Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 473–480

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia

Gabriel Yakub N.N. Adi, Michael Harley Tandio, Veronica Ong, Derwin Suhartono*

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
Email: {gabriel.adi, michael.tandio, veronica.ong}@binus.ac.id, dsuhartono@binus.edu

Abstract

This paper presents optimization techniques for automatic personality recognition (APR) based on Twitter in Bahasa Indonesia, the mother tongue of Indonesians. Foremost, we discuss Twitter and its utilization as a resource for many types of research. Several previous studies have been attempted to predict users' personality automatically. However, only a few of them have done their research for Bahasa Indonesia data. Therefore, this paper discusses the optimization of APR in Bahasa Indonesia. We evaluate a series of techniques implementing hyperparameter tuning, feature selection, and sampling to improve the machine learning algorithms used. The personality prediction system is built on machine learning algorithms. There are three machine learning algorithms used in this study, namely Stochastic Gradient Descent (SGD), and two ensemble learning algorithms, Gradient Boosting (XGBoost), and stacking (super learner). By implementing this series of optimization techniques, the current study's evaluation results show huge improvement by achieving 1.0 ROC AUC score with SGD and Super Learner.

Keywords: Automatic Personality Recognition; Big Five; Optimization; XGBoost; Super Learner

* Corresponding author. Tel.: +6221-534-5830; fax: +6221-530-0244.
E-mail address: dsuhartono@binus.edu

1877-0509 © 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the 3rd International Conference on Computer Science and Computational Intelligence 2018.
10.1016/j.procs.2018.08.199

1. Introduction

The massive growth of social networks involves billions of people to socialize around the internet by expressing their feelings, thoughts, and opinions. By 2017, statistics show there are 2.46 billion of social network users worldwide, meaning that 33.6% out of 7.3 billion of the worldwide population are social network users. On that note, social networks become an inseparable part of the internet. In the midst of the number of social networks available on the internet, Twitter has managed to become one of the most popular and active social networks. According to Statista, Twitter has 366 million monthly active users, and this number is constantly rising over the years.

Such massive amount of social network data could be utilized as resources for researchers from various fields to gain in-depth knowledge as well as improving services or products within many fields of interests, such as Computer Science. Predicting stock market [1], real-time event detection by social sensors [2], information filtering [3], spam detection [4] are few forms of utilization from Twitter data alone.

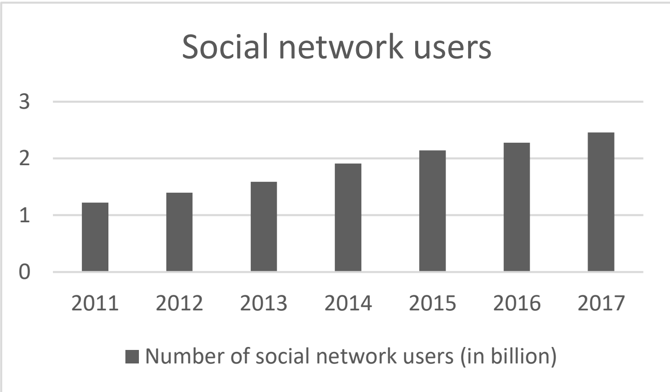

Figure 1 Number of worldwide social network users in billion.

One form of the utilization of knowledge extracted from social network data is Automatic Personality Recognition (APR). The purpose of APR is to automatically classify a person's personality trait using a personality model such as Big Five, and Myers-Brigg Type Indicator (MBTI) [5 6]. To automatically predict a user's personality, researchers have applied APR for various problems namely personality prediction from Facebook profile pictures [7], computational personality recognition [8], etc. However, there is still only a few studies done covering APR in Bahasa, the mother tongue of Indonesia, even though Indonesia is a strong social media consumer. An indication of high Twitter usage from Indonesia has been presented in a study which reported that 77.7% of internet users in Indonesia were on Twitter in the year 2012. It also reported that during the first quarter of the year 2014, 2.4% of the world's Twitter posts originated from Jakarta, the capital city of Indonesia [9].

With these considerations in mind, this research focuses on automatically recognize a person's trait from Twitter content in Bahasa Indonesia. Challenges to perform APR in Bahasa Indonesia are present as only limited tools, and a small number of datasets are available. To tackle such challenge, we perform APR with two ensemble learning methods, boosting and stacking, a standard machine learning algorithm, and their optimization to improve performance measured in receiver operating characteristic area under the curve (ROC AUC) score.

2. Literature Review

Automatic Personality Recognition uses well-known personality model as an approach to identify user's personality. One of them is the five-factor model (FFM), which also known as the Big Five (BIG5).

2.1. The Big Five Model

FFM is a personality model created by McCrae and Costa [5]. FFM is the dominant approach for representing the human trait structure today [10]**.** The model comprises five different personality traits which are, Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism [11] (OCEAN).
The following are the characteristics of each personality trait:

Table 1. BIG5 Personality Traits

Trait \ Label	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
High	Willing, obliging	Orderly, organized	Confident, bold	Negative emotional state, feelings of guilt, envy, anxiety	Non-conventional, inventive, inquisitive
Low	Reluctant, unaccommodating	Careless, equable	Diffident, shy	Emotional stability, calm	Conventional, consistent, vigilant

2.2. Automatic Personality Recognition

Several approaches to automatic personality recognition have been attempted in the past based on data from different social media platforms. One of the commonly used approaches in automated personality recognition is the use of LIWC, a pre-defined dictionary for text analysis, specifically on thoughts, feelings, personality, and motivations. Studies utilizing the LIWC have proved to be useful on different social media platforms[12], such as Twitter [13–15], Blogger [16], Facebook [17,18], YouTube [19], and Weibo [20]. While useful, LIWC's usage cannot be expanded to all languages, as it is only available in select languages. As the current study attempts to build a personality recognition system based on the Bahasa Indonesia language, the LIWC cannot be applied explicitly in this study. One of the approaches to solving this problem is the open-vocabulary approach, in which researchers attempt to build a machine learning model to identify user personality without a pre-defined dictionary such as LIWC. The open-vocabulary approach assesses the choice of words from a user. This approach has also been implemented by several studies [20–22]. Studies attempting to build an automated personality recognition system for Bahasa Indonesia have also been implemented by [23] and [24].

Besides text data, there have also been recent studies attempting to merge multiple data types for personality recognition. While [25] is only based on the Weibo social media, the study leverages multiple data types including Weibo posts, avatars, emoticons, and responsive patterns (the user's interactions with other users). Another study [26] uses of Instagram and Twitter data. The study analyzes text data (tweets, Instagram captions), user behavior features (number of followers and followings), as well as image data to identify a user's personality. A study by [27] built a personality recognition system by merging data from Facebook, Twitter, and YouTube. A set of text, demographics, and user behavior data were leveraged from Facebook, various user data such as text, age, and gender were extracted from Twitter, while text, audio-video features, and gender were utilized from YouTube platform.

Recent advances in deep learning have also been implemented in personality recognition to improvize its performance. [28], [25], [29] and [30] are among the studies which utilize deep learning methods such as Convolutional Neural Networks, Fully-connected Neural Networks, and Recurrent Neural Networks to improve text representation.

2.3. Optimization for Machine Learning

Recent studies for the machine learning optimization have been done, where the results show that there are methods that could be used such as ensemble learning, feature selection, hyperparameter tuning and sampling [31,32]. Ensemble learning is a method used to obtain a higher classifier accuracy by combining less accurate algorithm with a higher one. One of the very first methods is Bayesian averaging. Other recent algorithms are [33] bagging, boosting, and stacking. Boosting can be implemented using XGBoost, whereas stacking can be implemented using Super Learner [34,35].

Table 2. Recent ensemble learning methods

Bagging	Boosting	Stacking
A bootstrap aggregation to reduce the variance of an estimate by averaging the multiple estimates	An algorithm that boosts or converts a weak learner to strong learner	A technique that combines multiple classification models with a meta-classifier

The purpose of feature selection is to improve the interpretability and performance of a predictive model [31]. There are two categories for feature selection known as, filter methods and wrapper methods. Filter method uses general characteristics of the data to evaluate and to select the subsets of the feature. Wrapper method uses the performance of the chosen learning algorithm to evaluate each candidate feature subset [36].

Hyperparameter tuning is a way to optimize any learning algorithm, pre-processing and post-processing methods for any task. Several optimization techniques that can be used are random search, grid search, evolutionary algorithm, iterated F-racing, sequential model-based optimization [31], and Bayesian optimization. Bayesian Optimization uses Gaussian process as it's algorithm, and can perform faster than an expert at doing hyperparameter manually [37].

To overcome imbalance dataset, one of the most common technique is sampling. There are two methods, under-sampling and over-sampling. Under-sampling removes the majority of the class, while over-sampling uses the same minor class repeatedly to match the majority class quantity [38].

3. Data and Approach

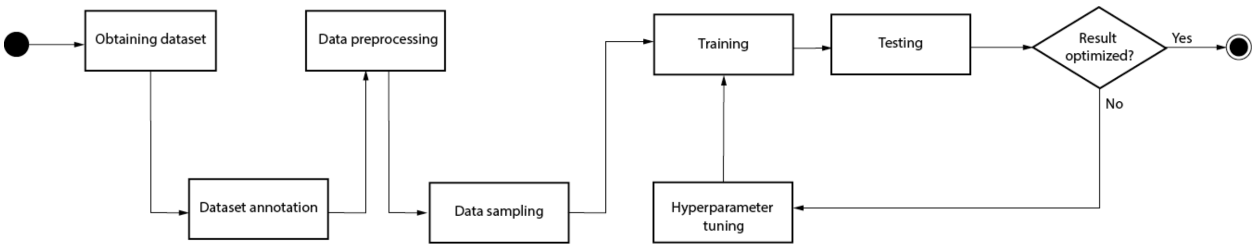

Figure 2 Overview of data and approach

3.1. Data training

The dataset used were taken from [24], comprising of 250 data containing Twitter data in Bahasa Indonesia. Each data represents a series of Twitter data from 1 user. In this study, the dataset was then expanded with 50 new data which have been extracted from Twitter using Twitter API. The users consisted of the new data are selected manually to get Bahasa Indonesia speaking users only. The implementation of this additional data was adapted and annotated by the same psychology experts from the previous study. This data extraction results in a dataset comprising of Twitter information from 300 users.

The dataset was then split into two parts, namely the features and the target variable. Twitter data is set as the input and annotation data as the target variable. Twelve features were selected as the input to train the machine learning model, whereas the output was separated into two labels, high and low. Each label is converted to a numerical representation of 1 (High) and 0 (Low).

Table 3. Input and output from twitter dataset

Input			Output
Tweets count	Retweets count	Replies count	HIGH
Followers count	Retweeted count	Hashtags count	LOW
Following count	Quotes count	URL count	
Favorites count	Mentions count	Tweet content	

The dataset then preprocessed to reduce noise using automated removal of several elements which involves replacing mentions with “[UNAME]” token, replacing hashtag with “[HASHTAG]” token, replacing ampersand with “[HASHAND]” and dot with “[HASHDOT]”, removing URL/hyperlink, removing emojis, removing retweet contents, and stop words omission. The list of stop words was taken from [39]. Next, tokenization was applied to the tweet content by taking the unigrams and bigrams of said content. The Term Frequency (TF) weighting scheme was implemented by counting each unigram’s and bigram’s number of occurrences to be used as features.

Due to the high level of imbalance on the dataset, sampling was done to gain 1:1 ratio between high and low label for each trait. Therefore, not all the data was used for the training. Tree-based feature selection was then applied using classifier built on Decision Tree to reduce the high dimensionality of the features.

Table 4. Dataset before sampling

	AGR	CON	EXT	NEU	OPN
HIGH	134	34	230	67	143
LOW	166	266	70	233	157

Table 5. Dataset after sampling

	AGR	CON	EXT	NEU	OPN
HIGH	134	34	70	67	143
LOW	134	34	70	67	143

3.2. Build prediction model

Five classifiers were built for the personality prediction system, one classifier for each of the trait in the Big Five model. Each classifier was built on a machine learning algorithm. The first algorithm is Stochastic Gradient Descent (SGD). The second and third algorithm are ensemble learning algorithms, namely boosting and stacking implemented using XGBoost (XGB), and Super Learner (SL). The Super Learner is built by stacking Logistic Regression, XGB, and SGD consecutively. All classifiers were run on Python.

Due to high imbalance level of the dataset, it was sampled to achieve 1:1 high and low label ratio. The dataset was then split to 70% of training and 30% for testing. The result is measured in ROC AUC score. Each classifier, except SL, was run on ten iterations of hyperparameter tuning adapted using Bayesian Optimization to find the optimal hyperparameter for each classifier.

4. Result and Discussion

In this research, we emphasize on the optimization of APR by implementing three different techniques. Whereas previous study highlight methods for performing APR on contents in Bahasa Indonesia. Series of training and testing was done using different scenarios to monitor the improvements of the machine learning performance. The system is tested on different scenarios with following actions:

1. No optimization algorithms
   Training and testing were done without any optimization technique.
2. Hyperparameter tuning (HPT)
   The training and testing were gone through 10 iterations of hyperparameter tuning using Bayesian Optimization. The hyperparameters which were tuned varies with different machine learning algorithms.
3. Hyperparameter tuning and feature selection (HPT + FS)
   The highest 1000 was selected from the input before it goes through the hyperparameter tuning iterations.
4. Hyperparameter tuning, feature selection and sampling (HPT + FS + Sampling)
   The input and output data was sampled to the lowest amount of output label to achieve 1:1 ratio between high and low value distribution. The sampled input then goes through features selection functions and finally hyperparameter tuning iterations.

Table 6. Evaluation Results

Scenario	Algorithm	ROC AUC					
		AGR	CON	EXT	NEU	OPN	Average
1	SGD	0.604	0.500	0.508	0.500	0.571	0.537
	XGB	0.613	0.500	0.533	0.503	0.572	0.544
	SL	0.739	0.500	0.500	0.500	0.612	0.570
2	SGD (HPT)	0.766	0.565	0.631	0.545	0.667	0.635
	XGB (HPT)	0.797	0.615	0.617	0.611	0.702	0.668
3	SGD (FS)	0.726	0.500	0.538	0.500	0.656	0.584
	XGB (FS)	0.716	0.500	0.514	0.510	0.623	0.573
	SL (FS)	0.934	0.700	0.780	0.620	0.923	0.791
4	SGD (FS + HPT)	0.794	0.611	0.585	0.672	0.790	0.690
	XGB (FS + HPT)	0.834	0.611	0.639	0.657	0.788	0.705
5	SGD (FS + HPT + Sampling)	0.986	1.000	1.000	1.000	0.966	0.990
	XGB (FS + HPT + Sampling)	0.887	0.937	0.785	0.8530.	0.770	0.846
	SL (FS + Sampling)	1.000	1.000	1.000	1.000	0.960	0.992

$$TP\ Rate = \frac{TP}{TP+FN}\ x\ 100\% \quad (1)$$

$$FP\ Rate = \frac{FP}{FP+TN}\ x\ 100\% \quad (2)$$

The evaluation metric used to measure the performance is ROC AUC. First, ROC Curve is generated by plotting True Positive Rate (TPR) against False Positive Rate (FPR). TPR and FPR are determined by computing the ratio of true positive of all positive data points (1) and false positive of all negative data points (2) respectively at multiple thresholds to cover the low and high values of FPR. AUC, then, is measured by calculating the area under the ROC curve. The maximum and minimum value of ROC AUC is 1.0 and 0.5 respectively, 1.0 would be a perfect score, whereas 0.5 is identical to guessing.

The evaluation results are shown in Table 6. The result from Table 6 shows that without any optimization, all the prediction models perform poorly even with ensemble learning algorithms, both boosting and stacking. This result is due to the highly imbalanced classes, and the high dimensionality of features on the dataset. The feature dimension of the input in the first two scenarios could reach over 40000 features generated by combining unigram and bigram alone. Meanwhile, using features without including n-grams does not provide enough feature to train the model. Therefore, this problem needs to be addressed.

The performance of prediction models, SGD and XGB, improved slightly with hyperparameter tuning as the average evaluation scores increase by around 0.1 margins. The utilization of feature selection, while not resulting in significant improvement (average of 0.15 points compared to the first scenario), significantly reduces the training time.

Finally, the system utilized all the optimization algorithm which combines FS, HPT, and sampling. This scenario results in the best improvement of all, with the highest score of 1.0 ROC AUC score. The last scenario successfully balances the dataset and reduce the dimensionality of the feature by only selecting the best 1000 features to decrease the noise of the data.

Sampling diminishes the dataset which causes missing information for the model to learn. This missing information leads to bias on such high score of ROC AUC since the model training, especially when training conscientiousness trait, exclude the majority of the dataset. The result might differ when populated with a larger and more balanced dataset.

5. Conclusion

In conclusion, we successfully improved machine learning algorithms to predict users' personality automatically. Hyperparameter tuning, feature selection, and sampling managed to tackle the imbalance and noise of the dataset. Though there might be biases in the result due to the extremely small size of the dataset after sampling. These approaches could be adapted by other studies to overcome the same problem. Further improvement could be made by expanding the dataset and utilizing deep learning algorithms to predict users' personality automatically.

Acknowledgements

Thanks to 3 experts from Psychology Department Bina Nusantara University who are involved in annotating the data for this research: Dr. Aryo E. Nugroho, Dr. Esther Widhi Andangsari and Dr. Nanang Suprayogi.

References

1. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011;2(1):1–8.
2. Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. Proc 19th Int Conf World Wide Web [Internet]. 2010;851–60. Available from: http://doi.acm.org/10.1145/1772690.1772777
3. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short Text Classification in Twitter to Improve Information Filtering. Proc 33rd Int ACM SIGIR Conf Res Dev Inf Retr SE - SIGIR '10 [Internet]. 2010;(July):841–2. Available from: citeulike-article-id:7573100%5Cnhttp://dx.doi.org/10.1145/1835449.1835643
4. Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on twitter. Collab Electron Messag anti-abuse spam Conf [Internet]. 2010;6:12. Available from: ceas.cc/2010/papers/Paper 21.pdf%5Cnhttp://ceas.cc/2010/papers/Paper 21.pdf
5. McCrae RR, Costa PT. Validation of the five-factor model of personality across instruments and observers. J Pers Soc Psychol. 1987;
6. Type TM, Mbti I, Briggs I. Myers-Briggs Type Indicator (MBTI). New York. 1985;
7. Celli F, Bruni E, Lepri B. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In: Proceedings of the ACM International Conference on Multimedia - MM '14 [Internet]. 2014. p. 1101–4. Available from: http://dl.acm.org/citation.cfm?doid=2647868.2654977
8. Farnadi G, Sitaraman G, Sushmita S, Celli F, Kosinski M, Stillwell D, et al. Computational personality recognition in social media. User Model User-adapt Interact. 2016;
9. Carley KM, Malik M, Kowalchuk M, Pfeffer J, Landwehr P. Twitter Usage in Indonesia. Cent Comput Anal Soc Organ Syst. 2015;1–54.
10. Roccas S, Sagiv L, Schwartz SH, Knafo a. The Big Five Personality Factors and Personal Values. Personal Soc Psychol Bull. 2002;28(6):789–801.
11. Pramodh KC, Vijayalata Y. Automatic personality recognition of authors using big five factor model. In: 2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016. 2017. p. 32–7.
12. Ong V, Rahmanto ADS, Williem, Suhartono D. Exploring Personality Prediction from Text on Social Media: A Literature Review. INTERNETWORKING Indones J. 2017;9:65–9.
13. Golbeck J, Robles C, Edmondson M, Turner K. 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. IEEE; 2011. p. 149–56.
14. Sumner C, Byers A, Boochever R, Park GJ. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: ICMLA '12 Proceedings of the 2012 11th International Conference on Machine Learning and Applications. IEEE; 2012. p. 386–93.
15. Arroju M, Hassan A, Farnadi G. 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. Toulouse, France; 2015.
16. Yarkoni T. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. J Res Pers. 2010;44(3):363–73.
17. Farnadi G, Zoghbi S, Moens M-F, De Cock M. Recognising personality traits using Facebook status updates. In: Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13). AAAI; 2013.

18. Tandera T, Hendro, Suhartono D, Wongso R, Prasetio YL. Personality Prediction System from Facebook Users. 2nd Int Conf Comput Sci Comput Intell 2017. 2017;605–11.
19. Farnadi G, Sushmita S, Sitaraman G, Ton N, De Cock M, Davalos S. A multivariate regression approach to personality impression recognition of vloggers. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition. ACM; 2014. p. 1–6.
20. Xianyu H, Xu M, Wu Z, Cai L. Heterogeneity-Entropy Based Unsupervised Feature Learning For Personality Prediction With Cross-Media Data. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). 2016. p. 1–6.
21. Iacobelli F, Gill AJ, Nowson S, Oberlander J. Large Scale Personality Classification of Bloggers. In: D'Mello S, Graesser A, Schuller B, Martin J-C, editors. Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9--12, 2011, Proceedings, Part II. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 568–77.
22. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality , Gender , and Age in the Language of Social Media : The Open-Vocabulary Approach. 2013;8(9).
23. Pratama BY, Sarno R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In: 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE; 2015. p. 170–4.
24. Ong V, Rahmanto ADS, Williem, Suhartono D, Nugroho AE, Andangsari EW, et al. Personality Prediction Based on Twitter Information in Bahasa Indonesia. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. Prague, Czech Republic; 2017. p. 367–372.
25. Wei H, Zhang F, Yuan NJ, Cao C, Fu H, Xie X, et al. Beyond the words: Predicting user personality from heterogeneous information. In: Proceedings of the tenth ACM international conference on web search and data mining. ACM; 2017. p. 305–14.
26. Skowron M, Tkalčič M, Ferwerda B, Schedl M. Fusing social media cues: personality prediction from twitter and instagram. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee; 2016. p. 107–8.
27. Farnadi G, Sitaraman G, Sushmita S, Celli F, Kosinski M, Stillwell D, et al. Computational personality recognition in social media. User Model User-Adapted Interact. 2016;26(2–3):109–42.
28. Majumder N, Poria S, Gelbukh A, Cambria E. Deep learning-based document modeling for personality detection from text. IEEE Intell Syst. 2017;32(2):74–9.
29. Siddique F Bin, Fung P. Bilingual Word Embeddings for Cross-Lingual Personality Recognition Using Convolutional Neural Nets. Learning. 2017;21:22.
30. Yu J, Markov K. Deep learning based personality recognition from Facebook status updates. In: Awareness Science and Technology (iCAST), 2017 IEEE 8th International Conference on. IEEE; 2017. p. 383–7.
31. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J Mach Learn Res. 2016;
32. Byrd RH, Chin GM, Nocedal J, Wu Y. Sample size selection in optimization methods for machine learning. In: Mathematical Programming. 2012.
33. Dietterich TG. Ensemble Methods in Machine Learning. MCS '00 Proc First Int Work Mult Classif Syst [Internet]. 2000;1–15. Available from: http://www.cs.orst.edu/~tgd
34. Chen T, He T. xgboost : eXtreme Gradient Boosting. R Packag version 04-2. 2015;
35. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. Stat Appl Genet Mol Biol. 2007;
36. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;
37. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. Adv Neural Inf Process Syst 25. 2012;
38. Longadge R, Dongre SS, Malik L. Class imbalance problem in data mining: review. Int J Comput Sci Netw. 2013;
39. Tala FZ. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. MSc Thesis, Append D. 2003;pp:39–46.