ICIC Express Letters
Volume **13**, Number **9**, September **2019**

ICIC International ©2019 ISSN 1881-803X
pp. 797–805

AUTOMATIC PERSONALITY RECOGNITION IN BAHASA INDONESIA: A SEMI-SUPERVISED APPROACH

Gabriel Yakub Novian Nugroho Adi[1], Michael Harley[1], Veronica Ong[1]
Derwin Suhartono[1] and Esther Widhi Andangsari[2]

[1]Computer Science Department
School of Computer Science
[2]Psychology Department
Faculty of Humanities
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ gabriel.adi; michael.tandio; veronica.ong }@binus.ac.id; { dsuhartono; esther }@binus.edu

Received February 2019; accepted May 2019

Abstract. *The knowledge of knowing peoples' personality is found to be useful for multiple purposes, including in the field of commercial and social studies. Recent studies of automatically assigning a personal profile to the users of social media have emerged within the last decade as using personality measuring instruments was considered impractical in the online context. Though being a part of a massive amount of activities in social media, only a few studies have reviewed the approach of automatically recognizing the Bahasa Indonesia speaking users. Previous studies indicated that dataset remains a serious limitation in the field of study, where obtaining a large amount of labelled data is expensive and time-consuming. With that in mind, we present a practical and effective approach to utilize the abundance of unlabeled data in a semi-supervised learning setting using the label propagation model. With the highest ROC AUC score of 0.93 as achieved by the Super Learner algorithm, the presented method could be considered as a sensible approach as a means of overcoming the dataset limitation.*

Keywords: Semi-supervised learning, Automatic personality recognition, The Big Five Personality Model, Bahasa Indonesia, Super Learner

1. **Introduction.** Over the last decade, social network penetration worldwide is ever increasing. In 2018, statistics displayed that 71% of Internet users are social media users[1]. The number is expected to grow for the next following years. Facebook, YouTube, WhatsApp, Instagram, and Twitter are among the most popular social media platform worldwide. Twitter alone has a considerable amount of 326 monthly active users throughout multiple countries in the world per 2018.

As the most popular online activities in the world, social media involves users to share their thoughts, opinions, and feelings through their posts on the platform. Through these activities, users generated a massive amount of contents in forms of texts, photos, sounds, and videos. Many researchers harness the user-generated data to automatically infer their personality in various approaches on multiple environments based on personality traits [1]. This type of study is typically referred to as Automatic Personality Recognition (APR).

One of the most popular forms of inferring the users' personality is through analyzing the users' writings as there exist strong correlations between human personality and their linguistic cues [2,3]. For example, researchers have developed a variety of approaches to automatically infer the users' personality using data from the *myPersonality* project [4-6], a dataset containing personality data from thousands of users from Facebook [7]. The

DOI: 10.24507/icicel.13.09.797

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

knowledge of individuals' personality provides insights that could be used for multiple purposes. These include making preferences predictions and enhancing recommendation systems [8], predicting stock market [9], and detecting real-time events [10]. The exploration of APR continues as researchers attempted APR in multiple languages, some of which are English [11,12], Chinese [13], and Tagalog [14].

Previous studies have been done to develop approaches on APR in Bahasa Indonesia [15-17]. However, there are limitations when performing such studies. One primary blocker in the field of the study is the dataset limitation. Depending on the context, datasets available to perform APR in Bahasa Indonesia are extremely limited and kept in private. As the nature of the available dataset requires text mining from social media and manual labelling from psychological experts [15,18], the process of expanding the dataset to contain a large amount of labelled users' data is expensive and time-consuming.

The focus of the study is to perform APR in Bahasa Indonesia. The study presents an implementation and evaluation of a semi-supervised approach to overcome the stated dataset limitation of performing APR in Bahasa Indonesia. We present the approach of utilizing the label propagation model to combine the use of both labelled data and additional unlabeled data in a semi-supervised learning setting.

We begin by briefly introducing the Big Five Personality Model and its personality measuring instruments. Equally important, we present related works and the existing trends on building an APR system. Furthermore, we explain the steps of implementing APR to predict the personality of Bahasa Indonesia speaking users. We present and discuss the classifiers performance on predicting the users' personality in a semi-supervised setting. Finally, we perform a comparative analysis on scenarios to prove whether or not satisfying a semi-supervised characteristic in regards of the dataset distribution helps the classifiers' performance to predict the user's personality.

2. Literature Review.

2.1. The Big Five Personality Model.

The Big Five, which is also commonly referred as Five-Factor Model, is the most widely accepted and used model to assign a personal profile to an individual [1,19]. As the name implies, the Big Five Personality Model separates human personality into five different dimensions [1,20,21]: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness to Experience.

Standardized instruments are developed to measure an individual's personality based on the Big Five Personality Model. The most comprehensive measuring instrument is the *NEO Personality Inventory, Revised* (NEO-PI-R) which comprised 240 questions [22]. The instrument, however, was considered too long. Since then, many studies devised newer and shorter measuring instruments to infer a person's personality based on the personality model. The most common measuring device used is the Big-Five Inventory (BFI) which is built upon 44 questions, and NEO Five-Factor Inventory (NEO-FFI) which comprised 50 items [23].

Methods of measuring an individual's personality using the stated measuring instruments are considered as impractical and time-consuming especially in the online context. People may be reluctant to fill lengthy questionnaires to get recommendation and personalization in online activities [1]. Therefore, new methods that automatically predict people's personality are studied among the researchers.

2.2. Automatic personality recognition and related works.

APR is a field of study that aims to automatically assign a personality profile to an individual from their digital footprints [4,11]. [25] pioneered the study of inferring individual's personality on social media. They examine the correlation between Facebook usage and the Big Five Personality Model. Although some correlations were found to be negative, they revealed trends between the Extraversion and Openness to Experience dimensions and the usage of Facebook.

Several approaches of APR have been conducted on different platforms [16]. One of the most commonly used methods is by utilizing text analysis tool Linguistic Inquiry Word Count (LIWC). The tool enables researchers to extract information from text using the pre-defined dictionaries in different categories encompassing linguistic dimensions, psychological processes, and grammatical categories [2,26]. For example, [6] utilized language features that are extracted using LIWC, social network features, time-related features, and content-based features to devise an approach of predicting personalities of 250 Facebook users of the *myPersonality* project. The use of *myPersonality* project on APR studies remains popular as many researchers utilized the data collected from the project [4,6,11,27].

The usage of user-generated contents from Twitter is also valuable from predicting the personality of individuals. The popular dataset of Twitter is derived from the *myPersonality* project as hundreds of participants link their Twitter account to their posts [1]. Using the dataset, [12] devised an approach that employed *Gaussian Process* and *ZeroR* to predict Twitter users' personality using language features produced by LIWC and M-RC psycholinguistic database along with additional features of Twitter use and sentiment analysis. Meanwhile, [28] attempted user behavioral analysis of 335 Twitter users of the *myPersonality* project participant.

In spite of the existing trends of utilizing text analysis tools, e.g., LIWC and MRC, and well-established datasets, in particular, the *myPersonality* project, the same method cannot be applied in the case of studies on APR in Bahasa Indonesia as it only supports a selected set of languages. Therefore, researchers who studied APR in Bahasa Indonesia used open-vocabulary approaches which centered to analyze the users' choice of words. [17] applied text tokenization, stop-words removal, and stemming on the users' writing and weighed each token using the *tf-idf* weighting scheme. With that in mind, they assumed that the meaning of every word remains accurate after being translated. The similar open-vocabulary approach was applied by [18]. The results showed that building an accurate APR system for Bahasa Indonesia is achievable without the use of predefined dictionaries (i.e., LIWC, MRC).

Furthermore, recent studies have also devised semi-supervised approaches to build APR models. [29] proposed an approach called PERSOMA (PERsonality prediction in SOcial Media datA) which was built upon three main modules. On the other hand, [30] conducted a semi-supervised APR approach to predict Microblog users' personality using a Local Linear Semi-Supervised Regression (LLSSR). From users' public information, they extracted 47 features that could be divided into several categories, namely the user's personal profile, the user's social circle, the user's social activities, and the user's social habit.

3. **Research Methodology.** We started the research by developing an updated dataset in which we combine both labelled and unlabeled data into a single union. Afterwards, we perform text preprocessing to reduce noises and to prepare the dataset to be able to be fed into the machine learning system. Subsequently, we implemented the label propagation model to classify (label) the unlabeled data by utilizing the labelled data. Finally, with all data set up, we perform the classification process in which we utilize two machine learning algorithms, namely Stochastic Gradient Descent and Super Learner. The research was built on *python* with the *scikit-learn* [31] and other numerous libraries. Figure 1 displays an overview of the research process.

3.1. **Data collection.** We utilized and expanded the Twitter dataset in Bahasa Indonesia as developed by [18] that comprises the data of 250 Twitter users. First, adopting the

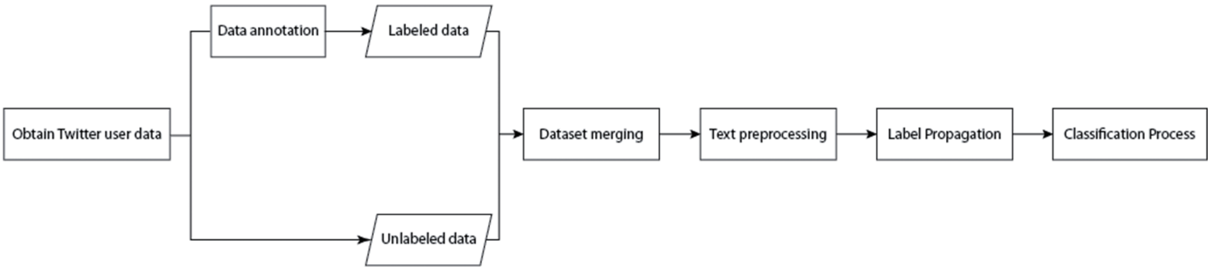

Figure 1. The research methodology overview

Twitter Use Information		
Tweet Count	Favorites Count	Quote Count
Followers Count	Retweet Count	Mention Count
Following Count	Retweeted Count	Reply Count
Hashtag Count	URL Count	

Figure 2. The final set of Twitter usage information extracted in the data collection stage

exact same steps as Ong et al., we added additional data to the dataset by retrieving the data of 258 Indonesian Twitter users. For each user, we obtained the latest 150 *tweets* along with a set of the user's Twitter usage information as displayed in Figure 2. Following the previous studies [16,32], the Twitter usage information set consists of the data in Figure 2.

Afterwards, we formatted the extracted data into an Excel file and sent to the same set of psychology experts as stated in the said study. We, then, continued to extract Indonesian Twitter data in which we managed to retrieve additional 450 users' data. Due to the psychology experts' time constraints, we were unable to make further annotation request. With the labelled and unlabeled data setup, we merged them into a union and converted them into a *Python* dataset. After merging, our final dataset comprises the data of 958 Indonesian Twitter users where 508 and 450 users are labelled and unlabeled consecutively.

3.2. **Data preprocessing.** We developed a preprocessing program to reduce noises of the existing dataset. Adopting and modifying the preprocess steps taken from previous studies [15,17,18], the program involves several preprocessing steps which are stated as follows.

1) **Replacing hashtags and mentions with tokens**
   The program replaces hashtags with the "[HASHTAG]" token, whereas mentions are replaced with "[UNAME]".
2) **Removing URL and emojis**
   Using pre-defined regular expressions, any URLs and emojis are omitted from the text.
3) **Removing stop words**
   The program filters stop words in Bahasa Indonesia. The list of Bahasa Indonesia's stop words is taken from [33].
4) **Tokenizing and converting text to *n*-gram range**
   For each of the users, the program tokenizes the user's writing into a collection of a single word. Afterwards, the tokenization result is converted to an *n*-gram range by combining unigram and bigram.
5) **Applying the term-frequency weighting scheme**

As we planned to implement APR using open-vocabulary approach, we weighed each element of the n-gram range collection using the normalized term-frequency weighting scheme. The normalized term-frequency weighting scheme, commonly abbreviated as *tf*, uses the function (1) to weigh each element.

$$tf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \tag{1}$$

$tf(t,d)$ = the weight of term t given document d; $f_{t,d}$ = the number of occurrences of term t in document d; $\max\{f_{t',d} : t' \in d\}$ = the raw frequency of the most occurring term in document d.

As the function displays, each term-frequency weight is scaled and translated such that the maximum value of each feature in the data is set to be 1.0.

6) **Feature selection**

The problem with an open-vocabulary approach, where we utilize the use of n-gram and *tf* weighting scheme, is that it generates a massive amount of features [15,17]. To reduce noises in the generated features as a means of improving the effectiveness of learning, we only took the top 1000 highest scoring features based on the computed ANOVA F-values.

3.3. **Label propagation.** We adopted the label propagation model as devised by [34] to estimate labels for each unlabeled data point using the labelled data. First, we denote l and u as the quantity of labelled and unlabeled data points in that order. Let $X_o = (x_1, y_1), \ldots, (x_l, y_l)$ be a collection of observed data, where $Y_o = \{y_1, \ldots, y_n\}$ is the pre-defined class for each of the observed data points. We refer X_o as the labelled data points since pre-defined classes are known for each element of the data collection. Then, let $(x_{n+1}, y_{n+1}), \ldots, (x_{l+u}, y_{l+u})$ be a collection of unlabeled data where $Y_u = \{y_{n+1}, \ldots, y_{n+u}\}$ is a set of classes that remain unobserved. Typically, the unlabeled data points should be larger in quantity than the labelled data points in. Therefore, we assume $l < u$. The goal of this model is to apply transductive inference that estimates Y_u from X_o and Y_o.

The model set up a fully-connected graph in which labelled and unlabeled data are connected. Afterwards, it defines the edge weight w_{ij}, where ij is an edge that connects node i to node j. The calculation of the edge weight w_{ij}, which in this case is based on the local Euclidean distance, is defined as (2).

$$w_{ij} = \exp\left(-\frac{d_{i,j}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_{d=1}^{D}\left(x_i^d - x_j^d\right)^2}{\sigma^2}\right) \tag{2}$$

w_{ij} = the weight of edge ij that connects node i to j; σ = the control parameter; d_{ij}^2 = the Euclidian distance from node i to j.

It is important to be noted that each node has a soft label which could be propagated to other closely located nodes. The higher the edge's weight of two connected nodes, the more probable that it would transition a label from a node to the other. The probabilistic label transition matrix, which we refer as matrix T whose dimension is $(l+u) \times (l+u)$ could be defined as:

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}} \tag{3}$$

T_{ij} = the transition probability to jump from node j to i; w_{ij} = weight of an edge that connects node i to j; $\sum_{k=1}^{l+u} w_{kj}$ = total weight of edges in the transition matrix T.

The model also defines label matrix $Y = (l+u) \times C$ with arbitrary initialization. The i-th row of the matrix represents the label probability distribution of node x_i. Finally, the algorithm proceeds with the following main steps.

1) **Propagate $\boldsymbol{Y} \leftarrow \boldsymbol{TY}$**
   In this step, each node propagates its labels for one step. Self-loop is included in the propagation process.
2) **Row-normalize $\boldsymbol{Y}$**
   The model applies row-normalization to matrix Y as a means of maintaining the label probability interpretation.
3) **Apply soft clamping to the initially labelled data**
   The last step applies to clamping to the nodes whose labels are initially defined. The stage aims to persist the initially labelled nodes that 'fade away' in the label propagation process.
4) **Repeat from 1 until $\boldsymbol{Y}$ converges**
   We utilized the label propagation model of *scikit-learn* to implement the explained algorithm. First, we define a specific class "-1" for each unlabeled data point. The class "-1" represents Y_u, a set of unobserved class that will be estimated by employing labelled data points. Next, we merged the labelled and unlabeled data into a single collection which could be denoted as $(x_1, y_1), \ldots, (x_{l+u}, y_{l+u})$. Finally, we run the label propagation model on the obtained collection of data to propagate the labels of the labelled data to every unlabeled data point.

3.4. **Classification process.** The classification process in this APR study is a multi-label classification as the personality model consists of multiple traits. Meaning, in the context of the Big Five Personality Model, each person has five labels for five personality traits as defined in the model. We transform each label into a binary form where a person scored as "HIGH" in a particular trait is denoted as "1", whereas "0" is denoted for the opposite.

With the classification case defined, we built one binary classifier for each trait of the Big Five Personality Model. Adapting from a previous study [15], the classifiers were built on two algorithms namely: Stochastic Gradient Descent (SGD), and Super Learner [35]. The two classifiers were proven to have a strong performance when solving problems that utilizes similar dataset.

As it may be noticed, our dataset labelled, and unlabeled distribution does not comply with the characteristic of semi-supervised learning. A semi-supervised learning characteristic stated that, typically, the unlabeled is assumed to be a lot more in quantity as compared to the labelled data points [34,36]. However, we only managed to get 450 unlabeled data which is lower in quantity than the 508 labelled data points. For that reason, we attempted a scenario where we randomly sampled down the labelled data points to significantly reduce the number and satisfy the stated semi-supervised learning characteristic. Overall, scenarios attempted in this study are as follows.

1) **Without undersampling**
   We utilized all data collected, in a sense, the said semi-supervised characteristic is not satisfied. In this scenario, 508 labelled and 450 unlabeled data points were put into the machine learning training and testing scheme.
2) **With undersampling**
   We undersampled to significantly reduce the labelled data points to an amount of 100 data, meaning, 100 and 450 labelled and unlabelled data were utilized in the scenario. This scenario was attempted as a means of complying the stated characteristic of semi-supervised learning.

We measure the performance of each classifier by calculating the ROC AUC score. ROC AUC is defined as the measurement of the area under the ROC curve [37]. First, we define the ROC curve by plotting True Positive Rate (TPR) against False Positive Rate (FPR) at multiple thresholds to cover the low and high values. Finally, the area under the ROC curve is measured to get the ROC AUC score.

4. **Results and Discussion.** In this section, we present and discuss the classifiers' performances on predicting users' personality in a semi-supervised approach. Before any training and testing scheme, we implemented semi-supervised model by classifying the labels of all unlabeled data using the label propagation model such that, every unlabeled data point that is denoted as the class "-1" will be transformed into either the "1" (HIGH) or the "0" (LOW) class. Afterwards, with all data classified, we performed k-fold cross-validation scheme with $k = 10$ and reported the evaluation performance score of all classifiers on the previously mentioned scenarios.

Table 1 displays the classifiers' performance where all data are utilized. Meanwhile, Table 2 shows the classifiers' performance of the scenario where we sampled down the labelled data to 100. Thoroughly, based on an ROC AUC interpretation score as presented by [38], the semi-supervised learning developed in the study successfully achieved good to excellent performances with the highest average score of 0.93 ROC AUC. The label propagation model implemented in study was able to capture the data distribution and classify unlabelled data as the machine learning performance shows a compelling performance on predicting the users' personality. Moreover, the result indicates that by satisfying the semi-supervised learning characteristic, the performance of both classifier algorithms was successfully improved. Though not significantly, undersampling the labelled data helped to reduce the probability of the occurrence of mislabeled training data. This, in turn, helps the label propagation model to use less noisy labelled data points to estimate and generalize the unobserved data better. For that reason, the ability of each classifier to predict the users' personality was boosted.

Table 1. ROC AUC achieved by each algorithm in the scenario where 508 labelled and 450 unlabeled were used

Classifier	**ROC AUC**					
	AGR	**CON**	**EXT**	**NEU**	**OPN**	**Average**
Super Learner	0.853	0.861	0.914	0.875	0.824	0.8654
SGDC	0.799	0.859	0.856	0.834	0.808	0.8312

Table 2. ROC AUC achieved by each algorithm in the scenario where the labelled data were sampled down to 100 users

Classifier	**ROC AUC**					
	AGR	**CON**	**EXT**	**NEU**	**OPN**	**Average**
Super Learner	0.937	0.935	0.929	0.944	0.913	0.9316
SGDC	0.905	0.919	0.908	0.944	0.855	0.9062

On the other hand, the result shows that the performance of an ensemble learning that is implemented using the Super Learner classifier is more superior as compared to a more traditional machine learning algorithm, Stochastic Gradient Descent (SGD). The Super Learner managed to get the highest average ROC AUC score of 0.931 in the undersampling scenario. Though not significantly better as compared to SGD, the use of an ensemble learning is still preferable as compared to a more traditional model. This finding remains consistent with the results presented in the previous study [15], where various supervised learning approaches were implemented and evaluated.

5. **Conclusions.** In conclusion, we successfully devised an effective and practical approach of semi-supervised learning in APR for Bahasa Indonesia. Evaluation results from the iterations of 10-fold cross-validation show that the prediction model that was built on Super Learner managed achieved the highest ROC AUC score of 0.93, while Stochastic

Gradient Descent Classifier achieved 0.90 ROC AUC. Though not significantly better, satisfying the stated semi-supervised approach characteristic was proven to improve the performance of the APR system to predict users' personality from data in Bahasa Indonesia as compared to the scenario where the characteristic was not satisfied. When labelled data are expensive and time-consuming to obtain, utilizing the abundance of unlabeled data in a semi-supervised setting using the label propagation model is a sensible and practical option to be performed.

Further exploration could be done with unsupervised learning approach as labelled data are scarce and expensive. Future studies could also devise a closed-vocabulary approach for APR in Bahasa Indonesia where text analysis tools are only available for selected languages. Closed-vocabulary methods allow more in-depth analysis to be studied as it considers the meaning of each word of the users' choice.

REFERENCES

[1] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. Moens and M. D. Cock, Computational personality recognition in social media, *User Modeling and User-Adapted Interaction*, vol.26, nos.2-3, pp.109-142, 2016.

[2] Y. R. Tausczik and J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, *Journal of Language and Social Psychology*, vol.29, no.1, pp.24-54, 2010.

[3] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar and M. E. P. Seligman, Automatic personality assessment through social media language, *Journal of Personality and Social Psychology*, vol.108, no.6, pp.934-952, 2014.

[4] K. C. Pramodh and Y. Vijayalata, Automatic personality recognition of authors using big five factor model, *2016 IEEE International Conference on Advances in Computer Applications (ICACA 2016)*, 2017.

[5] Y. Liu, J. Wang and Y. Jiang, PT-LDA: A latent variable model to predict personality traits of social network users, *Neurocomputing*, vol.210, pp.155-163, 2016.

[6] G. Farnadi, S. Zoghbi, M.-F. Moens and M. De Cock, Recognising personality traits using Facebook status updates, *Proc. of the Workshop on Computational Personality Recognition (WCPR13) at the 7th International AAAI Conference on Weblogs and Social Media (ICWSM13)*, 2013.

[7] D. J. Stillwell and M. Kosinski, myPersonality Project Website, *myPersonality Project*, 2015.

[8] R. Lambiotte and M. Kosinski, Tracking the digital footprints of personality, *Proc. of the IEEE*, vol.102, no.12, pp.1934-1939, 2014.

[9] J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, *Journal of Computational Science*, vol.2, no.1, pp.1-8, 2011.

[10] T. Sakaki, M. Okazaki and Y. Matsuo, Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proc. of the 19th International Conference World Wide Web*, pp.851-860, 2010.

[11] A. Laleh and R. Shahram, Analyzing facebook activities for personality recognition, *Proc. of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[12] J. Golbeck, C. Robles, M. Edmondson and K. Turner, Predicting personality from Twitter, *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd Inernational Conference on Social Computing (SocialCom)*, pp.149-156, 2011.

[13] K. H. Peng, L. H. Liou, C. S. Chang and D. S. Lee, Predicting personality traits of Chinese users based on Facebook wall posts, *The 24th Wireless and Optical Communication Conference (WOCC)*, 2015.

[14] E. P. Tighe and C. K. Cheng, Modeling personality traits of Filipino Twitter users, *Proc. of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp.112-122, 2018.

[15] G. Y. N. N. Adi, M. H. Tandio, D. Suhartono and V. Ong, Optimization for automatic personality recognition on Twitter in Bahasa Indonesia, *Procedia Computer Science*, vol.135, pp.473-480, 2018.

[16] V. Ong, A. D. S. Rahmanto, Williem and D. Suhartono, Exploring personality prediction from text on social media: A literature review, *Internetworking Indonesia Journal*, vol.9, no.1, pp.65-70, 2017.

[17] B. Y. Pratama and R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM, *2015 International Conference on Data and Software Engineering (ICoDSE)*, pp.170-174, 2015.

[18] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari and M. N. Suprayogi, Personality prediction based on Twitter information in Bahasa Indonesia, *Proc. of the 2017 Federated Conference on Computer Science and Information Systems*, pp.367-372, 2017.

[19] T. Ryan and S. Xenos, Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage, *Computers in Human Behavior*, vol.27, no.5, pp.1658-1664, 2011.

[20] C. J. Soto and J. J. Jackson, Five-factor model of personality, in *Oxford Bibliographies in Psychology*, D. S. Dunn (ed.), Oxford University Press, NY, 2013.

[21] R. R. McCrae and P. T. Jr. Costa, The five-factor theory of personality, in *Handbook of Personality: Theory and Research*, O. P. John, R. W. Robins and L. A. Pervin (eds.), Guilford Press, NY, 2008.

[22] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*, 1992.

[23] O. P. John and S. Srivastava, The Big Five trait taxonomy: History, measurement, and theoretical perspectives, in *Handbook of Personality: Theory and Research*, Guilford Press, NY, 1999.

[24] S. D. Gosling, P. J. Rentfrow and W. B. Swann, A very brief measure of the Big-Five personality domains, *Journal of Research in Personality*, vol.37, no.6, pp.504-528, 2003.

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

[26] J. W. Pennebaker, R. L. Boyd, K. Jordan and K. Blackburn, *The Development and Psychometric Properties of LIWC2015*, University of Texas at Austin, 2015.

[27] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock and S. Davalos, A multivariate regression approach to personality impression recognition of vloggers, *Proc. of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pp.1-6, 2014.

[28] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, Our Twitter profiles, our selves: Predicting personality with Twitter, *Proceedings – 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing (PASSAT/SocialCom 2011)*, 2011.

[29] A. C. E. S. Lima and L. N. de Castro, A multi-label, semi-supervised classification approach applied to personality prediction in social media, *Neural Networks*, 2014.

[30] D. Nie, Z. Guan, B. Hao, S. Bai and T. Zhu, Predicting personality on social media with semi-supervised learning, *Proceedings – 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology C Workshops (WI-IAT 2014)*, 2014.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikitlearn: Machine learning in python, *Journal of Machine Learning Research*, vol.12, pp.2825-2830, 2011.

[32] J. Golbeck, C. Robles, M. Edmondson and K. Turner, Predicting personality from Twitter, *Proceedings – 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing (PASSAT/SocialCom 2011)*, 2011.

[33] F. Z. Tala, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, Master Thesis, Universiteit van Amsterdam, 2003.

[34] Y. J. Zhu and Z. B. Ghahramani, *Learning from Labeled and Unlabeled Data with Label Propagation*, Technical Report CMU-CALD-02-107, Carnegie Mellon University, Pittsburgh, 2002.

[35] E. C. Polley and M. J. van der Laan, Super learner in prediction, *U.C. Berkeley Div. Biostat. Working Paper*, 2010.

[36] X. Zhu and A. B. Goldberg, Semi-supervised learning tutorial, *Synth. Lect. Artif. Intell. Mach. Learn.*, 2009.

[37] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letter*, vol.27, no.8, pp.861-874, 2006.

[38] T. Tape, Interpretation of diagnostic tests, *Ann. Intern. Med.*, 2001.