Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 19, No. 4, December 2019, pp. 283-289
http://doi.org/10.5391/IJFIS.2019.19.4.283

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Identifying Personality Traits for Indonesian User from Twitter Dataset

Nicholaus Hendrik Jeremy, Cristian Prasetyo, and Derwin Suhartono
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

Abstract

Social media allows the user to convey their actual self and share their life experiences through numerous ways. This behavior in turn reflects the user's personality. In this paper, we experiment to automatically predict user's personality based on Big Five Personality Trait on Twitter. Our focus is towards Indonesian user. Not only word n-gram, Twitter metadata is also used in a certain combination to determine the feature that will be used to predict the personality. Our research also attempts to find optimum setting based on the number of n-gram, classifier, and twitter metadata. Our experiment yields 0.7482 at most on F-Measure. We conclude that among all scenario, twitter metadata is the least impactful feature, while word n-gram impacts the most.

Keywords: Social media, Personality prediction, Twitter, Big five

1. Introduction

Social media has become an integral part in our life. Its usage emerges from several aspects, which are the needs of doing social interaction and exchanging information. The way we represent ourselves in social media through posts indirectly expresses our personality as a person. Research shows that social media user tends to express their actual personality rather their fabricated or false-image self [1].

Indonesia is a heavily populated country and the citizen's consumption of internet increased every year. A survey done in March 2019 by Polling Indonesia and Indonesian Internet Service Provider Association (originally *Asosiasi Penyelenggara Jasa Internet Indonesia*) shows that there is an increase of internet user in Indonesia by 10.12%, reaching 171.7 million user.

Social media user contributes the most from sharing information in either text, image, audio, or video. The abundant amount of user and activities per day provides researcher a large data to be tested. Doing manual prediction could be laborious due to the reason stated. Provided the technology available today, automatic prediction utilizing computer is possible. In fact, over time computer able to outperform human in personality prediction [2].

Large amount of base internet user causes different researchers to tackle on different problem.One of the most often problem occurred when tackling this topic is the lack of dataset. Research done by [3] shows that this problem is possible to be solved by automatically retrieving data from Twitter that is related with Myers-Briggs Type Indicator (MBTI). From personal observation, MBTI is more widely known than Big Five, making automatic data retrieval easier in larger amount.

Received: Aug. 13, 2019
Revised : Dec. 18, 2019
Accepted: Dec. 21, 2019

Correspondence to: Derwin Suhartono
(dsuhartono@binus.edu)

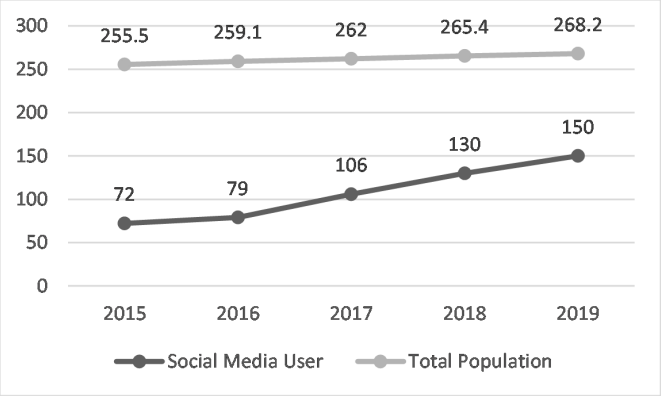


Figure 1. Number of social media user compared to total population in Indonesia in million according to Hootsuite and We Are Social [24–28]. All data is retrieved per January of each year.

Another problem tackled is language usage. Research done by [4] attempted to compare Big Five to MBTI in two different language scenario: mixed and English-only. It is found that not only that MBTI performs better than Big Five, attempt on mixed language performs better as well, although the difference is barely noticeable because the amount of English dataset is very dominant. It is noted, however, that feature selection may affect the result.

We attempt on creating an automatic personality prediction based on social media content and activities. To specify the problem, we scope it to exclusively Indonesia citizen. We start by briefly explaining the personality model and its relation with social media. After that, we propose our data preparation and learning methods.

2. Literature Review

2.1 Personality Model

The personality model used is Big Five, which consist of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) [5, 6].

- Openness (OPN) relates to how acceptance they are towards unusual behavior or unique experiences. People with high score of openness are creative, abstract, and imaginative.
- Conscientiousness (CON) relates to self-control against impulsive actions. People with high score of conscientiousness are discipline, cautious, and dutiful.
- Extraversion (EXT) relates to someone's approach to social world. People with high score of extraversion are communicative, easy to approach, and assertive.
- Agreeableness (AGR) relates to someone's faith to others. People with high score of agreeableness are trustful, honest, and modest.
- Neuroticism (NEU) relates to someone's behavior when facing negative experience. People with high score of neuroticism are emotionally unstable and vulnerable.

Table 1. Personalities in big five and its description according to their score

Personality	High	Low
OPN	Adventurous, abstract	Prefer regularity, conventional
CON	Disciplined, reliable, strict	Disorganized, impulsive, laid back
EXT	Friendly, joyous	Solitude, independent
AGR	Cooperative, honest	Sceptical, suspicious
NEU	Self-conscious, prone to negativity	Contained, calm

2.2 Personality in Social Media

There are some characteristics for each trait that can be observed from user's activity. Longer period of browsing through social media and user with high neuroticism is found to be associated with higher risk of social isolation [7]. The period of time used, however, is unlikely related to high score of neuroticism but extraversion [8], although it has been found otherwise as well [9]. Under the same research, it is found that people with high score of conscientiousness are less likely to be active in social media. This holds true in our dataset.

2.3 Personality Prediction

Preceding researches have attempted on predicting personality on various conditions, such as using multiple social media platform [10], using different personality model [11], or using key feature other than linguistic cues [12]. Most of research that uses linguistic cues as key feature uses Linguistic Inquiry and Word Count (LIWC) [13], a dictionary that allows word counting based on their groups. However, not all language are supported by LIWC, with Indonesia being one of them. Other similar dictionary, MRC Psycholinguistic Database [14], used

in [10, 15], is also not available in Indonesia. In this paper we try to replicate exactly every feature that preceding papers used. Not using LIWC nor MRC might causes significant difference of evaluation score result compared to the result from the actual research, knowing that such linguistic cues are major features.

Under the same institution there has been conducted a research [29] with similar topic but utilizes boosting classifier. The research shows that using boosting algorithm, specifically XGBoost, may improve the overall performance by significant amount. Further research under the same team unpublished manuscript (V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model,") focuses on using XGBoost as classifier and AUROC (Area Under the ROC Curve) as performance metric. The research achieves higher perfomance when predicting agreeableness and openness than the rest of the personality traits, although it is possible that their dataset imbalance affects the result.

3. Dataset

The dataset used was taken from an unpublished manuscript (V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model,"). It consists of 508 twitter data from indonesian users and 5 classes represent big five personality traits, which for each class consists of two labels: HIGH and LOW. Three psychology experts labeled the dataset using voting system - for each trait of one user the label that gets most vote becomes the label. Each user has processed tweets and Twitter metadata as their feature. The bolded value represents the baseline of the dataset.

3.1 Preprocessing Tweets

Tweets are first preprocessed using method proposed by [16]. The result is then used to create word n-gram. Word n-gram is a sequential order of n words from a sentence. In this paper, we use bigram and trigram. The result is sorted in descending manner, as appearance that is more frequent means that the n-gram represents the class better. The amount of both unique bigram and trigram is 428,115 with the most amount of occurrence of 8,699.

Table 2. Class distribution

	High	Low
AGR	278 (**54.7%**)	230 (45.3%)
CON	131 (25.8%)	376 (**74.2%**)
EXT	363 (**71.5%**)	145 (28.5%)
NEU	221 (43.5%)	287 (**56.5%**)
OPN	272 (**53.5%**)	236 (45.5%)

3.2 Listing the Metadata

Twitter metadata is obtained through its API. Other than tweet content, for this paper, we use the amount of follower, following, favorite, tweet, hashtag, retweet, retweeted, mentions, mentioned, and links. Some element is then calculated with

Table 3. Twitter metadata used compared to related works

Our metadata	**List of metadata from** [4]	**List of metadata from** [3]	**List of metadata from** [17]
Amount of follower	Followers tweets ratio	Amount of tweets	Amount of followers
Amount of following	Favorite tweets to tweets ratio	Amount of followers	Amount of following
Amount of tweets	Hashtag to words ratio	Total of tweets and rewteets	Amount of mentions[b]
Amount of favorites	Retweets to retweeted ratio	Amount of favorites	Amount of replies[b]
Amount of retweets	Listed count[a]	User's gender[a]	Amount of hashtags[b]
Amount of retweeted	Link color[a]	Listed count[a]	Amount of urls[b]
Amount of mention	Text color[a]	-	Average word per tweet
Amount of quote	Border color[a]	-	Density of social network
Amount of replies	Background color[a]	-	-
Amount of hashtag	Default profile picture[a]	-	-

[a] Color hex code, listed count, profile picture, and user gender is not stored in the dataset. It is possible that the user has changed any of it or they have their account suspended. Revising the account risks not only more resource, but also requires revising the manual personality labelling done by the expert, as their personality may have changed [18, 19].

[b] The metadata uses both sum and average per tweet [17].

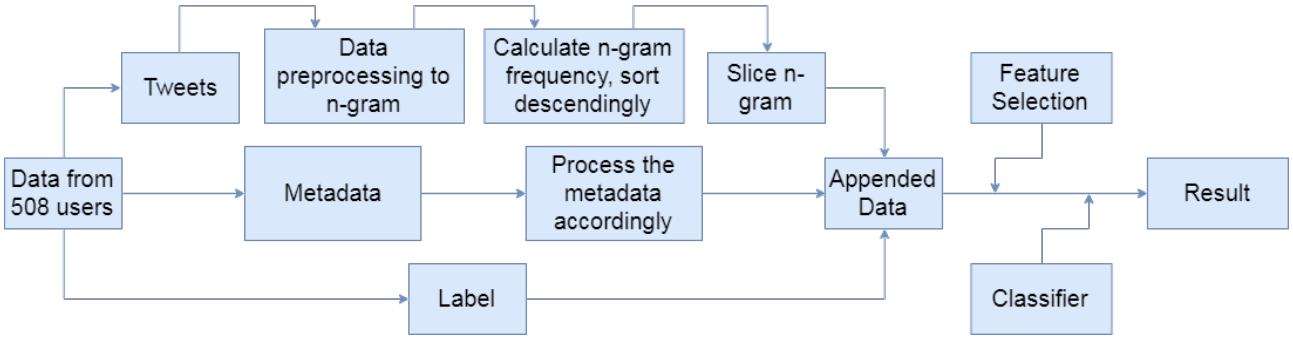


Figure 2. Flow of the experiment.

other element accordingly. In this paper, we compare three combination of metadata from different papers.

4. Methodology

We train our dataset in three different cases. Before training, we use correlation-based feature subset selection [20] to minimize the dimension. The classifier we use will be stated on each case. To separate the training data from test data we use 10-fold cross validation. The evaluation metrics used are precision, recall, and f-measure. We do not use accuracy to prevent accuracy paradox [21, 22] and therefore not using baseline value as it displays accuracy.

4.1 Word n-gram

We use partial amount instead of the whole n-gram, considering that we have sorted the frequency by descending, thus making latter feature sparser than and not as relevant as the feature with higher frequency. We conduct experiment with and without metadata appended. All training is done with Naïve Bayes.

4.2 Twitter Metadata

We compare the result between our selected metadata feature and other works' which can be seen in table 3. The goal is to see if we can make a list of metadata that able to outperform other works specifically for our dataset. The amount of word n-gram used is 1000. All training is done with Naïve Bayes.

4.3 Classifier

We experiment on various classifier to see which one works best for this specific dataset. None of the classifier is tuned, meaning that we use their default value for each hyperparameter. The classifiers used consist of k-Nearest Neighbors (k-NN), tree-based (J48 and Random Forest), Support Vector Machine (SVM), and Naïve Bayes. The dataset used consists of 1000 word n-gram appended with our proposed Twitter metadata.

5. Results

We show the average of each metric: precision (P-AVG), recall (R-AVG), and f-measure (F-AVG). We divide the result based on the experiment case. The bold value represents the highest value in a column.

Table 4. Result on how different word n-gram amount and appended metadata affects the result

Dataset	P-AVG	R-AVG	F-AVG
1000	0.7684	0.725	0.718
1500	0.7728	0.736	0.7306
2000	0.7808	0.7424	0.7372
2500	0.7874	0.749	0.7436
3000	0.7876	0.749	0.7428
3500	**0.7932**	**0.7542**	0.7468
4000	**0.7932**	0.7526	0.7452
4500	0.7926	0.7506	0.7426
5000	0.7912	0.7506	0.7426
1000 + metadata	0.745	0.7164	0.7152
1500 + metadata	0.7458	0.7216	0.7192
2000 + metadata	0.7536	0.729	0.727
2500 + metadata	0.7738	0.7454	0.7414
3000 + metadata	0.7712	0.7446	0.7408
3500 + metadata	0.782	0.7524	**0.7482**
4000 + metadata	0.7814	0.7478	0.7422
4500 + metadata	0.783	0.7466	0.7399
5000 + metadata	0.778	0.7404	0.7332

5.1 Word n-gram

We observe the result of different amount on two different increments: 100-500 with increment of 100 (scenario 1) that is not displayed on the table, and 1000-5000 with increment of 500 (scenario 2). During scenario 1, any amount that is appended with metadata able to outperform most of the time compared to dataset without appended metadata. However, during scenario 2, the dataset without appended metadata more often than not outperform in certain metric. Looking at average F-Measure, we get dataset with appended metadata possess the highest score. We also get that at most we are able to reach 0.7932, 0.7542, and 0.7482 respectively for each metrics observed, obtained around the word n-gram amount of 3000-4000. Conscientiousness and extraversion on every metric reaches highest score among all, which might be caused by the imbalance of dataset.

5.2 Twitter Metadata

We observe that our selection of metadata unable to outperform preceding research with the significant difference in precision and recall and a slight one in f-measure. Since each paper has their own amount of metadata feature, the result of attribute selection is different as well. We suspect that our selection of metadata heavily affects the important linguistic feature during reduction, resulting low score among other list of feature. It is also found that conscientiousness and extraversion reaches the highest score while openness scores the least with its difference to agreeableness is quite significant.

5.3 Classifier

We observe that Random Forest and sequential minimal optimization performs the best among our selection of classifier with Naïve Bayes positioned the third. J48 and k-NN are close to reach 0.7, although both J48 and Random Forest are both tree-based classifier. This case also happens in [23] where Random Forest performs better than J48 by 10%. As mentioned, none of the classifiers used here are tuned, therefore it is possible for k-NN to be significantly sensitive towards outliers in the data. Random Forest is also more consistent in result throughout all metric than SMO.

Table 5. Result on how different combination of metadata affects the result

Dataset	P-AVG	R-AVG	F-AVG
Not appended	**0.7684**	0.725	0.718
Appended with our own list of metadata	0.745	0.7164	0.7152
Appended based on [4]	0.7642	**0.73**	**0.722**
Appended based on [17]	0.7526	0.7234	**0.722**
Appended based on [3]	0.7534	0.7206	0.7178

Table 6. Result on how different combination of metadata affects the result

Classifier	P-AVG	R-AVG	F-AVG
J48	0.7002	0.7032	0.6994
k-NN	0.6996	0.7004	0.6996
Naïve Bayes	0.745	0.7164	0.7152
Random Forest	0.744	**0.7484**	**0.744**
SMO	**0.7662**	0.747	0.7218

6. Conclusion

In this paper, we attempt to find optimal setting to perform personality prediction with Twitter as our source of the data using big five as the personality model, focusing on Indonesian user. We analyze three comparisons: amount of n-gram, twitter metadata, and classifier used.

- With the highest F-average of 0.7482, we obtain the optimal result around the usage of 3000-4000 word n-grams.
- While our list of Twitter metadata does not perform better than other lists, we find the difference is not that significant. The impact of not using list of metadata is not as well significant, although better.
- Random forest and SMO performs well with our dataset, suspecting that there are some outliers affected sensitive classifiers prone to it, such as k-NN.

The major issue of this research is the dataset. Not only there are heavy imbalances in some traits, the amount of dataset is perceived lacking. As a counterpoint, the lack of dataset is compensated by the validity of labeling as it is done manually by experts. In our future research, we expect to use even more dataset that is more balanced than what we currently have.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," *Psychological Science*, vol. 21, no. 3, pp. 372-374, 2010. https://doi.org/10.1177/0956797609360756

[2] Y. Wu, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036-1040, 2015. https://doi.org/10.1073/pnas.1418680112

[3] B. Plank and D. Hovy, "Personality traits on twitter—or—how to get 1,500 personality tests in a week," in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Lisbon, Portugal, 2015, pp. 92-98. http://dx.doi.org/10.18653/v1/W15-2913

[4] We Are Social, "Digital 2019: Indonesia," Available https://datareportal.com/reports/digital-2019-indonesia

[5] We Are Social, "Digital in 2018 in Southeast Asia," Available https://slideshare.net/wearesocial/digital-in-2018-in-southeast-asia-part-2-southeast-86866464

[6] We Are Social Singapore, "Digital in 2017: Southeast Asia," Available https://www.slideshare.net/wearesocialsg/digital-in-2017-southeast-asia

[7] We Are Social Singapore, "Digital in 2016," Available https://www.slideshare.net/wearesocialsg/digital-in-2016/224

[8] We Are Social Singapore, "Digital, Social & Mobile in 2015," Available https://www.slideshare.net/wearesocialsg/digital-social-mobile-in-2015

[9] F. Celli and B. Lepri, "Is big five better than MBTI? A personality computing challenge using Twitter data," in *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, 2018.

[10] P. T. Costa and R. R. McCrae, "The NEO personality inventory," *Journal of Career Assessment*, vol. 3, no. 2, pp. 123-139, 1985.

[11] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175-215, 1992. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[12] E. O. Whaite, A. Shensa, J. E. Sidani, J. B. Colditz, and B. A. Primack, "Social media use, personality characteristics, and social isolation among young adults in the United States," *Personality and Individual Differences*, vol. 124, pp. 45-50, 2018. https://doi.org/10.1016/j.paid.2017.10.030

[13] L. E. Annisette and K. D. Lafreniere, "Social media, texting, and personality: a test of the shallowing hypothesis," *Personality and Individual Differences*, vol. 115, pp. 154-158, 2017. https://doi.org/10.1016/j.paid.2016.02.043

[14] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1289-1295, 2010. https://doi.org/10.1016/j.chb.2010.03.018

[15] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Modeling and User-Adapted Interaction*, vol. 26, no. 2-3, pp. 109-142, 2016. https://doi.org/10.1007/s11257-016-9171-0

[16] 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 *Proceedings of 2012 11th International Conference on Machine Learning and Applications*, Boca Raton, FL, 2012, pp. 386-393. https://doi.org/10.1109/ICMLA.2012.218

[17] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, 2018, pp. 606-611.

[18] 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. https://doi.org/10.1177/0261927X09351676

[19] M. Coltheart, "The MRC psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, no. 4, pp. 497-505, 1981. https://doi.org/10.1080/14640748108400805

[20] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of per-

sonality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457-500, 2007.

[21] V. Ong, A. D. Rahmanto, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi, "Personality prediction based on Twitter information in Bahasa Indonesia," in *Proceedings of 2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Prague, Czech Republic, 2017, pp. 367-372. https://doi.org/10.15439/2017F359

[22] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing tasks in Indonesian Twitter messages," *Journal of Physics: Conference Series*, vol. 801, article no. 012072, 2017. https://doi.org/10.1088/1742-6596/801/1/012072

[23] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Proceedings of 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing*, Boston, MA, 2011, pp. 149-156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

[24] N. Haan, R. Millsap, and E. Hartka, "As time goes by: change and stability in personality over fifty years," *Psychology and Aging*, vol. 1, no. 3, pp. 220-232, 1986. https://doi.org/10.1037/0882-7974.1.3.220

[25] B. W. Roberts and D. Mroczek, "Personality trait change in adulthood," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 31-35, 2008. https://doi.org/10.1111/j.1467-8721.2008.00543.x

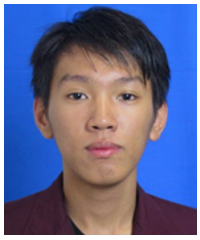
[26] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, 1999.

[27] B. J. M. Abma, "Evaluation of requirements management tools with support for traceability-based change impact analysis," M.S. thesis, University of Twente, Enschede, The Netherlands, 2009.

[28] F. J. Valverde-Albacete, J. Carrillo-de-Albornoz, and C. Pelaez-Moreno, "A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks," in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Heidelberg: Springer, 2013, pp. 41-52.

[29] F. A. S. Borges, R. A. Fernandes, A. M. Lucas, and I. N. Silva, "Comparison between random forest algorithm and J48 decision trees applied to the classification of power quality disturbances," in *Proceedings of the International Conference on Data Mining (DMIN)*, Las Vegas, NV, 2015, pp. 146-147.

Nicholaus Hendrik Jeremy is an undergraduate from Bina Nusantara University (BINUS). On March 2019 until February 2020, he has the responsibility to be a research assistant at BINUS School of Computer Science. His research interest includes natural language processing and personality prediction.
E-mail: nicholaus.jeremy@binus.ac.id



Cristian Prasetyo is a student of Computer Science of Bina Nusantara University (BINUS). On March 2019 until February 2020, he has the responsibility to be a research assistant at BINUS School of Computer Science. His research interest includes artificial intelligence, natural language processing, and linguistics.
E-mail: nicholaushendrik@gmail.com

Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He got his PhD degree in computer science from Universitas Indonesia in 2018. His research fields are natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.
E-mail: dsuhartono@binus.edu