Systematic Literature Review of Profiling Analysis Personality from Social Media

Idris¹, E Utami², A D Hartanto³ and S Raharjo⁴

¹,²,³Magister of Informatics Engineering Universitas Amikom Yogyakarta, Yogyakarta Indonesia, ⁴ Department of Informatics Engineering, IST AKPRIND Yogyakarta Indonesia

¹idriscoklat@gmail.com,²emma@nrar.net,³anggit@amikom.ac.id, ⁴wa2n@akprind.ac.id

Abstract. The success or failure of a company is usually supported by the presence of reliable resources, especially human resources. Recruitment and placement of employees in the right position will have a significant impact on a company. Human nature and character are very diverse in their forms. DISC theory classifies personality into four types namely dominance, influence, steadiness, and compliance. The difference in the character of each type will of course affect the behavior style, how to deal with life pressures, and also how to communicate both directly and with social media. Through social media, a person can vent his feelings through the posts he uploaded. From these posts an analysis of the personality character he is possessed can be carried out. This systematic literature review is used to analyze and focus on techniques for conducting personality profiles through the use of social media. From this literature review of analysis, it can be obtained that the various personality classification methods and algorithms can provide a good level of accuracy.

1. Introduction

An organization, agency, or company will run well if supported by reliable resources. One important resource is human resources. Appropriate employee recruitment will affect the company's performance. If the performance increases, the target can be achieved, the production rate, and sales also increase. Appropriate employee recruitment process requires several stages, one of which is the prospective test to find out how high a match between the prospective employee and the employee's character needed by the company. Personality tests are made to provide a general picture of a person's style and behavior as completely as possible. There are several methods used to find out a person's character and personality. The first test method is MBTI (Myers-Briggs Type Indicator) that divided into four dichotomies namely introversion-extraversion, intuition-sensing, feeling-thinking and perception-judging which then the results of the test will be included in one of sixteen personalities. The second method is the OCEAN method or Big Five Personality Model which analyzes thirty traits in personality under openness, conscientiousness, extraversion, agreeableness, and neuroticism. The third personality test method is the DISC (Dominance, Influence, Steadiness, and Compliance).

The DISC method provides an overview of a person's style that can predict future behavioral tendencies. It is obtained by evaluating the main personality factors that exist in a person. If a complete personality test kit often contains hundreds of questions and takes a long time to complete, the DISC test kit is only a questionnaire containing 24 questions and can be completed in about 15 minutes. The questionnaire can be presented on paper or in the form of applications so that the determination of the value can be automated so that the process is much faster and efficient. This
DISC test usually measures how one's behavior style, how to communicate, how to deal with pressure, and so on. People who know themselves and understand DISC will be easy to adapt to others' world who are different from him, the friction will decrease and the stress level decreases so that productivity rises and life success also increases [1].

In the past decade, internet users around the world have grown significantly [2]. We can find information about many things on the internet, especially information about someone through their social media pages. Data on social media also vary in number, such as status, tweets, books you like, the music you like, etc. So, we can analyze the data and process it into more useful information. Twitter is a frequently used object of research because of its fast nature because of the limited number of characters to a maximum of 280 characters, so there is no "read more" feature because all characters have appeared on the screen. Therefore, information that comes very quickly. Profiling is a technique that processes personal and non-personal data that is used to develop predictive knowledge in building a profile that is then applied based on decision making [3]. The data retrieved can be in the form of text data, images, user activities, etc.

2. Related Works

Some studies that examine the relationship between personality and posting on social media are conducted by [4] which collect Facebook status data on 100 applicants and a maximum of 5 statuses taken from each account. By using the backpropagation algorithm, personality classifications get an error value with a min error of 0.00001% in tests 2 and 3. The results of tests that have been carried out the performance of the system with the highest accuracy of 84.00% are obtained in the proportion of 70:30% data tested. The research done by [5] analyze the tweet to find out someone's personality that can be utilized in employee recruitment. By using the Naïve Bayes method, this study yields very good accuracy in classifying personality after comparison with the results of the classification from experts. Research [6] conducts research aimed at classifying personality types through social media using K-Nearest Neighbor and Big Five theory. The results of the study showed that the accuracy rate reached 90%. Of the five personalities, the class that has the highest accuracy value is O, E, A, N which is 100% while the class that has the lowest accuracy is in C which is only 60%. This is because C has a tendency similar to class E. Research conducted by [7] aimed at finding out the character of prospective employees based on the tweet to be processed to produce MBTI personality traits as a reference for the placement of prospective employees. From these results it is known that the K-Nearest Neighbor (KNN) algorithm can be implemented in the personality classification system. The total data used is 160 data divided into two by sharing 50% for training data and 50% for test data. The final results show that, the KNN algorithm produces an accuracy value of 66%. The classification of the personality of Twitter users into extrovert or introvert classes by using the Support Vector Machine (SVM) algorithm is carried out by [8]. The result is the use of the SVM method to classify the personality of Twitter users into extroverts and introverts using as many as 17 features that have been adapted to the characteristics of Twitter social media. After doing a grid search and cross-validation, the best parameter pairs are C = 1048576 and γ = 3.814697265625e-06 in the data distribution of 80:20. The test results with the best model obtained an accuracy of 88.89%.

This research aims to find out the relationship between Big Five personality and online self-disclosure shows that there is one dimension of Big Five personality, which is the openness to experience which affects online self-disclosure on social media users. The effective contribution magnitude of the Big Five personality dimension is the openness to experience of 0.193 (19.3%) and can still affect other variables that affect self-expression in addition to personality variables such as gender, ethnicity, nationality, age, dyadic effect, large topic groups, and recipient relationships. Of the five Big Five personalities who have a significant positive relationship is the openness to experience. The dimensions of personality extraversion, agreeableness, conscientiousness and neuroticism do not show a significant relationship [8].

Classification of one's personality type into the Big Five personality theory based on tweets that have been made by Twitter users in Indonesia is done [9]. The results of this study indicate that it is possible to determine the personality type of Twitter users without literacy tools (LIWC, MRC), but
based on the diction written by the user of the object of this study. The current study compares two different classifications, Support Vector Machine and XGBoost. Both algorithms are tested in different scenarios involving the occurrence of minimum n-gram, n-gram weighting scheme, use of LDA features, and removal of stopwords. Evaluation using 10-fold cross-validations shows that the personality prediction system based on SVM managed to achieve the highest average accuracy of 76.2310%, while XGBoost reached 97.9962%.

Research [10] aims to build a Big Five personality prediction model of Twitter users using Support Vector Regression (SVR). The analysis was conducted on the social behavior of Twitter users and the use of linguistics when writing tweets and biographies to determine the features that best form the learning model predicting the user's personality. Overall, the personality of Big Five Twitter users can be well predicted using a combined model of social behavior features and linguistic features with the open-vocabulary bigram method. Researchers in [11] conducted research aimed at finding out whether there were differences in communication privacy management on Twitter among extroverted and introverted teens. The results of hypothesis testing in this study indicate the number 0.01. This shows that there are differences between extrovert and introvert types, the results of the study show that extroverted types express their privacy more on Twitter than introverted types.

The classification of Facebook users' personalities into the five factors personality was carried out by [12] using the Boost-Decision Tree algorithm and produced an Accuracy of 82.2%. Researchers got 117 participants and requested that the participants fill in some data and then get 100 valid participants. Then the researcher analyzes user habits through user activities on his Facebook, such as the number of likes, posts, friends, etc. The researcher uses several algorithms, namely Naïve Bayes, Neural Network, Decision Tree, SVM, Boosting-Naïve Bayesian, Boosting-neural network, Boosting-decision tree, and Boosting-SVM. Then compare all the algorithms and note that the Boosting-decision tree has the best accuracy. Next is research [13] that focuses on developing a lexical resource, which consists of dictionary abbreviations, contractions, non-standard words, and emoticons that are often used on social media. This Lexical resource will later be used to improve the accuracy of results in studies using neural network algorithms and data derived from social media. Lexical is made because of problems encountered when processing NLP, namely the number of slang (non-standard language). This research uses Twitter and English, Spanish, Dutch and Italian dictionaries as data sources. The results of this study indicate that by using these lexical resources, the accuracy quality of the Neural Network algorithm increases when used to run profiling.

The use of Twitter as an object of research is carried out by [14]. This study aims to provide alternatives to human resources in obtaining prospective employee personality data through his Twitter account. This study uses the Naïve Bayes Classifier algorithm with W-IDF Weighted-Inverse Document Frequency weighting) to classify the personality of prospective employees into the DISC personality model, as well as using training data and testing data of 120 accounts. Researchers refer to word bags that have been psychologically verified and have an accuracy of 36.67%. Research on Twitter posts is also carried out by [15] which aims to analyze whether profiling on Twitter can be an alternative for human resources in exploring the personality of prospective employees. Text processing was selected in this study and divided into four scenarios, namely stemmed-not weighted, stemmed-not weighted, stemmed-weighted, and stemmed-weighted. The results for each scenario are as follows: stemmed-not-weighted notes get an accuracy of 37.41%, stemmed-not weighted get an accuracy of 30.21%, stemmed-weighted notes get an accuracy of 35.97%, and stemmed-weighted gets an accuracy of 30.93%.

### 3. Methodology

In the process of recruiting prospective employees in a company, one's personality is very influential to determine the position occupied. In order to find out a person's personality is usually done by a personality test. Personality profiling analysis from social media can be used as a reference for recruiting and placing prospective employees so that an appropriate position is obtained. In profiling analysis, various studies try to analyze various types of data from one's social media. Some go through status [4], tweet [6][7][16][9][10][11][14][15], user activity [12], etc. to find the best accuracy. The methodology used for profiling analysis is explained in the following Figure 1.
Figure 1. Methodology used for profiling analysis.

Profiling analysis starts from collecting user data. Data collection can be done by distributing questionnaires on the internet. The questionnaire contains questions from the model chosen to classify personality (e.g., a questionnaire that contains 24 questions from the DISC model). The questionnaire is presented in a web form or web application so that everyone can access it. Prospective participants will answer the questions given and enter their social media account username after they submit the questionnaire, the system will determine the personality types of participants based on the questions they have answered so that researchers get participant social media account data and personality type information. After getting enough participants, the next step is to select the participants to get a valid participant. The screening contains rules made based on research[17] which includes checking participants' social media accounts whether or not there are, counting the total tweets or status in participant social media, if less than 500 tweets, then it is less valid. Another rule is that the system will record the time the user passes when filling out the questionnaire, if the time is more than 10 seconds then the participation is declared valid. After getting the participant and the participant's personality type data, the next process is data mining. Data mining is a series of processes to explore the added value of information that has not been known manually from a database/dataset [18]. In the data collection various ways are done, namely scraping [19], crawling[20], and streaming [21][22]. Scraping is the process of extracting information from a website automatically by parsing the HTML tag and only retrieving the information needed [19]. While crawling and streaming are data mining processes that are carried out through the Application Programming Interface (API) of social media.

Next is the process of cleaning the data that has been filtered. This process is commonly called preprocessing. Preprocessing is done through several steps, namely: Case folding, URLs elimination, username elimination, hashtags elimination, Tokenizing or separating sentences into words, Normalizing words, Stemming, POS Tagging, and POS Filtering. After preprocessing, training data and testing data is built. The data created must adjust the input of the algorithm. After the training data and testing data are created, the next step is to test the level of accuracy. Accuracy testing can be done by the k-fold cross-validation method.

4. Result And Discussion
In this section, the results of profiling analysis are based on a literature review that is used to create a personality profiling analysis design on social media for the recruitment process of prospective employees. This research uses several types of social media such as Facebook and Twitter as the source of the data as well as several algorithms to classify into several types of personalities to conduct comparative analysis. However, for this study the researchers limited one social media platform, Twitter, for data acquisition. Twitter was chosen because of its fast nature because of the limited number of characters to a maximum of 280 characters, so there is no "read more" feature because all characters have appeared on the screen. Therefore, information that comes very quickly. The collection of Twitter user tweets will be classified based on the personality of DISC.
In this literature review various algorithm techniques are used to analyze personality. The results of studies that have been carried out to analyze someone's Facebook status with the backpropagation algorithm for classifying their personalities produce good system performance and only get error values with very small error min of only 0.00001% and have the highest accuracy value of 84.00%. The results of personality analysis using the Naïve Bayes Classification method showed very good accuracy compared to the results of the classification from experts. For the MBTI personality analysis, the KNN algorithm also produces a pretty good accuracy even though it is only in the range of 66%. However, it can still be concluded that the KNN algorithm can be used to classify MBTI personalities as a basis for employee recruitment. Furthermore, other researchers used the SVM, XGBoost, and SVR methods to analyze Twitter user account tweets that were used to classify Big Five and Introvert-Extrovert personality types. The results obtained also indicate that these methods can produce a good level of accuracy so that they can classify personalities according to their respective research objectives.

Some research results that analyze Twitter user account tweets and also analysis of one's Facebook status shows that a person's personality can be known by analyzing what they reveal on social media either through Facebook or Twitter even though in classification using different algorithms but able produce a good level of accuracy. Thus, the personality analysis of one's Twitter account and Facebook can be used as a reference for the recruitment and placement of prospective employees. A detailed analysis of the comparative details of several researchers used as a review in the analysis of personality profiling on social media as a basis for recruiting prospective employees can be seen in the following Table 1.

Table 1. The Comparation Analytics of Review Process

| Reference | The Objective of The Paper | Conclusion | Suggestion/Weakness |
|-----------|----------------------------|------------|---------------------|
| Ref.[4]   | Knowing the personality of a job applicant through Facebook status consisting of social words, positive emotions, and negative emotions | The results of tests produced the highest accuracy of 84.00% using the proportion of 70:30% data splitting | In the study, the use of hidden layer neurons can not determine which neuron is best to determine the value, because neurons are trial and error |
| Ref.[5]   | Analyze tweets to find out someone's personality so that it can be useful in the employee recruitment process | Naïve Bayes Classification Algorithm can produce great accuracy in personality classification after comparing with the classification done by the experts | For the development of this research, it is necessary to increase the amount of training data and test data in conducting experiments so that the result can be more accurate |
| Ref.[6]   | The purpose of this study is to classify personality types through social media using K-Nearest Neighbor and Big Five theory | KNN algorithm can classify the personality of Twitter account users based on the Big 5 personality theory with an accuracy level that reaches 90%. | Future studies are expected to apply the clustering method to dig deeper data with the addition of several parameters such as the types of accounts that are followed, retweet types, and certain vocabulary that describe a person's personality. |
| Reference | The Objective of The Paper | Conclusion | Suggestion/Weakness |
|-----------|----------------------------|------------|---------------------|
| Ref.[7]   | To find out the character of prospective employees based on the tweet processed to produce MBTI personality traits. | This research produces an accuracy value of 66% with 53 data that correctly classified and 27 incorrectly classified. | Further research can be continued by using other accuracy testing methods, such as k-fold cross-validation. |
| Ref.[16]  | To classify Twitter user personality into an extrovert or introvert class by using the SVM method. | Used 17 The features that are adjusted to the twitter characteristics. After doing a grid search and cross-validation, the best parameter pairs are $C = 1048576$ and $\gamma = 3.814697265625e-06$ in 80:20 data distribution. The best model obtained an accuracy of 88.89% of the test result. | This research can be improved by using other features that have not been used, for example words contained in a tweet. With the increasing number of features used, it is recommended to implement the feature selection step using the F-score. Also, the amount of data used is recommended to be expanded, for example reaching 10,000 accounts to obtain the desired corpus. |
| Ref.[8]   | This study aims to determine the relationship between Big Five personalities with the online self-disclosure. | The results showed that there is one dimension of the big five personalities, which is the openness to experience which affects online self-disclosure on social media users. The amount of effective contribution of that personality dimension is 0.193 (19.3%) and there are still other variables that influence self-expression in addition to personality variables such as gender, ethnicity, nationality, age, dyadic effect, topic size, and recipient relationship. | Research subjects are advised to improve their openness to experience characteristics, such as being more open with some information and wanting to learn something new. And for similar studies, it is advisable to look at other factors that influence self-disclosure in addition to external and internal factors and it is expected to study more sources of references that related to this research. |
| Ref. [9]  | Aims to classify a person's personality type into the Big Five personality theory, based on tweets that have been made by the users in Bahasa. | The results of this study indicate that it is possible to determine the personality type of Twitter users without literacy tools (LIWC, MRC), but based on the diction written by the user of the object of this study. This study compares 2 algorithms, namely Support Vector Machine and XGBoost. Evaluations using 10-fold cross-validations show that SVM managed to achieve the highest average accuracy of |
|           |   |   | • The future development of this study is expected to be able to utilize a larger training and testing dataset, which allows the system to understand a wider variety of tweets.  
|           |   |   | • Improving the n-gram normalization function can also improve system accuracy because it allows the system to recognize and rate more
| Reference | The Objective of The Paper | Conclusion | Suggestion/Weakness |
|-----------|---------------------------|------------|---------------------|
| Ref.[10]  | This study aims to build a Big 5 personality prediction model of Twitter users using SVR. | 76.2310%, while XGBoost reached 97.9962%. | To use the closed-vocabulary method, the construction of a word dictionary needs to involve experts in the field of linguistics, especially Bahasa and English. Construction time can be extended so that more vocabulary is covered. |
| Ref.[11]  | This study aims to determine whether there are differences in communication privacy management on social media Twitter in teens with extrovert and introvert personality types | Hypothesis test results show the number 0.01. This shows that there are differences in communication privacy management between extravert and introvert personality types. The results showed that the extravert type revealed more of its privacy on social media compared to the introvert type. | In further research, data retrieval should not only be done with a questionnaire instrument but also can be done with interviews and observations so that the data obtained is more in-depth as supporting data, especially in this study observing individual behavior as users of social media. |
| Ref.[12]  | This study aims to recognize a person's Five-Factors personality based on an analysis of their user activity on Facebook | This study involved 100 participants. By using the boost-decision tree algorithm, it produces an accuracy of 82.2%. | In future studies, it is expected to use participants who come from different countries and different conditions |
| Ref.[13]  | This study aims to introduce lexical resources that can be used for the preprocessing stage in social media profiling research. | The resources come from the dictionary of slang words, contractions, abbreviations, and emoticons commonly used on social media. Each dictionary is made for English, Spanish, Dutch, and Italian. | In further research, we must expand the dictionary which contains non-standard words and is entered manually for each language. In the same way as in Spanish. |
| Ref.[14]  | This study aims to provide alternatives to human resources in obtaining prospective employee personality data through their Twitter account. | Twitter's classification process uses the W-IDF weighting method and the Naive Bayes Classifier algorithm produces a not-so-good accuracy of 36.67% by comparing with the results of psychologists. | Increase the amount of data training words that have been labeled as well as Twitter accounts that have been validated by a psychologist. This is useful for testing the accuracy of the methods used in this study. |
The Objective of The Paper

This study aims to analyze whether profiling on Twitter through text processing tweets can be an alternative for human resources in exploring the personality of prospective employees. This research is divided into four scenarios, namely not-stemmed-not-weighted, stemmed-not-weighted, not-stemmed-weighted, and stemmed-weighted.

Conclusion

The results for each scenario are as follows: not-stemmed-not-weighted gets an accuracy of 37.41%, stemmed-not-weighted gets an accuracy of 30.21%, not-stemmed-weighted gets an accuracy of 35.97%, and stemmed-weighted gets an accuracy of 30.93%.

Suggestion/Weakness

Future studies are suggested using more variables, not just text. Examples are user hashtags, user demographics, emoji usage, user behavior, and user location.

5. Conclusion

This study analyzes profiling through the activities of social media users. Important information about a person can be extracted through social media activities such as Facebook and Twitter. Posts on social media can be used as a data source for analysis so that important information about a person can be obtained. The results of the personality analysis can later be utilized by a company to help recruit and place prospective employees in suitable positions, this is because one of the factors that support the back and forth of a company is the quality of its human resources. Thus, it can be concluded that the personality analysis of social media can be used as one of the references of a company to recruit and recruit new job candidates.

Acknowledgements

Universitas AMIKOM Yogyakarta for the support by giving all the facility which is needed for the study. This research also can be done by good cooperation to provide which hopes can useful for others.

References

[1] E. Shin, *The DISC Codes: Cara Cepat Menguasai Kode Sukses Manusia*. Bor Monda Tour, 2013.
[2] “Digital 2020 - We Are Social.” https://wearesocial.com/digital-2020 (accessed May 17, 2020).
[3] F. Amapola, F. Bosco, G. Cafiero, E. D. Angelo, and Y. S. Unici, “Defining Profiling Working Paper,” *SSRN eLibrary*, pp. 1–40, 2013.
[4] D. Lhaksmana, KM; Nhita, Fhira & Anggraini, “Klasifikasi Keprribadian Berdasarkan Status Facebook Menggunakan Metode Backpropagation,” *e-Proceeding Eng.*, vol. 4, no. 3, pp. 5174–5183, 2017.
[5] M. Z. Sarwani and W. F. Mahmudy, “Analisa Twitter Untuk Mengetahui Karakter,” *Semin. Nas. Sist. Inf. Indones.*, no. November, pp. 2–3, 2015.
[6] N. Hutagalung, “Klasifikasi Tipe Keprribadian Pengguna Sosial Media Berdasarkan Teori Big Five Menggunakan K-Nearest Neighbor,” pp. 14–15, 2018.
[7] Y. I. Claudy, R. S. Perdana, and M. A. Fauzi, “Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN),” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 8, pp. 2761–2765, 2018, [Online]. Available: https://www.researchgate.net/publication/322959490.

[8] N. M. Hikmah, “Hubungan Kepribadian Big Five Dengan Pengungkapan Diri Secara Online Pengguna Media Sosial,” 2017.

[9] V. Ong et al., “Personality prediction based on Twitter information in Bahasa Indonesia,” in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, 2017, doi: 10.15439/2017F359.

[10] A. T. Damanik and Masayu Leylia Khodra, “Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression,” J. Cybermatika, 2015.

[11] H. Gómez-Adorno, I. Markov, G. Sidorov, J.-P. Posadas-Durán, M. A. Sanchez-Perez, and L. Chanona-Hernandez, “Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts,” Comput. Intell. Neurosci., vol. 2016, p. 1638936, 2016, doi: 10.1155/2016/1638936.

[12] A. D. Hartanto, E. Utami, S. Adi, and H. S. Hudnanto, “Job seeker profile classification of twitter data using the naïve bayes classifier algorithm based on the DISC method,” in 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISSEE 2019, 2019, doi: 10.1109/ICITISSEE48480.2019.9003963.

[13] A. Souri, S. Hosseinpour, and A. M. Rahmani, “Personality classification based on profiles of social networks’ users and the five-factor model of personality,” Human-centric Comput. Inf. Sci., vol. 8, no. 1, 2018, doi: 10.1186/s13673-018-0147-4.

[14] X. Liu and T. Zhu, “Deep learning for constructing microblog behavior representation to identify social media user’s personality,” PeerJ Comput. Sci., 2016, doi: 10.7717/peerj-cs.81.