[a,*], Imelda Atastina[a]

[a]*School of Computing, Telkom University, Bandung, Indonesia, 40257*

Abstract

Social media has become one of the most important things in daily life to communicate, show expression and exchange information. Facebook is one of the most widely used social media. This research focuses on classifying the personality of Facebook users into one of the Big Five Personality Traits. there are 170 volunteers who are Facebook users who have been asked to fill out the Big Five Inventory questionnaire and have allowed their data to be scraped. Based on the data collected, the classifier is built using data mining techniques using Support Vector Machine (SVM) that aim to find out someone's personality based on a Facebook account without having to fill in any questionnaire. The best accuracy results in this study with a classification model that has been built at 87.5% using the Radial Basis Function (RBF) kernel.

Keywords: Personality; Big Five Personality Traits; Data Mining; Classification; Support Vector Machine (SVM)

1. Introduction

Social media is one place for someone to express themselves. The ease of accessing social media makes the majority of people do various kinds of activities on their social media accounts, such as telling stories about daily life, sharing experiences, communicating, and many things. Facebook is a social media that is widely used with a number of users reaching 1.8 billion users, and around 800 million users are active in doing activities on Facebook by spending approximately 40 minutes a day accessing Facebook[1]. Facebook's diverse user status makes a group of researcher's studies whether a person's personality type can be determined through the use of social media.

There are several personality models that can be used to predict personality. Commonly used models include the MBTI (Myers Briggs Type Indicator), DISC (Dominance Influence Steadiness Conscientiousness), and Big Five Personality. After several experiments and a review process from several literatures, the Big Five Personality Model was used in this study because this model is the most popular and accurate way to predict a person's personality traits.

* Corresponding author.
E-mail address: wmaharani@telkomuniversity.ac.id

10.1016/j.procs.2021.12.023

From the research of Cobb-Clark et al[2], Big Five personality traits are more consistent in time, because the methods owned by the Big Five will remain consistent over time so the data can be used at any time and that makes this model the best to use. As the name implies, Big Five Personality has five traits, including Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each person must have the most dominant personality among the five personalities[3].

There are research that examines personality classifications using the Support Vector Machine method, which explains the use of a Multi-Class Support Vector Machine that is good at solving personality traits cases that have five classes[4]. SVM can generalize more accurately on unseen cases relative to classifiers that aim to minimize the training error so it can generate better prediction on multi-class cases. Based on the results given by these methods, this research will use the Support Vector Machine (SVM) method because this method has the ability to generalize problems well so as to produce high accuracy and a smaller error rate than the pre-existing methods[5].

In this research, the purpose is to build a model of Support Vector Machine that can find the user's personality on Facebook, without asking them to fill out questionnaires and interviews related to personality according to activities they do on their Facebook, so that in the future, the results of this research can be used by companies to find the best human resources in accordance with the fields that are being needed by the company. Therefore, this study applies existing methods in data mining techniques to build the user's personality classifier.

2. Related Work

There are many studies that analyzing the personality activities in social networks. This research used some previous studies as the references to build the personality classifier model. The five-personality trait model was first created by Tupes and Christal as a fundamental characteristic that exists from the results of analyzes of other personality tests that have been done before[6]. Other research[7] continued the Big Five model research and found the Big Five traits are characterized by the following:

- Openness to Experience: Someone who has a trait of openness usually has a creative idea, imaginative, and full of curiosity. This is shown from the high scores on art
- Conscientiousness: Conscientiousness Trait has a responsible, diligent and organized nature. Someone who has this trait is usually reliable, good planner and hard worker.
- Extraversion: Someone who is easy to get along with, energetic, cheerful, and friendly are the characteristics of Extraversion. People who have extraversion trait can usually be friends with anyone.
- Agreeableness: The owner of Agreeableness trait is usually a cooperative person, likes to help others, is always optimistic and can be trusted.
- Neuroticism: Someone who has Neurotic Trait is usually a person who is sensitive, moody, easily anxious, and easily touched when experiencing a negative experience.

Another study[8] analyzed the relationship of personality between users interaction from different perspective by investigating the relationship between Big Five Personality with shyness, selfishness, and being alone nature of Facebook users. The results show that Facebook users who are active on social media tend to be more extroverted on social media but almost all feel socially alone, compared to non-Facebook users.

In other work of Maxim Stankevich et al[3], the classification is done using the Support Vector Machine method and the Random Forest method where the results are compared and the results are only slightly different. Previous studies[9] classify the personality based on the activities of Facebook user. The model to test the accuracy are using the Naïve Bayes Classifier, Decision Tree, and Support Vector Machine methods by classify the personality using Big Five Personality Traits and comparing each of the traits into two classes, i.e. Yes or No. The previous studies just only used the classification with binary class. In this paper, the research will develop the multiclass classification.

3. Proposed Method

The steps to be taken in this study are shown in Fig.1.

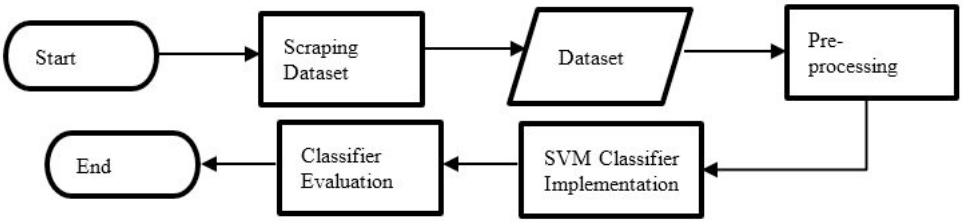

Fig. 1. Steps of Work Process

3.1. Scraping Dataset

The dataset used in this research was obtained with tools called ultimate-Facebook-scraper. The tools help to collecting data from each user who agreed to become a volunteer in this research. Before scraping, first, distributed the BFI (Big Five Inventory) questionnaire which was used to find out the most dominant personality of the volunteers.

BFI (Big Five Inventory) is one of the Big Five personality measuring devices consisting of 44 items of statements developed by John in 1990. This Big Five measurement tool has been translated in various countries by many researchers [10]. Each item related to each trait of Big Five. The volunteers will be asked to fill the questionnaire with score 1 if they agree to the statements and score 5 if they disagree with the statements. For example, there are items that related to Agreeableness trait. If the volunteer fill score 1 in this statement, then it will add 5 point to the Agreeableness. If the volunteer score 5 (disagree) to the items that related to Openness traits, then it will add 1 point to the Openness traits. The max points will be 220. After The 44 items has been calculated, the highest point received will be the most dominant traits that the volunteer has. The results of the BFI questionnaire will be used to determine the class attributes of each volunteer in classifying personality as data train.

After volunteers fill out the BFI questionnaire, the next step will be scraping the data on the Facebook account of each volunteers. There are 170 volunteers that their Facebook data will be scrapped. There are some data obtained from scraping, the data are Gender, Friends, Following, Games and Applications, Posts, Tagged Photos, Uploaded Photos, and Likes.

Table 1. Data of Facebook that Used in Building the model.

Attributes	Description	Type of Attributes
Gender	Gender of the users	Categorical
Friends	The total of friends on Facebook	Numeric
Following	The total of users following on Facebook	Numeric
Games and Applications	The total of users Games and Applications on Facebook	Numeric
Posts	The total of users posts on Facebook	Numeric
Tagged Photos	The total of tagged photos on Facebook	Numeric
Uploaded Photos	The total of uploaded photos on Facebook	Numeric
Likes	The total of liked pages on Facebook	Numeric

The features that selected in Table 1 have a high correlation with personality traits [11]. For example, according to the results of BFI questionnaire that user fills, it comes that the user got Agreeableness traits. that has 974 Friends, 2 Followings, 6 Games Applications, 1233 Posts, 73 Tagged Photos, 188 Uploaded Photos, and 122 Liked Pages is labeled to have Agreeableness Traits. These data will be the training data to build the classifier.

3.2. Data Preprocessing

Fig.2. is a step by step of preprocessing that will be used for this system. The data that has been collected will be processed first so the data will be of good quality and the results of the classification will be good. In this research, the

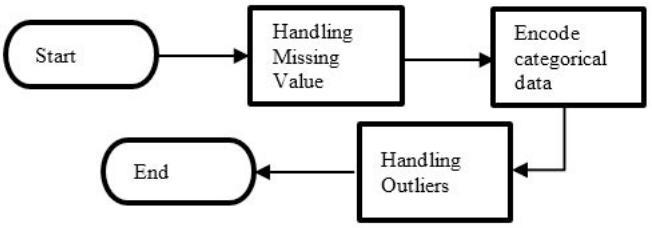

Fig. 2. The Steps of the Data Preprocessing

preprocessing step will be using the data mining techniques, there are handling missing value, encoding categorical data, and detection of outlier.

The first step in preprocessing is to overcome the problem of missing value. The basic strategy that will be used to overcome incomplete datasets is to remove all rows and / or columns that contain missing values. However, this can affect the value due to the possibility of losing data that might be valuable. This study uses the imputation algorithm to impute the missing values to infer them from the known part of the data [12]. The data that has the value of NaN will be replaced with the mean value of the attribute.

In the next step, coding will be done on categorical data where categorical type data will be converted into numerical data so that the data can be used in building classifiers.

Last, now the outlier detection and handling will be carried out. Outliers are values from a different set of data than other data and not describe the characteristics of the data. Outliers detected are candidates for deviant data which can cause model specification errors, biased parameter estimates, and incorrect results [13]. The outliers will be replaced with the mean value of the attribute.

After the preprocessing step has been completed, the data that have been preprocessed will be classified into five classes. The data will be split randomly into 80% as the data train and 20% as the data test to build the model because after several test, the best split data is on 80%: 20%. The algorithm used for the classification is Support Vector Machine (SVM).

3.3. Classify Using Support Vector Machine (SVM)

The basic principle possessed by the Support Vector Machine (SVM) method is a linear classifier, where SVM can classify data that can be separated linearly, but SVM has been developed to be able to overcome non-linear problems by applying the concept of kernel tricks in the workspace. The purpose of SVM is to get the best hyperplane that separates two classes in linear SVM and several classes in non-linear SVM [14].

This research will use several SVM kernels which will be tested among which kernels produce the best accuracy. kernels that will be used include Linear, Radial Base Function (RBF), and Polynomials. the output will be a classification model. Each kernel is calculated by following formula:

- Linier

$$K\left(X_i, X_j\right) = X_i^T X_j \tag{1}$$

- Polinomial

$$K\left(X_i, X_j\right) = \left(X^T X_j + 1\right)^d \tag{2}$$

- Radial Basis Function(RBF)

$$K\left(X_i, X_j\right) = \exp\left(-\gamma \|X_i - X_j\right)^2 \tag{3}$$

The method that will used to calculated the Support Vector Machine is using the Sequential Training Method. It is one of SVM algorithm that has a simpler algorithm and faster time needed to solve the SVM method. The Sequential Training algorithm is as follows:

- Initialize

$$\alpha_i = 0, C = 1, \gamma = 1, l = 100 \tag{4}$$

Calculate Matrix

$$D_{ij} = Y_i Y_j \left(K \left(X_i \cdot X_j \right) + \lambda^2 \right) \tag{5}$$

The Matrix will be used in the Sequential Training. Adjust the equation that bolded in the equation 5 based on the kernel that will be used.

- Sequential Training SVM
Do the steps (a),(b), and (c) below for i=1,2,. . . ,l
(a)

$$E_i = \sum_{j=1}^{l} \alpha_j D_{ij} \tag{6}$$

(b)

$$\delta\alpha_i = \min \{ \max \left[\gamma \left(1 - E_i \right), -\alpha_i \right], C - \alpha_i \} \tag{7}$$

(c)

$$\alpha_i = \alpha_i + \delta\alpha_l \tag{8}$$

- Return to step-2 until the α value converges (no significant change).

4. Experiments and Results

The testing in these steps are the test tuning the parameters used in the process of calculating the Support Vector Machine (SVM) method. The tuning parameters aims to get the most optimal parameters that will provide the best accuracy of each kernels, so the models will be good to use on classifying the personality. There are three parameters that will be tuned. They are Complexity, Gamma, and Max Iteration.

4.1. Testing C (Complexity) Variable

In this test, testing the value of C (Complexity) is done to find out how much influence C (Complexity) on the level of accuracy. In this test there are several C (Complexity) values to be tested. values to be used include 0.01, 0.1, 0.5, 1, 10. There are also several supporting variables that will be used in this test, like gamma = 0.1, max iteration = 100, and degree = 1.

In the Fig. 3. shows the highest level of accuracy obtained at a C value of 1, with an accuracy of 87.5% using the RBF kernel. C values are used to control the trade-off between margins and classification errors, but in tests that have been done, in linear and polynomial kernels, C values do not affect accuracy. this is because the data tested with the C value have a classification error that has no effect on margins, while the RBF kernel does. In the Linear Kernel, the accuracy is same for all C values, at 50%, while in the Polynomial kernel, the accuracy changes when the C value is 1, so the accuracy becomes 58.3%.

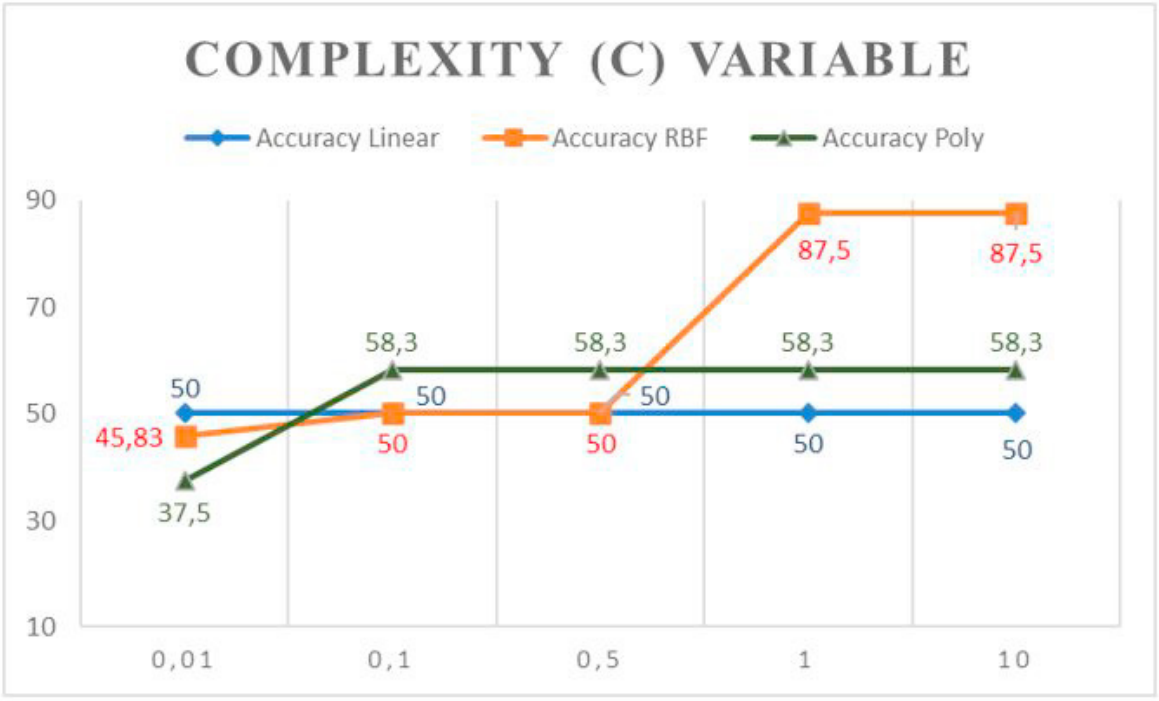

Fig. 3. Complexity (C) Variable Accuracy Test Results

4.2. Testing Gamma (γ) Variable

In this test, testing the value of Gamma (γ) is done to find out how much influence Gamma (γ) on the level of accuracy. in this test there are several Gamma values to be tested. values to be used include 0.01, 0.1, 0.5, 1, 10. There are also several supporting variables that will be used in this test, like C = 1, max iteration = 100, and degree = 1.

From testing the parameters of the gamma constant showed in Fig. 4., the greater the gamma value, the accuracy will decrease. Different accuracy is obtained, where the accuracy obtained by each kernel decreases when the gamma value is 0.1 with the accuracy value obtained in the linear kernel by 50% , the RBF kernel by 87.5% and the polynomial kernel by 58% and decreases when the value gamma added.

Basically, the gamma function is to set the learning rate. This is what causes the higher the gamma value, the learning rate will be higher, and if the learning rate is higher, the level of accuracy will be reduced. Therefore, all the kernel that used shown to have lower accuracy as the higher of the gamma value.

4.3. Testing Iteration Max Variable

In this test, testing the value of Iteration Max is done to find out how much influence Iteration Max on the level of accuracy. in this test there are several Itermax values to be tested. values to be used include 0.01, 0.1, 0.5, 1, 10. There are also several supporting variables that will be used in this test, like C = 1, max iteration = 100, and degree = 1.

In this scenario showed in Fig. 5., the RBF kernel has the highest accuracy results of 87.5% in each itermax value, whereas in a linear kernel, the highest accuracy is 62.5% i.e. when the itermax value is 250, then for the polynomial kernel, the highest accuracy of 50% is obtained at an itermax value of 250.

The Iteration Max is to set how many iterations that will be done. Each kernel has different algorithm and each time to solve the SVM, so with the limitation of iteration make some kernel doesn't seems to have differences of results. In this test, the greater the iteration value does not guarantee the higher accuracy value, because the alpha value has not reached convergence. decreased accuracy when iteration increases due to unbalanced support vector ratios, and some data is still located far from the ideal hyperplane.

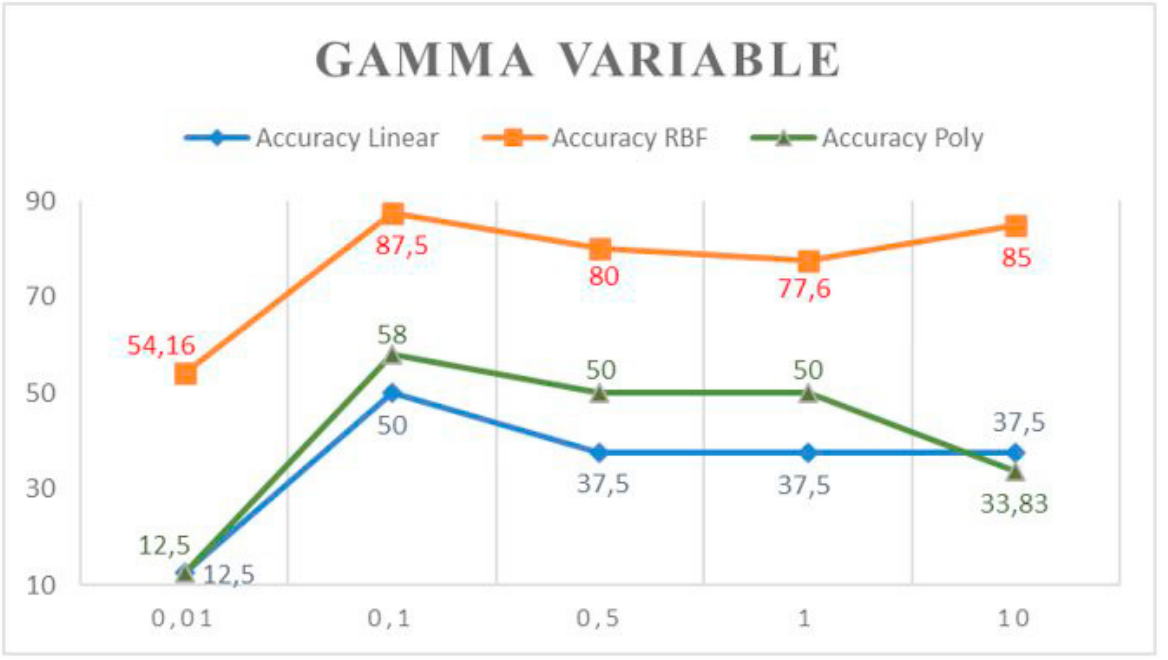

Fig. 4. Gamma Variable Accuracy Test Results

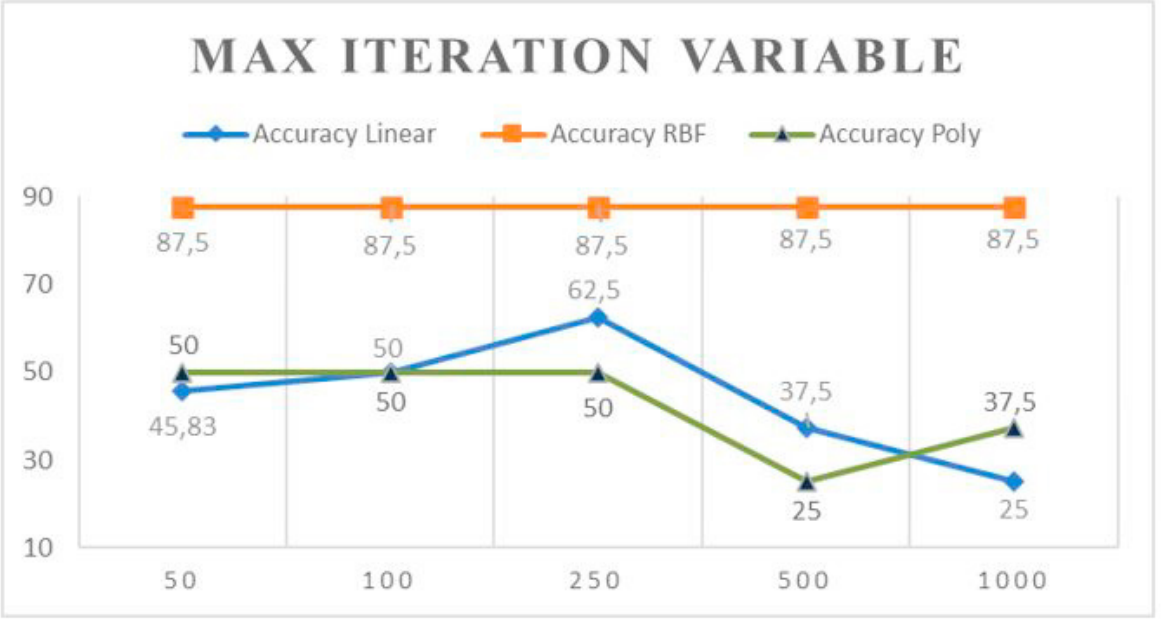

Fig. 5. Max Iteration Variable Accuracy Test Results

5. Conclusion

In this paper, classification of users' personality according to Big Five Personality traits by using the Support Vector Machine method has been trained and tested. The classification has been done using 170 Facebook user data. The method of classification is by made the Big Five Traits as the classes of the classifier, so it performs the Multi-

Class SVM. Work evaluation of this paper is to find the best accuracy that can be provided by tuning the parameters of SVM and comparing the results as the purpose of this research is to build the best Support Vector Machine to help classification of personality. After several test, it found that the parameters that fits best with the SVM model in this paper was using Radial Basis Function (RBF) Kernel, Complexity (C) = 1, Gamma = 0,1, and degree = 1 that provided the accuracy value 87,5%. Based on the results obtained from tests that have been done, The SVM model that has been built can be used in the future to predict new data classes by making the new data as the data test and tested using the data train along with the best parameters that have been obtained.

References

1. Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D.. Personality and patterns of facebook usage. In: *Proceedings of the 4th annual ACM web science conference*. 2012, p. 24–32.
2. Cobb-Clark, D.A., Schurer, S.. The stability of big-five personality traits. *Economics Letters* 2012;**115**(1):11–15.
3. Stankevich, M., Smirnov, I., Ignatiev, N., Grigoryev, O., Kiselnikova, N.. Analysis of big five personality traits by processing of social media users activity features. In: *DAMDID/RCDL*. 2018, p. 162–166.
4. Vaidhya, M., Shrestha, B., Sainju, B., Khaniya, K., Shakya, A.. Personality traits analysis from facebook data. In: *2017 21st International Computer Science and Engineering Conference (ICSEC)*. IEEE; 2017, p. 1–5.
5. Gunn, S.R., et al. Support vector machines for classification and regression. *ISIS technical report* 1998;**14**(1):5–16.
6. Tupes, E.C., Christal, R.E.. Recurrent personality factors based on trait ratings. *Journal of personality* 1992;**60**(2):225–251.
7. McCrae, R.R., John, O.P.. An introduction to the five-factor model and its applications. *Journal of personality* 1992;**60**(2):175–215.
8. Ryan, T., Xenos, S.. Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in human behavior* 2011;**27**(5):1658–1664.
9. Souri, A., Hosseinpour, S., Rahmani, A.M.. Personality classification based on profiles of social networks' users and the five-factor model of personality. *Human-centric Computing and Information Sciences* 2018;**8**(1):24.
10. Denissen, J.J., Geenen, R., Van Aken, M.A., Gosling, S.D., Potter, J.. Development and validation of a dutch translation of the big five inventory (bfi). *Journal of personality assessment* 2008;**90**(2):152–157.
11. Tadesse, M.M., Lin, H., Xu, B., Yang, L.. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access* 2018;**6**:61959–61969.
12. Guan, N.C., Yusoff, M.S.B.. Missing values in data analysis: Ignore or impute? *Education in Medicine Journal* 2011;**3**(1).
13. Ben-Gal, I.. Outlier detection. In: *Data mining and knowledge discovery handbook*. Springer; 2005, p. 131–146.
14. Vapnik, V.. *The nature of statistical learning theory*. Springer science & business media; 2013.