2017 16th IEEE International Conference on Machine Learning and Applications

ANALYZING FACEBOOK ACTIVITIES FOR PERSONALITY RECOGNITION

Laleh Asadzadeh
Department of Computer Science
Southern Illinois University
Carbondale, IL, 62901-6899
Email: asadzadeh@siu.edu

Shahram Rahimi
Department of Computer Science
Southern Illinois University
Carbondale, IL, 62901-6899
Email: rahimi@cs.siu.edu

***Abstract*—Facebook is the largest and the most popular online social network application that records large amount of users' behavior expressed in various activities such as Facebook likes, status updates, posts, comments, photos, tags and shares. One of the major attractions of such a dataset relates to the predictability of the individuals' psychological traits from their digital footprints. Such predictions help researchers and service providers to improve personalized offering of products and services. The goal of this work is to investigate the predictability of Facebook users' personality traits, measured by BIG5 test as a function of their digital records of behavior such as Facebook likes. This research is based on a dataset of 92,255 users who provided their Facebook likes and the results of their BIG5 personality test. As the Facebook likes data includes 600 attributes, the proposed model uses LASSO algorithm to select the best features and to predict Facebook users' BIG5 personality traits. The best accuracy level of these predictions is achieved for Openness and Extraversion, the lowest accuracy level is obtained for Agreeableness while the accuracy levels of Conscientiousness and Neuroticism are in the middle.**

Keywords*—Automatic Personality Recognition, BIG5 Personality Model, Facebook, LASSO.

I. INTRODUCTION

With the rapid growth of social networks millions of people are involved in expansive amount of comments, feelings, ideas, news, pictures, and videos. According to the Facebook newsroom, the total number of the Facebook daily active users was 1.038 billion in 2015, which makes this social network one of the best social data resources where researchers from various fields can reexamine theories or improve products and services in their fields of interest. For this to take place, the machine learning techniques are the most powerful tools to help researchers extract knowledge from the social data.

Theoretically, personality models predict "individual's characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms—hidden or not—behind those patterns" [1]. In addition to the natural curiosity of human beings about judgment of others' personality, research has shown that there is a significant relation between individuals personality traits and their behavior and preferences [2]. Therefore, personalized products and services compete over the fastest and most accurate automatic personality recognition solutions in social media. Automatic Personality Recognition (APR) consists of the automatic classification of individuals' personality traits, that can be compared against gold standard labels, obtained by the means of personality tests such as Myers Briggs Type Indicator (MBTI), Minnesota Multiphasic Personality Inventory (MMPI), and BIG5 [3]. Nowadays, the volume of the personal information that unintentionally is self-disclosed through the social media puts the automatic personality recognition problem on the map.

Being able to recognize personality traits of customers, improves the intelligent recommender systems such as "Tell me What I Need" (TWIN) [4]. APR also boosts ad targeting [5] while researchers have proven that people with high openness and low neuroticism responded more favorably to a targeted advertisement. Research shows that recognizing personality characteristics and their relations with the academic motivation helps in developing the more effective teaching strategies [6]. APR also improves the dating websites as it has been proven that there is a certain association between the personality traits and the relationship compatibility [7].

The most popular datasets researchers used to study the relationship between the personality traits and the digital footprints on Facebook, come from myPersonality Facebook application [1]. This application set up by David Stillwell at Psychometric Center of the Cambridge University in 2007 and provides access to various psychological tests and has attracted over six million users. In return, users may donate their Facebook profile information to research. This project results in datasess which are being shared with the academic community at myPersonality.org [2].

To automatically predict the personality traits of the Facebook users, researchers have chosen different approaches. Some researchers have approached APR as a classification problem and have applied classification methods such as Support Vector Machine (SVM) [8, 9, 10, 11], K Nearest Neighbor (kNN) [2, 10, 11], Naïve Bayes (NB) [2, 10, 11], Decision Trees (DT) [10], Sequential Minimal Optimization for Support Vector Machine (SMO) [12, 9], Bayesian Logistic Regression (BLR) [12, 10], Multinomial Naïve Bayes (MNB) [12], and Rule Learning, among others. However, other researchers have used various Multiple/Multivariate Linear Regression (MLR) to predict the personality traits scores of the Facebook users [13, 14]. Table I has summarized the most common machine learning algorithms that have been used for APR in Facebook.

[1] http://tests.e-psychometrics.com

[2] http://mypersonality.org

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.00-29

With the consideration of the above mentioned, the definition and applications of Automatic Personality Recognition in social network, the importance of Facebook as one of the most popular social networks, and various machine learning algorithms applied to Facebook data, this work is on the analysis of the Facebook likes using LASSO algorithm on a new myPersonality datasets to predict BIG5 personality traits of the Facebook users where their Facebook likes are described in 600 dimensions (600 features). Among all the Facebook features such as status updates, likes, image uploads, comments, number of friends, etc., Facebook likes is the most frequent activity users have. Facebook likes also cover a broad range of topics like politics, entertainment, shopping, etc.

Consequently, this work has chosen Facebook likes as the best Facebook feature that reveals users thoughts and preferences. Although, Kosinski et al. [15] have previousll applied LASSO algorithm on Facebook likes for APR, this work differs since the dataset utilized here has described a user preferences in 600 weighted topics while the dataset used in the previous work each users' likes in only 100 topics with 0/1 values. Therefore, the dataset used in this work is considerably more descriptive and the evaluations are more specific.

II. Literature Review

A. Automatic Personality Recognition (APR) in Social Media

Generally, researchers studying APR on social media investigate the correlation patterns between the personality and variety of the user's data captured from multiple sources. Several hypotheses have been raised regarding the relationship between personality and the Facebook profile features.

The first personality recognition work on social media was accomplished by Ross et al. in 2009 [16]. They observed several correlations between personality and the Facebook features, including positive correlation between extraversion and Facebook use, number of Facebook friends and associations with Facebook groups or negative correlation between conscientiousness and overall use of the Facebook. One limitation associated with this research was the small size of their homogeneous data sample (97 female students of the same major of the same university). Besides that, this study was based on the self-reports of the Facebook profiles the students made rather than direct observation. Therefore, this research only found one significant correlation between extraversion and group membership.

In 2011, Golbeck et al. [13] attempted to predict personality from Facebook profile information using machine learning algorithms but again their sample size was so small that limited the reliability and generalizability of their results however, they used very rich set of features (74 features for 167 records) such as words in status updates. Following these works the scientific community in personality recognition has grown rapidly [15, 17]. Although, the growing number of works in personality recognition in promoting, it is difficult to compare the reported performance and results, as almost each of these works perform their experiments on different datasets, and use different evaluation procedures [17].

B. MyPersonality Datasets in Literature

Recently, several researches have used myPersonality dataset for investigation of the relationship between the Facebook users' activities and their personality profiles [2, 13]. In 2011, Hagger-Johnson et al. [2] used the *interests* and *activities* sections of the 694 Facebook profiles to compare the users' personality and Sensational Interests Questionnaire (SIQ) scores to indicate the users' unusual violate interest. This study shows that Facebook users' *interests* and *activities* are valid indicators of sensational content. Bachrach et al. (2012) [13] have utilized myPersonality Facebook users' activities and have demonstrated that there is a significant negative correlation between neurotisicm and the number of friends. In addition, they have demonstrated that Agreeableness is positively correlated with the number of tags. Farnadi et al. (2013) [2] have used myPersonality Facebook users' status updates and their demographic profiles and reported several interesting observations of the Facebook profiles and personality scores. For instance, they observed that users with high extravertion and openness scores are more emotional in their status posts than those with high score of neuroticism. Most of the research projects on APR have utilized various machine learning algorithms. The next section discusses these algorithms.

C. Machine learning Algorithms applied to APR in Facebook

Researchers have different approaches to solve the APR problem in Facebook. Some researchers have utilized classification algorithms to assign five labels to the users displaying high/low level of the five personality traits from the BIG5 personality model. However, others have applied various versions of regression algorithms to predict the users' Big5 personality test scores. Clustering algorithms are mainly used as data preprocessing like feature selection. Table I summarizes the algorithms, datasets, performance evaluation methods, and accuracy of the related works.

III. METHODOLOGY AND IMPLEMENTATION

This section explains how the proposed model in this work has approached the problem of Automatic Personality Recognition (APR) in Facebook.

A. Description of the dataset

Among $3,137,694$ Facebook users who have taken the BIG5 personality test using myPersonality application, $92,225$ have allowed this application to record their psychological and Facebook profile information for anonymous use in research. BIG5 personality model describes the humans personality traits in 5 dimensions: *Openness to experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. It designates 5 numbers in $[1,5]$ to each user corresponding to each personality dimension. In this work, dataset X represents Facebook likes of $92,225$ users as a weighted combination of 600 topic sets. The Y dataset, on the other hand, describes the personalty of each user.

Table I. APR in Facebook: Support Vector Machine(SVM); Sequential Minimal Optimization for Support Vector Machine (SMO); Bayesian Logistic Regression (BLR); Multinomial Naïve Bayes (MNB); Naive Bayes (NB); k Nearest Neighbor (kNN); R: Square-root of the coefficient of determination; Decision Tree (DT); Logistic Regression(LR); Rule Learner (RIP)

Literature Review: Machine Learning Algorithms Applied to APR in Facebook					
Year	Author(s)	Data	#users	Algorithm(s)	Accuracy (avg.)
2011	Golbeck et al. [13]	Ego-Network	1100	Regression: M5	$MSE = 0.117$
2012	Celli [17]	Ego-Network, Status	23	Classification: SMO	$F - Score \simeq 0.56$
2013	Farnadi et al. [2]	Freq. of status updates	250	Classification: SVM, kNN, NB	$F - Score(SVM) \simeq 0.57$ $F - Score(kNN) \simeq 0.53$ $F - Score(NB) \simeq 0.5$
2013	Verhoeven et al. [8]	Status	10K	Classification: SVM	$F - Score \simeq 0.65$
2013	Markovikj et al. [9]	Status	250	Classification: SVM, SMO	$ROCArea \simeq 0.9$
2013	Alam et al. [12]	Status		Classification: SMO, BLR, MNB	$F - Score(SMO) \simeq 0.57$ $F - Score(BLR) \simeq 0.56$ $F - Score(MNB) \simeq 0.58$
2013	Appling et al. [14]	Status	250	Regression	$\lvert Correlation\rvert \simeq 0.05$
2013	Kosinski et al. [15]	like	58K	Regression	$Correlation \simeq 0.34$
2013	Schwartz et al. [20]	Status	71,968	Regression	$R \simeq 0.37$
2013	Tomlinson et al. [21]	Status	250	Regression	(Conscientiousness only) $Correlation \simeq 0.26$ $RMSE \simeq 0.74$
2014	Celli et al. [10]	Profile images	100K	Classification: NB, SVM, DT, LR, kNN, RIP	$F - Score(NB) \simeq 0.61$ $F - Score(SVM) \simeq 0.7$ $F - Score(LR) \simeq 0.62$ $F - Score(kNN) \simeq 0.73$ $F - Score(RIP) \simeq 0.61$
2014	Farnadi et al. [11]	Status, Ego-Network, Freq. of status updates	250	Classification: SVM, kNN, NB	$F - Score \simeq 0.63$
2014	Eftekhar et al. [16]	photo-related activities	115	Regression	$R^2 \simeq 0.06$
2015	Peng et al. [18]	Status, #Friends	222	Classification: SVM	$Precision \simeq 0.76$
2015	Ghavami et al. [22]	like	65	Classification: SVM, kNN, DT	$F - Score(SVM) \simeq 0.54$ $F - Score(kNN) \simeq 0.51$ $F - Score(DT) \simeq 0.38$
2016	Wang et al. [19]	Status	250	Classification: probability-based predicting model	$F - Score \simeq 0.8$

B. Structure of the model

The presented model considers Automatic Personality Recognition as a regression problem that receives the Facebook likes of the users and predicts their BIG5 personality trait scores. The proposed model has training/learning phase, hyper-parameter tuning phase, and prediction/test phase. The input of the model in the training phase is 75% of the data for both X (X_{Train}) and Y (Y_{Train}). The remainder of X, X_{Test}, is used as the input of the prediction phase and the remainder of Y, Y_{Test}, is compared with the output of the prediction phase to estimate the accuracy of the model. This model uses *Cross-Validation* technique to evaluate the results.

C. Least Absolute Shrinkage and Selection Operator Algorithm (LASSO)

LASSO is a *regression* model that performs both *feature selection* and *regularization* to improve the accuracy and interpretability of the model. Suppose $X = \{X_1, X_2, ..., X_m\} \subset \Re^n$ is the set of input objects, $Y = \{y_1, y_2, ..., y_m\} \subset \Re$ is the set of actual values associated with X, where for $1 \leq i \leq m$, $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$, and $y_i \in \Re$. LASSO estimates $(\hat{\beta_0}, \hat{\beta}) \in (\Re, \Re^n)$, where $\hat{\beta} = (\beta_1, \beta_2, ..., \beta_n)$, such that:

$$(\hat{\beta_0}, \hat{\beta}) = \mathbf{argmin}\left\{\sum_{\mathbf{i=1}}^{\mathbf{m}}(\mathbf{Y_i} - \beta_{\mathbf{0}} - \sum_{\mathbf{j=1}}^{\mathbf{n}}\beta_{\mathbf{j}}\mathbf{x_{ij}})^{\mathbf{2}}\right\} \qquad (1)$$

subject to $\sum_{j=1}^{n} |\beta_j| \leq k$

Here $k \in \Re$ is the parameter which defines the coefficient of regularization and should be tuned to optimize the accuracy of the model. When $\mathbb{X}$ is standardized one can rewrite Eq.(1) as follows:

$$\hat{\beta} = \mathbf{argmin}\left\{\sum_{\mathbf{i=1}}^{\mathbf{m}}(\mathbf{Y_i} - \sum_{\mathbf{j=1}}^{\mathbf{n}}\beta_{\mathbf{j}}\mathbf{x_{ij}})^{\mathbf{2}}\right\} + \lambda\sum_{\mathbf{j=1}}^{\mathbf{n}}|\beta_{\mathbf{j}}| \qquad (2)$$

Large value of the coefficient $\lambda \in \Re$ causes some of the coefficients β_j be 0. Intuitively, choosing λ is similar to choosing the number of the features of the input data (predictors). *Cross validation* is a good tool for estimating optimal value of λ.

The prediction phase of the model challenges the model by predicting the Y_p values corresponding to the X_{Test} and comparing them with the actual values Y_a in Y_{Test}. This phase evaluates the accuracy of the model.

D. Accuracy Metrics

Two famous metrics for evaluating the accuracy of the regression models are *Mean Squared Error (MSE)*, Eq.(3) , and *Pearson Correlation Coefficient*, Eq.(4). MSE calculates the mean distance between the actual scores of the users' personality profiles Y_a and the predicted scores Y_p.

$$\mathbf{MSE(Y_p, Y_a) = \frac{\sum_{k=1}^{m}(y_{pk} - y_{ak})^2}{m}} \tag{3}$$

Pearson Correlation Coefficient measures the linear correlation between the predicted value Y_p and the actual value of Y_a.

$$\mathbf{r(Y_a, Y_p) = \frac{\sum_{i=1}^{m}(y_{ai} - \overline{y_a})(y_{pi} - \overline{y_p})}{\sqrt{\sum_{i=1}^{m}(y_{ai} - \overline{y_a})^2}\sqrt{\sum_{i=1}^{m}(y_{ai} - \overline{y_p})^2}}} \tag{4}$$

IV. Implementation and Results

This project is implemented in R due to powerful R packages in statistical analysis, machine learning, and data mining. After standardizing the data, our model receives X_{Train}, Y_{Train} to train the model and fit the $\hat{\beta}$ values that minimize the value of error in Eq.(2) corresponding to a sequence of λ values.

To find the optimized value of the hyper-parameter λ in LASSO model, the *cross validation* method has been used. Fig. 3 illustrates the *10-fold cross validation* of the model and compares the value of the error for each λ. This figure illustrates the minimum and the maximum values of λ that gives the least value of error as well as the number of active variables (predictors). For example, if $log(\lambda) = -5$, the number of the active predictors is 479. The best value of λ for the fitted model is 0.001 and $MSE \simeq 0.16$. The correlation of the predicted values and the Y_{Test} has been shown in table II that can be compared with the results of Kosinski et al. [15] where Pearson Correlation Coefficient (r) of "Openness" (r = 0.43), "Extraversion" (r = 0.40), and the remaining personality traits were predicted with somewhat lower accuracy (r = 0.17 to 0.30).

According to Eq.(2), large values of λ results in more zero coefficients of β_j, and therefore, fewer the number of the predictors. Fig. 1 illustrates the number of the non-zero coefficients $\hat{\beta}$ against λ. Fig. 2 represents the value of the coefficients as the function of the L1 Norm (sum of the absolute values of the coefficients).

Table II. Accuracy of the model based on LASSO algorithm, $min(\lambda) = 0.001$

Accuracy of LASSO model		
Personality Trait	Pearson Correlation Coefficient	Mean Squared Error(MSE)
Openness	0.38	0.024
Conscientiousness	0.29	0.030
Extraversion	0.34	0.038
Agreeableness	0.22	0.030
Neuroticism	0.27	0.038

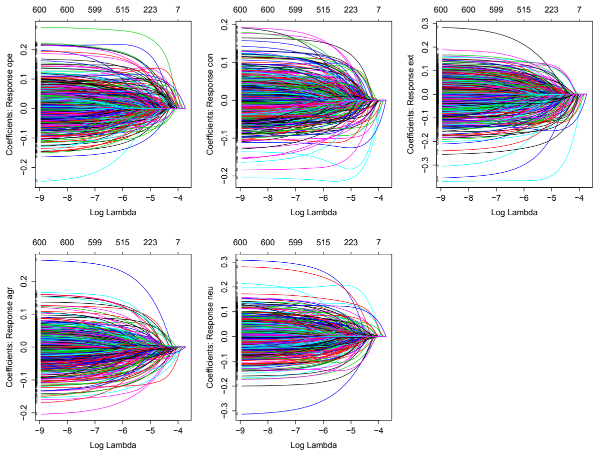

Figure 1. The number of the non-zero coefficients decreases as the value of lambda increases.

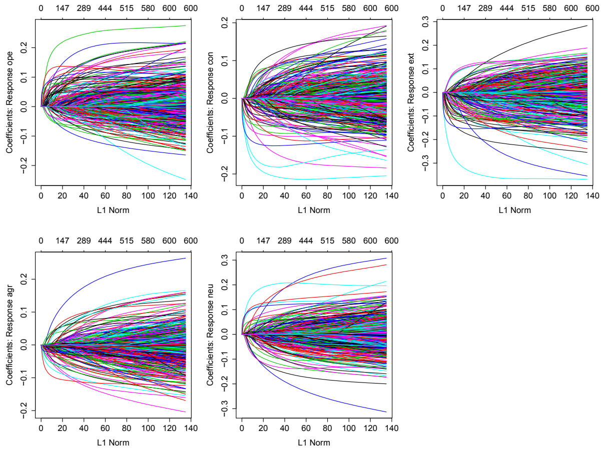

Figure 2. Y coefficients as function of the λ

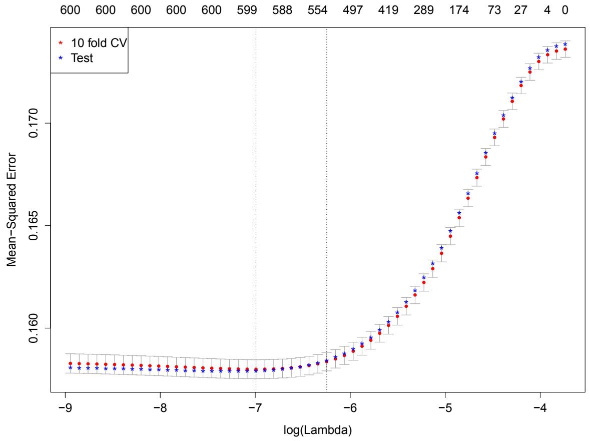

Figure 3. The value of the coefficients as the function of the L1 Norm of the coefficients

V. CONCLUSION AND FUTURE WORK

The goal of this project was providing a predictive model that makes valid judgments about personality traits of the social network users based on their activities. This model utilizes LASSO algorithm as one of the most accurate algorithms in the literature. The presented model can be extended to study and reexamine theories from social sciences such as economics, sociology, psychology, law, etc.. The recommender systems can also take advantage of this model to improve their services and products by adding psychological profiles of the users to their activity profiles and provide better recommendations. The recommender systems specifically benefit from the psychological profiles of the new users since the lack of the history of the activities of the new users causes the computer models not to be able to predict the preferences of these users accurately. One important future work is to improve our model by using fuzzy logic. In such model we would convert each personality trait score to actual terms in natural language. Consequently, the outcome of the model would be more informative.

ACKNOWLEDGMENT

The authors would like to thank David Stillwell, Michal Kosinski, and the myPersonality project (mypersonality.org) for providing the data used in this work.

REFERENCES

[1] Bandura, A. (2001). *Social cognitive theory: An agentic perspective* . Annual review of psychology, 52(1), 1-26.

[2] Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013, January). *Recognising personality traits using facebook status updates.* InProceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13). AAAI.

[3] McCrae, R. R., & John, O. P. (1992). *An introduction to the five-factor model and its applications* . Journal of personality, 60(2), 175-215.

[4] Roshchina, A., Cardiff, J., & Rosso, P. (2015). *TWIN: Personality-based Intelligent Recommender System* . Journal of Intelligent & Fuzzy Systems,28(5), 2059-2071.

[5] Chen, J., Haber, E., Kang, R., Hsieh, G., & Mahmud, J. (2015, April). *Making use of derived personality: The case of social media ad targeting* . InNinth International AAAI Conference on Web and Social Media.

[6] Komarraju, M., & Karau, S. J. (2005). *The relationship between the big five personality traits and academic motivation.* Personality and individual differences, 39(3), 557-567.

[7] Donnellan, M. B., Conger, R. D., & Bryant, C. M. (2004). *The Big Five and enduring marriages..* . Journal of Research in Personality, 38(5), 481-504.

[8] Verhoeven, B., Daelemans, W., & De Smedt, T. (2013). *Ensemble methods for personality recognition.* Proceedings of WCPR13, in conjunction with ICWSM-13.

[9] Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013, June). *Mining facebook data for predictive personality modeling.* In Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA.

[10] Celli, F., Bruni, E., & Lepri, B. (2014, November). *Automatic personality and interaction style recognition from facebook profile pictures.* In Proceedings of the ACM International Conference on Multimedia (pp. 1101-1104). ACM.

[11] Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013, January). *How well do your facebook status updates express your personality.* InProceedings of the 22nd Edition of the Annual Belgian-Dutch Conference on Machine Learning, BENELEARN.

[12] Alam, F., Stepanov, E. A., & Riccardi, G. (2013, June). *Personality traits recognition on social network-facebook.* In Proceedings of the Workshop on Computational Personality Recognition (pp. 6-9).

[13] Golbeck, J., Robles, C., & Turner, K. (2011, May). *Predicting personality with social media.* In CHI'11 extended abstracts on human factors in computing systems (pp. 253-262). ACM.

[14] Appling, D. S., Briscoe, E. J., Hayes, H., & Mappus, R. L. (2013, June). *Towards automated personality identification using speech acts.* In Seventh International AAAI Conference on Weblogs and Social Media.

[15] Kosinski, M., Stillwell, D., & Graepel, T. (2013). *Private traits and attributes are predictable from digital records of human behavior.* Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

[16] Eftekhar, A., Fullwood, C., & Morris, N. (2014). *Capturing personality from Facebook photos and photo-related activities: How much exposure do you need?.* Computers in Human Behavior, 37, 162-170.

[17] Celli, F. (2012). *Adaptive Personality Recognition from Text* (Doctoral dissertation, University of Trento).

[18] Peng, K. H., Liou, L. H., Chang, C. S., & Lee, D. S. (2015, October). *Predicting personality traits of Chinese users based on Facebook wall posts.* In Wireless and Optical Communication Conference (WOCC), 2015 24th (pp. 9-14). IEEE.

[19] Wang, M., Zuo, W., & Wang, Y. (2015). *A Novel Adaptive Conditional Probability-Based Predicting Model for User's Personality Traits.* Mathematical Problems in Engineering, 2015.

[20] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). *Personality, gender, and age in the language of social media: The open-vocabulary approach.* PloS one, 8(9), e73791.

[21] Tomlinson, M. T., Hinote, D., & Bracewell, D. B. (2013, July). *Predicting conscientiousness through semantic analysis of facebook posts.* In Proceedings of the Workshop on Computational Personality Recognition (pp. 31-34).

[22] Ghavami, S. M., Asadpour, M., Hatami, J., & Mahdavi, M. (2015, May). *Facebook user's like behavior can reveal personality.* In Information and Knowledge Technology (IKT), 2015 7th Conference on (pp. 1-3). IEEE.