Proceedings of the Federated Conference on Computer Science and Information Systems pp. 367–372
DOI: 10.15439/2017F359
ISSN 2300-5963 ACSIS, Vol. 11

Personality Prediction Based on Twitter Information in Bahasa Indonesia

Veronica Ong[1], Anneke D. S. Rahmanto[1], Williem[1], Derwin Suhartono[1], Aryo E. Nugroho[2], Esther W. Andangsari[2], Muhamad N. Suprayogi[2]
[1]School of Computer Science, Computer Science Department, Bina Nusantara University, Jakarta, Indonesia
[2]Faculty of Humanities, Psychology Department, Bina Nusantara University, Jakarta, Indonesia
Email: {veronica.ong, anneke.rahmanto, williem002}@binus.ac.id, dsuhartono@binus.edu, aryonugroho@binus.ac.id, {esther, msuprayogi}@binus.edu

***Abstract*—The sheer usage of social media presents an opportunity for an automated analysis of a social media user based on his/her information, activities, or status updates. This opportunity is due to the abundant amount of information shared by the user. This fact is especially true for countries with high number of active social media users such as Indonesia. Extraction of information from social media can yield insightful results if done correctly. Recent studies have managed to leverage associations between language and personality and build a personality prediction system based on those associations. The current study attempts to build a personality prediction system based on a Twitter user's information for Bahasa Indonesia, the native language of Indonesia. The personality prediction system is built on Support Vector Machine and XGBoost trained with 329 instances (users). Evaluation results using 10-fold cross validation shows that the system managed to reach highest average accuracy of 76.2310% with Support Vector Machine and 97.9962% with XGBoost.**

I. Introduction

STATISTICS show that 1 in every 3 minutes of Internet usage is spent on social media [1]. The sheer usage of social media means that a lot of information are shared by users during their social media usage. Information can be shared explicitly or implicitly. One of the information that can be analyzed from social media usage is user's personality.

Recent studies on automated personality assessment (hereinafter personality prediction) have been conducted in the past on several social medias. The current study focuses on Twitter. Twitter has gained high popularity over the years. Statistics show that the number of active users on Twitter are constantly rising each quarter and reaching up to 313 million active Twitter users as of June 2016 [2], [3].

Unlike other studies which focuses on English as the prediction system's main language, this study focuses on Bahasa Indonesia, the mother tongue of Indonesia. There are several statistics which show that Indonesia has high Twitter usage. The first among them is a study by [4], in which they mentioned that 2.4% of worldwide tweets are posted by users of Jakarta, the capital city of Indonesia. Mr. Roy Simangunsong, Indonesia's Twitter Country Head, reported that 77% of Indonesians are active on Twitter every day [5]. An observation by eMarketer on November 2015 also shows the rise of Twitter users in Indonesia from year 2014-2015 and is predicted to keep rising until 2019 [6].

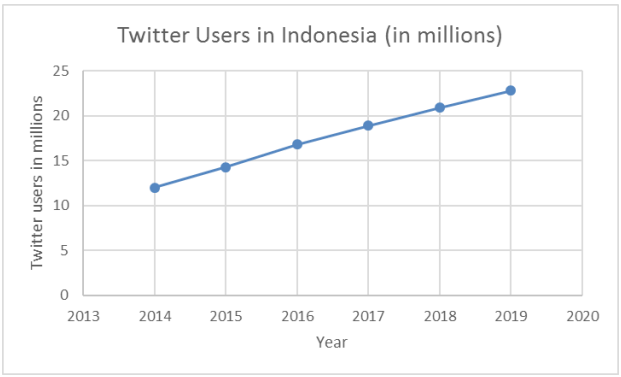


Fig. 1 Number of Twitter users in Indonesia per year 2014 – 2019 in millions

The personality prediction system for this study is built to classify a user's personality based on The Five Factor Model, a personality model by McCrae and Costa, which divides an individual's personality into 5 traits, namely Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness. The contributions of this paper are the personality prediction system built for the Bahasa Indonesia language, set of scenarios which contribute to the system's accuracy, and the comparison of 2 machine learning algorithms implemented into the prediction model.

II. Related Works

Previous studies have attempted to implement personality prediction on Twitter. [7] and [8] built a personality prediction system for The Five Factor Model. A personality prediction system was also built for the Dark Triad personality model in [9]. [7], [8], and [9] built the personality prediction system for English using tools such as LIWC (Linguistic Inquiry and Word Count) and MRC Psycholinguistic Database. Another study by [10] also used LIWC to build a personality prediction system on Facebook. These tools are predefined categories of words which can be

This work was supported by grant from Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

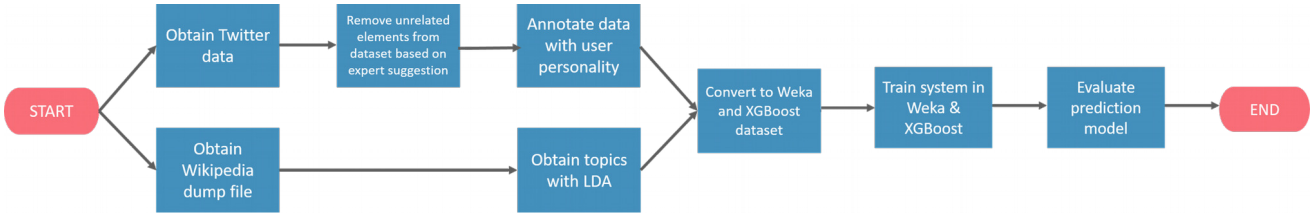


Fig. 2 Overview of methodology

used to assess the tendency of a user to talk about a certain category. Such tools have also been utilized to create a prediction system in non-English languages such as Spanish, Dutch, Italian [11], and Chinese [12].

A literature review on personality prediction by [13] states that among the literatures that they examined, more than half utilized such tools to build their personality prediction system. Despite its usefulness, LIWC and MRC have language limitations—it doesn't support all languages. The tools are not supported in Bahasa Indonesia, so another approach must be applied to this study.

Other recent studies have come up with another approach by assessing the tendency of a user's choice of words. This is done by counting the usage frequency of a certain n-gram by a user. This method has been implemented in the past with data from various social platforms such as blogs [14] [15] and Facebook [16][17]. Said method has also been applied for non-English languages such as Chinese [18] and Bahasa Indonesia [19].

In [19], they managed to build a personality prediction system for The Five Factor Model using myPersonality, which is a corpus consisting of status updates from users which have been labelled with The Five Factor Model personality traits. The dataset is translated into Bahasa Indonesia to build a prediction system in said language. Therefore, this study attempts to apply the personality prediction task on an original, non-translated Bahasa Indonesia corpus.

III. The Five Factor Model

Personality is regarded as the main factor of what causes an individual to act a certain way in online interactions [20]. The Five Factor Model is one of the most widely used concepts in studies observing the association between personality and social media use [21][22][20][23][24]. Results from these studies show that the Five Factor Model can indeed act as a predictor in social media use.

The Five Factor Model is a hierarchical structure of personality traits which consist of 5 main dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience [25]. In fact, the Five Factor Model is commonly used among psychologists to comprehensively describe personality. The naming of personality traits is done through a series of literature reviews and studies. One of the examples is Neuroticism, which corresponds to low scores of Emotional Stability [25]. The neuroticism trait cannot be viewed as someone with psychopathological characteristics, but someone who is unsatisfied with his/her life [26] or an individual who tends to experience psychological distress [20]. Individuals with low Extraversion (Introversion) are viewed as reserved, not unfriendly, and independent individuals. They also prefer to be alone without having social anxiety [26]. Extraverted people are interpreted as individuals who tend to be sociable and experience positive emotions [20]. Individuals with high Openness scores are individuals who are open to new ideas while cautiously implementing them. On the other hand, individuals with low Openness scores have smaller scope of interest [26]. People with Agreeableness trait are trusting of others, sympathetic, and cooperative [20]. Individuals with high scores on the Conscientiousness trait are active in planning and organizing their activities, while an individual with a low Conscientiousness score is usually more laid-back in their work [26].

The previous study between Facebook and the Five Factor Model shows that individuals with Extraversion trait have higher number of friends [20][24]. Introverted individuals tend to present more personal information on their social media [20]. Individuals with high Neuroticism show the tendency to post photos of themselves more compared to those with low scores [20]. This study however contradicts results from a previous study by [24]. Individuals with high Openness score are known to be more expressive on their Facebook profiles [20]. People with high Conscientiousness trait have more friends and have tendency to post pictures compared to individuals with low Conscientiousness [20]. This result too, contradicts the results from [24]'s study. Finally, more observation is required regarding the correlation between Agreeableness trait individuals towards their social media usage [20].

IV. Methodology

This study consists of 3 main tasks: data collection, preprocessing, and building the prediction model. Figure 2 shows an overview of the method applied in this study.

A. Data Collection

Preparation of dataset is done to obtain the training dataset and testing dataset. The dataset acquired contains Twitter user information and a maximum of 100 of the user's latest tweets. Users are chosen based on the following criteria:

TABLE I.
TRAINING DATASET DISTRIBUTION

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
High	134	92	150	202	163
Low	195	237	179	127	166

TABLE II.
TESTING DATASET DISTRIBUTION

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
High	19	16	21	24	16
Low	11	14	9	6	14

1. User posts on Twitter at least once a month.
2. User uses Bahasa Indonesia as their main language.

The user information extracted covers 12 features:

1. Number of tweets
2. Number of followers
3. Number of following
4. Number of favorites
5. Number of retweets from extracted tweets
6. Number of retweeted tweets from extracted tweets
7. Number of quote tweets from extracted tweets
8. Number of mentions from extracted tweets
9. Number of replies from extracted tweets
10. Number of hashtags from extracted tweets
11. Number of URLs from extracted tweets
12. Average time difference between each tweet

A total of 359 data were collected, where 1 data represents Twitter data from 1 user. 329 data were utilized as training data, and the remaining 30 data as testing data. The dataset was then annotated with "high" or "low" label for each personality trait by 3 psychology experts. A "high" label indicates that the user has a high level of a certain personality trait, while "low" label represents that the user has a low level of a certain personality trait. Thus, each user consists of 5 labels, each label representing the level (high or low) of each personality.

Table 1 shows the distribution of the training dataset, while table 2 shows the distribution of the testing dataset.

B. Preprocessing

To preprocess the extracted information from Twitter, a series of automated and manual removal of elements were applied. Automatic element removal involves omitting retweets, replacing mentions with "[UNAME]" token, replacing hashtag with "[HASHTAG]" token, removing hyperlinks/URLs, and removing emoji. After applying automatic element removal, several manual element removals were applied to reduce the noise in the training data (e.g. non-Bahasa Indonesia content, non-Twitter content).

Next, tokenization is applied to the resulting dataset from the previous step, which produces a series of unigram and bigram. The occurrence of each unigram and bigram is counted. Each n-gram goes through a series of n-gram normalization functions to reduce the occurrence of unrecognized words (e.g. misspelled or slang words). The n-gram normalization functions applied were adapted from [27] and [28].

Omission of stop words was also applied in scenarios which require said action. The scenarios are further explained in section 4.3. The list of stop words was adapted from [29].

Finally, the system also utilized LDA (Latent Dirichlet Allocation) generated topics. Topics were generated using a Bahasa Indonesia Wikipedia dump file. The dump file contains the content of every article available on the Bahasa Indonesia version of Wikipedia. This file is loaded into the LDA algorithm with Gensim [30] to produce 100 topics, where each topic consists of 20 words.

The final output of the dataset consists of the 13 features presented in section 4.1 representing the user information, and the frequency of each n-gram.

C. Build Prediction Model

The personality prediction system consists of 5 classifiers. Each classifier is tasked with the prediction of 1 personality trait. The system is trained with 329 instances of the output from the preprocessing step. Classifiers built on the Support Vector Machine and XGBoost are trained with the same dataset. The Support Vector Machine classifier was run on Weka, while XGBoost was run on R.

After the training process, the system is evaluated using 10-fold cross validation and loading the 30-instance testing dataset into the system. The evaluation measure used for evaluation is accuracy.

The personality prediction system is tested on different scenarios with the following actions:

1. Minimum occurrence of n-gram (minimum occurrence=1 or minimum occurrence=2)

TABLE II.
SCENARIOS FOR EVALUATION

Scenario	Minimum occurrence of n-gram		n-gram weighting scheme		LDA topic features		Stop words omission	
	1	2	Boolean	TF	Use LDA	Don't use LDA	Omit	Don't omit
1	✓		✓		✓		✓	
2	✓		✓		✓			✓
3	✓		✓			✓	✓	
4	✓		✓			✓		✓
5	✓			✓	✓		✓	
6	✓			✓	✓			✓
7	✓			✓		✓	✓	
8	✓			✓		✓		✓
9		✓	✓		✓		✓	
10		✓	✓		✓			✓
11		✓	✓			✓	✓	
12		✓	✓			✓		✓
13		✓		✓	✓		✓	
14		✓		✓	✓			✓
15		✓		✓		✓	✓	
16		✓		✓		✓		✓

Refers to the number of times an n-gram appears in the list of extracted tweets.
If minimum occurrence is set to 1 for a scenario, then the system will take all the user's existing n-grams into consideration for the prediction.
If the scenario's minimum occurrence is set to 2, then the system will only take n-grams that appear at least twice into consideration for the prediction.

2. n-gram weighting scheme (Boolean or TF weighting)
   Refers to how an n-gram's weight is calculated.
   If the weighting scheme for a scenario is Boolean, then the n-gram's weight is set to 1 if it appears in the list of tweets, and 0 if otherwise.
   However, if the weighting scheme for a scenario is TF, then the n-gram's weight is set to the number of times it appears in the list of tweets.
3. LDA topic features (use LDA topic features or don't use LDA topic features)
   Refers to whether LDA-generated topic features are used in a scenario.
4. Stop words omission (omit stop words or don't omit stop words).
   Refers to whether stop words are omitted from the list of n-grams in a scenario.

Combining these actions results in a total of 16 scenarios, which are shown in table 3. Each row represents a single scenario. The checked cells on said table are the actions used in the row of the corresponding scenario.

V. RESULT AND DISCUSSION

The system is evaluated with a held-out test set of 30 data and 10-fold cross validation. A test set evaluation is included to make sure the system still performs the same way as when evaluated with 10-fold cross validation.

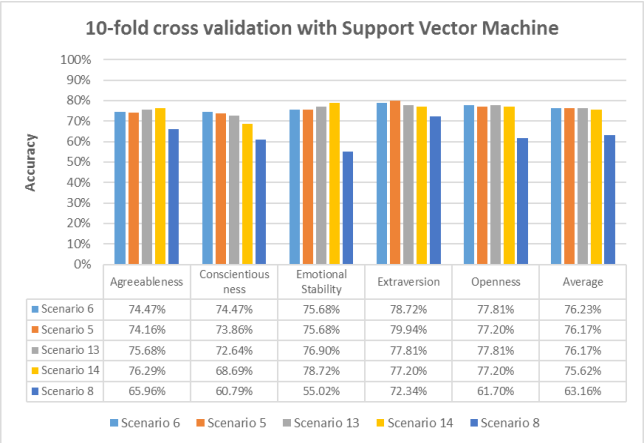
The evaluation results are shown on Figures 3 to 6. Due to the large number of scenarios tested, only the top 5 average accuracies for each evaluation are presented in the figures below.

The results from Figure 3 show that the highest average accuracies are dominated by scenarios 6, 5, 13, and 14. The 4 scenarios have 2 things in common: the usage of TF weighting scheme and LDA topic features. The highest average accuracy is 79.9392%, which is achieved on the Extraversion personality trait with scenario 5.

The results from Figure 4 are dominated by scenarios 14 and 13. The common feature shared by both scenarios are that they utilize TF weighting scheme and LDA topic features. Evaluation on test set with Support Vector Machine managed to achieve 90%, the highest accuracy for the Agreeableness trait with scenario 6, and Extraversion trait with scenarios 13 and 14.

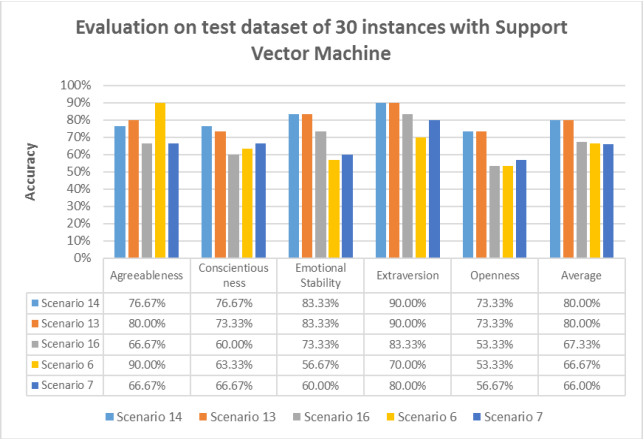
Figure 5 presents the results from 10-fold cross validation with XGBoost, which are dominated by scenarios 6, 5, 14, and 13. The mentioned scenarios also have the same thing in common as the previous evaluations: usage of TF weighting scheme and LDA topic features. In this evaluation, Emotional Stability with scenarios 5 and 6 managed to achieve highest accuracy, which is 98.7900%.

Figure 6 presents the accuracy of the XGBoost classifier when evaluated with a 30-instance test dataset. The evaluation results are dominated by scenarios 13 and 14, where both scenarios utilize TF weighting scheme and LDA topic features. 100% accuracy is achieved on the Emotional Stability and Extraversion personality trait with scenario 13, and on Openness with scenario 14.



	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness	Average
Scenario 6	74.47%	74.47%	75.68%	78.72%	77.81%	76.23%
Scenario 5	74.16%	73.86%	75.68%	79.94%	77.20%	76.17%
Scenario 13	75.68%	72.64%	76.90%	77.81%	77.81%	76.17%
Scenario 14	76.29%	68.69%	78.72%	77.20%	77.20%	75.62%
Scenario 8	65.96%	60.79%	55.02%	72.34%	61.70%	63.16%

Fig. 3 Accuracy of Support Vector Machine using 10-fold cross validation



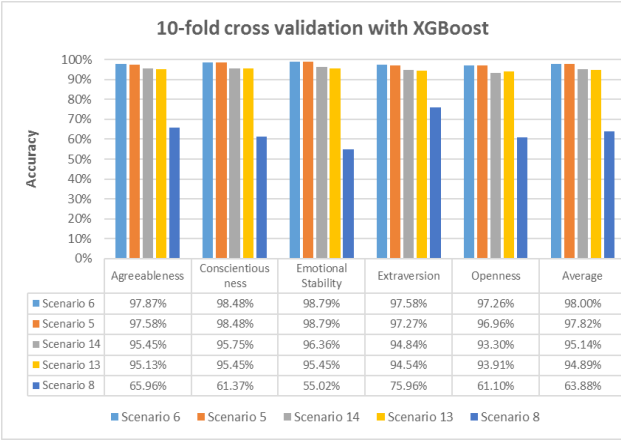
	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness	Average
Scenario 14	76.67%	76.67%	83.33%	90.00%	73.33%	80.00%
Scenario 13	80.00%	73.33%	83.33%	90.00%	73.33%	80.00%
Scenario 16	66.67%	60.00%	73.33%	83.33%	53.33%	67.33%
Scenario 6	90.00%	63.33%	56.67%	70.00%	53.33%	66.67%
Scenario 7	66.67%	66.67%	60.00%	80.00%	56.67%	66.00%

Fig. 4 Accuracy of Support Vector Machine using test dataset

The TF weighting scheme managed to achieve higher accuracy as it provides the system with information of how many times the word occurs from a user with a particular type of personality. The Boolean weighting scheme doesn't contain this information since the values only show whether a particular word is used by the user.

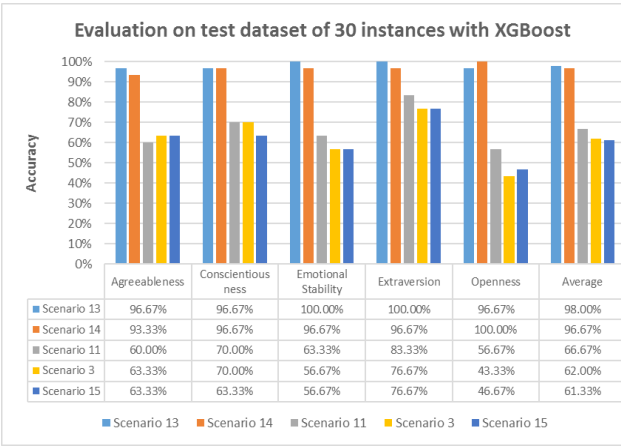
The LDA topic features also contributed to the system's accuracy because it does not restrict the system to assess by the user's choice of words, but also by the user's choice of topics.

Results from XGBoost show a significant increase in accuracy compared to Support Vector Machine, even when evaluated on 10-fold cross validation or a prepared test set. This is also consistent with other literatures which claim that XGBoost managed to achieve the best prediction when compared to other algorithms [31][32][33]. The creator of XGBoost also reports that XGBoost was used by the top 10 winning teams in KDD Cup 2015 [34].



	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness	Average
Scenario 6	97.87%	98.48%	98.79%	97.58%	97.26%	98.00%
Scenario 5	97.58%	98.48%	98.79%	97.27%	96.96%	97.82%
Scenario 14	95.45%	95.75%	96.36%	94.84%	93.30%	95.14%
Scenario 13	95.13%	95.45%	95.45%	94.54%	93.91%	94.89%
Scenario 8	65.96%	61.37%	55.02%	75.96%	61.10%	63.88%

Fig. 5 Accuracy of XGBoost classifier using 10-fold cross validation



	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness	Average
Scenario 13	96.67%	96.67%	100.00%	100.00%	96.67%	98.00%
Scenario 14	93.33%	96.67%	96.67%	96.67%	100.00%	96.67%
Scenario 11	60.00%	70.00%	63.33%	83.33%	56.67%	66.67%
Scenario 3	63.33%	70.00%	56.67%	76.67%	43.33%	62.00%
Scenario 15	63.33%	63.33%	56.67%	76.67%	46.67%	61.33%

Fig. 6 Accuracy of XGBoost classifier using test dataset

VI. Conclusions and Improvements

In this study, we have presented a personality prediction system for Bahasa Indonesia based on a Twitter user's information. Results of this study show that personality prediction in Bahasa Indonesia is indeed possible without using a tool with predefined words (LIWC, MRC), but by assessing a user's choice of words. The current study compares 2 different classifiers: Support Vector Machine and XGBoost. Both classifiers are tested under different scenarios which involve minimum occurrence of n-gram, n-gram weighting scheme, usage of LDA topic features, and omission of stop words. Evaluation using 10-fold cross validation showed that the personality prediction system built on Support Vector Machine managed to achieve a highest average accuracy of 76.2310%, while XGBoost achieved 97.9962%.

Evaluation results using 10-fold cross validation and 30-instance test dataset also showed that usage of LDA topic features and TF frequency weighting scheme contributed greatly to the personality prediction system's accuracy.

The results also showed that even when tested under the same scenario and same dataset, the personality prediction system built on XGBoost managed to perform significantly better than on Support Vector Machine.

Future developments of this study may utilize a larger training and testing dataset, which will allow the system to immerse itself in a wider variety of tweets. Improving n-gram normalization functions may also increase the system's accuracy since it allows the system to recognize and assess more words.

ACKNOWLEDGMENT

The authors would like to thank Mr. Tri Swasono Hadi for his participation in labelling the personality traits for each data. Mr. Tri is a practitioner in clinical psychology.

This research and publication is fully supported by grant named "Penelitian Produk Terapan" from Ministry of Research, Technology and Higher Education of the Republic of Indonesia with contract number 039A/VR.RTT/VI/2017

REFERENCES

[1] GlobalWebIndex, "GlobalWebIndex Social Report Q4/2016," 2016.

[2] Twitter Investor Relations, "Q414 Selected Company Metrics and Financials," 2014. .

[3] Twitter Investor Relations, "Q216 Selected Company Metrics and Financials," 2016. .

[4] K. M. Carley, M. M. Malik, M. Kowalchuck, J. Pfeffer, and P. Landwehr, "Twitter usage in Indonesia," 2015.

[5] CNN Indonesia, "Twitter Rahasiakan Jumlah Pengguna di Indonesia," 2016. .

[6] eMarketer, "Southeast Asia Has Among the Highest Social Network Usage in the World," 2015. .

[7] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011, pp. 149–156.

[8] A. Wijaya, I. Prasetia, N. Febrianto, and D. Suhartono, "Sistem Prediksi Kepribadian 'The Big Five Traits' Dari Data Twitter," Bina Nusantara University, 2016.

[9] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012, vol. 2, pp. 386–393.

[10] G. Farnadi, S. Zoghbi, M. Moens, and M. De Cock, "Recognising Personality Traits Using Facebook Status Updates," *Work. Comput. Personal. Recognit. Int. AAAI Conf. weblogs Soc. media*, pp. 14–18, 2013.

[11] M. Arroju, A. Hassan, and G. Farnadi, "Age, Gender and Personality Recognition using Tweets in a Multilingual Setting," in *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*, 2015.

[12] D. Wan, C. Zhang, M. Wu, and Z. An, "Personality Prediction Based on All Characters of User Social Media Information," pp. 220–230, 2014.

[13] V. Ong, A. D. S. Rahmanto, Williem, and D. Suhartono, "Exploring Personality Prediction from Text on Social Media: A Literature Review," *Internetworking Indones. J.*, vol. 9, no. 1, pp. 65–70, 2017.

[14] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large Scale Personality Classification of Bloggers," in *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9--12, 2011, Proceedings, Part II*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 568–577.

[15] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *J. Res. Pers.*, vol. 44, no. 3, pp. 363–373, 2010.

[16] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality , Gender , and Age in the Language of Social Media : The Open-Vocabulary Approach," vol. 8, no. 9, 2013.

[17] Y. Liu, J. Wang, and Y. Jiang, "PT-LDA: A Latent Variable Model to Predict Personality Traits of Social Network Users," *Neurocomputing*, 2015.

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

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

[20] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Comput. Human Behav.*, vol. 26, no. 6, pp. 1289–1295, 2010.

[21] J. L. Skues, B. Williams, and L. Wise, "The effects of personality traits, self-esteem, loneliness, and narcissism on Facebook use among university students," *Comput. Human Behav.*, vol. 28, no. 6, pp. 2414–2419, 2012.

[22] T. Ryan and S. Xenos, "Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage," *Comput. Human Behav.*, vol. 27, no. 5, pp. 1658–1664, 2011.

[23] T. Correa, A. W. Hinsley, and H. G. De Zuniga, "Who interacts on the Web?: The intersection of users' personality and social media use," *Comput. Human Behav.*, vol. 26, no. 2, pp. 247–253, 2010.

[24] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with Facebook use," *Comput. Human Behav.*, vol. 25, no. 2, pp. 578–586, 2009.

[25] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.

[26] I. B. Weiner and R. L. Greene, "Revised NEO Personality Inventory," *Handb. Personal. Assess.*, pp. 315–342, 2008.

[27] A. R. Naradhipa and A. Purwarianti, "Sentiment classification for Indonesian message in social media," in *Cloud Computing and Social Networking (ICCCSN), 2012 International Conference on*, 2012, pp. 1–5.

[28] G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation," pp. 7–8, 2014.

[29] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," *Inst. Logic, Lang. Comput. Univ. van Amsterdam, Netherlands*, 2003.

[30] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," *Proc. Lr. 2010 Work. New Challenges NLP Fram.*, pp. 45–50, 2010.

[31] V. Ayumi, "Pose-based human action recognition with Extreme Gradient Boosting," in *Research and Development (SCOReD), 2016 IEEE Student Conference on*, 2016, pp. 1–5.

[32] I. Babajide Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, vol. 21, no. 8, p. 983, 2016.

[33] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting."

[34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.