Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 1, February 2024, pp. 69-75

DOI: https://doi.org/10.52436/1.jutif.2024.5.1.1166
p-ISSN: 2723-3863
e-ISSN: 2723-3871

PERSONALITY DETECTION ON TWITTER USER USING XGBOOST ALGORITHM

Adinda Putri Rosyadi[*1], Warih Maharani[2], Prati Hutari Gani[3]

[1,2,3]Informatics, Faculty of Informatics, Universitas Telkom, Indonesia
Email: [1]adindarosyadi@student.telkomuniversity.ac.id, [2]wmaharani@telkomuniversity.ac.id ,
[3]pratihutarigani@telkomuniversity.ac.id

(Article received: June 28, 2023; Revision: July 23, 2023; published: January 31, 2024)

Abstract

Personality is a person's identity that is addressed to the public. The Big Five personality is the most commonly used personality model. Detecting a person's personality is still a difficult task today. Because personality detection still often requires humans to fill out lengthy questionnaires to evaluate various personality traits. Therefore, a system that is able to identify personality easily and specifically is needed. By using social media, individuals often express their feelings. Twitter is the most popular social networking platform today. In this research, we use the XGBoost Algorithm, a powerful machine learning method, to create a personality detection system that improves upon existing approaches. Our research aims to determine how well the XGBoost algorithm can recognize Big Five personality features in Twitter users. We achieved encouraging results through in-depth investigation and experimentation. The XGBoost algorithm successfully developed a model that can recognize all Big Five personality trait labels but with different precision, recall and f1-score values. The highest value was obtained for the Extroversion label with a precision of 0.92, recall of 1.00 and f1-score of 0.96. Meanwhile, the lowest value is owned by the Agreeableness label with a precision value of 0.29, recall 0.29, and f1-score of 0.29. This research demonstrates the potential of the XGBoost Algorithm for personality discovery on social media platforms, providing a fast and accurate method to identify distinctive characteristics. Overall, the results of this study demonstrate the efficiency of the XGBoost Algorithm in the context of personality recognition, opening the door for further development in understanding and evaluating human behavior through social media platforms such as Twitter.

Keywords: *Big Five, Personality, Twitter, XGBoost Algorithm.*

1. INTRODUCTION

Personality is one of the factors that influence human behavior and actions. According to experts, personality is an attribute that displays what, why, and how an individual behaves in their environment [1]. Many types of personalities have been discovered, such as the Big Five, MBTI, and DISC. Among these types, the Big Five is the most commonly used model by researchers. It has a genuinely integrative understanding of personality structure across essential personality traits and clinical psychiatry [2]. The Big Five, as described by Costa & McCrae in Utami [3], is a test tool that can enhance trait theory, which describes the model to find key characteristics in describing personality. The Big Five Model is also frequently used in psychological research to evaluate an individual's contribution to an organization. One such study was carried out by Febriyanti, et al. [4], who employed the Big Five Model to gauge the social well-being of their workforce. Furthermore, Celli and Leprib [5] proved that the Big Five offers more insightful data and rules than the MBTI. The Big Five consists of Neuroticism (emotional stability), Extraversion (sociability), Openness (intellectualism), Agreeableness (sensitivity level), and Conscientiousness (self-discipline) [6].

Detecting someone's personality is a complicated process that often involves filling out lengthy questionnaires evaluating many personality traits. However, the results may be inaccurate or inappropriate if respondents need to understand the questions fully or are unwilling to provide honest answers [7]. Therefore, a system is required to detect someone's personality without the need for lengthy questionnaires and with accurate results. In this technological era, detecting someone's personality can be seen through their online interactions. Because social media users have a lot of information available about them, a system is needed to detect individuals through the social media they use [8]. According to Nuo Han et al., social media can reveal a person's personality. Therefore, the authors claim that social media expression can disclose personality traits because social media users frequently express their feelings [9].

One of the social media platforms widely used today is Twitter. Ruby [10] writes that Twitter is the 15th most commonly used social media platform globally, with 450 million active Twitter users as of 2022. This indicates that Twitter is Indonesia's most

widely used social media platform. According to Kemp [11] the number of active Twitter users in Indonesia as of 2022 has reached 18.45 million. A large amount of active user data indicates that the number of tweets generated is also the same. Using tweets, humans can communicate with each other to express their feelings.

Many researchers have researched personality detection in Twitter users. However, the results of these studies may only be considered approximately accurate. For example, Angsaweni's study showed that the accuracy rate for identifying the Big Five personality traits using the AdaBoost method was only 53.57%[12]. In Lydia's study, SVM is used to identify DISC-type personalities, but the model accuracy is just 53% [13]. Pratama [14] conducted a similar study using a different method, Random Forest and achieved an accuracy rate of 69.23%. Given the accuracy rates achieved in these studies, there is a need for a classification model that can provide higher accuracy rates.

Therefore, this study will use the XGBoost algorithm. XGBoost (Extreme Gradient Boosting) is one of the derivative models of the Gradient Boosting Model that has fast and accurate calculations, making it the best-performing model among the Decision Tree Models available. This algorithm was previously used by Nasution and his team [15] to calculate the accuracy rate of diabetes classification, and it was found that XGBoost had the highest accuracy rate (90.10%) compared to the comparative model, Naive Bayes (79.68%). Kurniawanda and Tobing [16] also conducted research using XGBoost to analyze sentiment comments on Instagram, achieving an accuracy rate of 75.20%. Qi [17] proved in his research on text-based classification of theft crimes using the XGBoost algorithm and produced a precision value of 0.96, recall 0.96 and f1-score of 0.96. Therefore, the application of the XGBoost Algorithm in this research is to perform personality identification on Twitter users with good performance results. In addition, the XGBoost Algorithm is capable of predicting the personalities of Twitter users based on the actual personalities of the users.

2. RESEARCH METHODOLOGY

2.1. Research Stages

In this study, a system will be developed to detect personality traits in Twitter users using the XGBoost algorithm. The research data used was obtained from Twitter through crawling. The system design used in this study is illustrated in Figure 1.

2.2. Data Collection

This study used Big Five personality test data from Twitter users who were willing to participate as respondents. The information gathered consists of username-accompanied Indonesian comment data. The Twitter API that has been offered is used to retrieve the data. The outcomes of data crawling are saved in documents with the .csv extension, where personality categorization based on the big five personalities is still being done. The Big Five Personality Theory is largely accepted by the general public and is frequently utilized in recruitment. Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism are the five dimensions of the Big Five model of personality.

2.3. Preprocessing

The obtained data will be preprocessed before the extraction and classification of features. This procedure aims to select pertinent and suitable data for classification. The preprocessing phase includes five steps: Tweets containing URLs, keywords, emoticons, and usernames are cleaned by removing or deleting words. Case folding converts capital letters to lowercase characters. Stopword Removal eliminates prevalent words and only retains data-representing words. Stemming, which looks for the root form of words within the tweet data. Tokenizing, which deconstructs sentences into individual words, makes the words visible.

2.4. Feature Extraction

In this section, feature extraction is carried out by weighting with TF-IDF (Term Frequency-Inverse Document Frequency) and adding the weight values to the emotion features. This phase attempts to transform raw input data into meaningful and condensed representations of features that can be used in the classification phase.

TF – IDF method was used in this research to determine the importance of the weight of each word in the text. IDF measures the informativeness of the term, while TF measures the ratio of the occurrence of a word in a document [12]. The method of calculating TF-IDF is shown in (1).

$$tfidf_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

For the emotion feature will be using EmoLex Dictionary. The EmoLex Dictionary contains terms pertaining to eight fundamental emotions: anger, anticipation, disgust, fear, sadness, joy, surprise, and trust, as well as two polar emotions: positive and negative. EmoLex was used because an evaluation of 2,216 linguists revealed 84.7% agreement on word-emotion associations pertaining to fundamental emotions and 82% agreement on word-emotion associations pertaining to emotion polarity [14]. Words associated with a particular emotion are assigned a value of 1 by EmoLex. Meanwhile, terms that are not associated with an emotion are assigned a value of 0.

2.5. XGBoost Algorithm

One of the current implementations of the best gradient-boosting trees is XGBoost (Extreme Gradient Boosting). The XGBoost research project, released by Tianqi Chen on March 27th, 2014, has become the most popular machine learning algorithm for classification and regression. It is significantly faster than previous ensemble tree-based algorithms due to its parallel tree-boosting feature, designed for efficiency and scalability. XGBoost gained notoriety for its remarkable accuracy after dominating several machine learning competitions [18].

It is important to employ a ready-to-use XGBoost [19] library for classification known as XGBClassifier in order to implement this algorithm in the research aimed at detecting personality traits in Twitter users. XGBoost also includes the learning_rate (model update employs step size reduction), max_depth (depthiest point of a tree), and n_estimators (the quantity of enhanced trees) parameters, which can be used and configured according to your requirements [20]. After calling the library, input the dataset that has been divided into multiple portions, such as 90:10, 80:20, 70:30, 60:40, and 50:50. The goal of evaluating the XGBoost algorithm with various split shapes is to identify the optimal classification outcome.

To use the XGBoost algorithm, numeric data input is required because XGBoost cannot process textual information data. Therefore, if the available dataset is in text form, preprocessing and weighting are necessary for the text in the dataset [21]. The TF-IDF (Term Frequency-Inverse Document Frequency) method can be used to perform weighting on text data.

2.6. Grid Search Cross Validation

A machine learning model's model selection process is facilitated by optimizing with hyperparameter tuning utilizing Grid Search. Grid Search makes it simple to verify each model parameter without having to perform manual validation one at a time, together with Cross Validation. It produces precise and ideal prediction outcomes when used in conjunction with understanding and intuition [22].

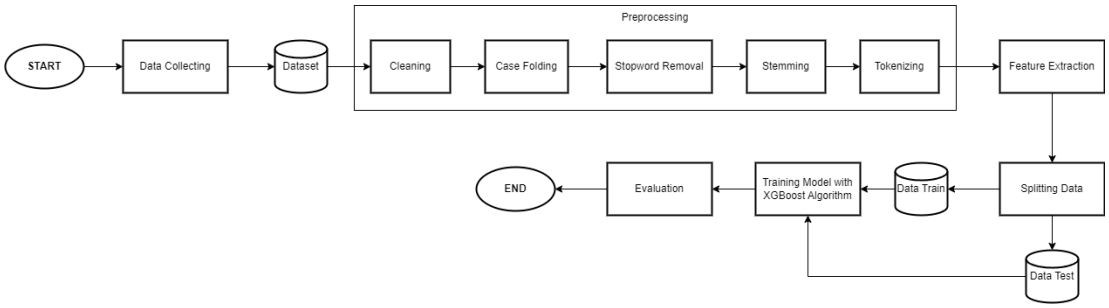

Figure 1. System Flowchart

2.7. SMOTETomek Hybrid Sampling

SMOTETomek is one of the Hybrid Sampling methods used to balance the dataset. SMOTETomek is a combination of SMOTE (Synthetic Minority Over-sampling Technique) and Tomek's under sampling link technique. SMOTE is an oversampling technique that generates new synthetic minority class samples, whereas Tomek Links is an under-sampling technique that eliminates some of the majority class samples that are adjacent to the minority class samples. The combination of these two techniques seeks to improve the efficacy of handling class data with imbalances [23].

2.8. Evaluation

In this section, the evaluation process will involve the use of Confusion Matrix to classify the classification results. The Confusion Matrix is a two-dimensional performance measurement and classification matrix. The matrix consists of two identifiers, positive and negative, and four distinct permutations. These are True Positive, True Negative, False Positive, and False Negative combinations.

Table 1 illustrates the shape of the Confusion Matrix.

Table 1. Confusion Matrix

Actual Class	**Predicted Result**	
	True	False
Positive	TP	FP
Negative	TN	FN

Furthermore, the results of the confusion matrix can be utilized to calculate the classification model's accuracy, precision, recall, and f1-score. This evaluation will determine how effective the XGBoost algorithm is at detecting personality traits among Twitter users. Accuracy (2) can be described as the proportion of right predictions relative to the total number of forecasts. On the other hand, the evaluation metrics won't be able to represent the classifier's effectiveness if the data are uneven and skewed. Because of this, additional measures including precision (3) and recall (4) are required in order to calculate the evaluation. If the recall and precision of the numbers are somewhat close to one, then the categorization will be more accurate. In order

to take into account both precision and recall, the F1-Score value is necessary. The F1-score (5) is a representation of the harmonic recall and precision on average. The best possible F1-Score is a 1, and the worst possible score is 0 [24].

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2)$$

$$precision = \frac{TP}{(TP + FP)} \quad (3)$$

$$recall = \frac{TP}{(TP + FN)} \quad (4)$$

$$f1 - score = \frac{2 \times (precision \times recall)}{(recall + Tprecision)} \quad (5)$$

3. RESULT

This section contains the results of this research, which can be made especially the application of the method used, either simply by presenting the existing data in the study. This section will elucidate the findings of the study utilizing the XGBoost algorithm to detect Twitter users' personalities. This study's data consists of 260 users, with each user's tweet data collected to a maximum of around 1,000 tweets. In addition, the data is supplemented with personality labels that have been calculated and adjusted according to the Big Five Personality characteristics.

Table 2 illustrates the data distribution for this research. According to the table, Openness has the most data, whereas Extraversion has the least. The data distribution is observed to be imbalanced, which will impact the model's accuracy. Therefore, this study will concentrate on the precision, recall, and F1-Score values of the XGBoost algorithm.

Table 2. Data Distribution

Label Personality	Amount Data
Openness	107
Agreeableness	78
Neuroticsm	48
Conscientiousness	23
Extroversion	6

Afterwards, the collected data will undergo preprocessing to filter and extract the data that is pertinent to the classification needs. Table 3 displays the results of the study's data preprocessing.

Table 3. Process of Preprocessing

Process	Result
Actual Data	2021-03-04 12:31:09,1367452552321585155,"Akhirnya berhasil juga doski potong rambutnya sendiri ðŸ˜‚ðŸ˜
Cleaning	Akhirnya berhasil juga doski potong rambutnya sendiri
Case Folding	akhirnya berhasil juga doski potong rambutnya sendiri
Stopword Removal	akhirnya berhasil doski potong rambutnya
Stemming	akhir berhasil doski potong rambut
Tokenizing	[‘akhir’, ‘berhasil’, ‘doski’, ‘potong’, ‘rambut’]

The next step is feature extraction using TF-IDF and Emotion Feature. In this study, word weighting with TF-IDF and an emotion feature are used to evaluate the value of words in user data that may contain specific emotions associated with the user's personality. In this study, the evaluation scenario will be implemented twice. In the first test scenario, TF - IDF and emotion features will be used as extraction features. Then, for the second test scenario, two new features will be added to the first test scenario, namely sentiment features and social features. Using the XGBoost Algorithm, these two scenarios will be juxtaposed to determine which model has superior performance for classifying Twitter users' personalities.

3.1. First Test Scenario

In the first test scenario, the use of datasets that had features extracted from them using TF-IDF and emotion features were put into the XGBoost Algorithm model in the form of split data in the ratios 90:10, 80:20, 70:30, 60:40, and 50:50 respectively. This test also makes use of GridSearchCV to determine which hyperparameters would be most beneficial to incorporate into the XGBoost Algorithm model. The findings of the tests indicate that, in comparison to the other split forms, the accuracy of the 90:10 data split form that makes use of GridSearchCV is the highest, coming in at 38.46% with parameter learning_rate = 0.2, max_depth = 5, and n_estimators = 400. **Error! Reference source not found.** depicts the results of the confusion matrix for the 90:10 split data Table 5 displays the precision, recall, and F1-score values for the results of the test scenario.

Table 4. Confusion Matrix of First Test Scenario

Predict Label / Actual Label	Agr	Cons	Neu	Open
Agr	2	1	2	2
Cons	2	0	0	0
Ext	0	0	1	0
Neu	1	1	2	1
Open	2	2	1	6

Table 5. Evaluation Result of First Test Scenario

Label Personality	Precision	Recall	F1-Score
Agreeableness	0.29	0.29	0.29
Conscientiousness	**0.00**	0.00	0.00
Extroversion	**0.00**	0.00	0.00
Neuroticsm	0.33	0.40	0.36
Openness	**0.67**	0.55	0.60

3.2. Second Test Scenario

Following the completion of the first test scenario, the imbalanced data distribution led to the production of less-than-ideal outcomes. Therefore, the data balancing process in the second situation is conducted by utilizing the Hybrid Sampling approach

with SMOTETomek. The results of data distribution after balancing are shown in

Table 6.

Table 6. Distribution Data after Balancing

Label Personality	Amount Data
Openness	98
Agreeableness	100
Neuroticsm	104
Conscientiousness	103
Extroversion	106

The difference between the quantity of data presented in

Table 2 and

Table 6 is significant. It can be seen that the Extroversion label oversamples with 103 data out of an initial 6 data points. While Openness experiences under sampling with a large amount of data, 98 out of the initial 107 data points. So that the optimal results are obtained with a 90:10 split data form using GridSearchCV with an accuracy value of 73.08%.

Table 7 depicts the outcomes of the data balancing process with SMOTETomek for a 90:10 data shape supported by GridSearchCV and hyperparameter values of XGBoost learning_rate = 0.3, max_depth = 7, and n_estimators = 100 and Table 8 displays the values for precision, recall, and f1 score.

Table 7. Confusion Matrix of Second Test Scenario

Predict Label / Actual Label	Agr	Cons	Ext	Neu	Open
Agr	7	0	0	0	3
Cons	0	8	0	1	1
Ext	0	0	11	0	0
Neu	0	0	0	8	3
Open	3	2	1	0	4

Table 8. Evaluation Result of Second Test Scenario

Label Personality	Precision	Recall	F1-Score
Agreeableness	0.70	0.70	0.70
Conscientiousness	0.80	0.80	0.80
Extroversion	**0.92**	1.00	0.96
Neuroticsm	0.89	0.73	0.80
Openness	**0.36**	0.40	0.38

3.3. Third Test Scenario

In this final scenario test experiment, we will compare three different types of parameter values to determine which types of parameters can provide the XGBoost Algorithm with optimal values. The results of the first and second scenario test experiments will be used to determine Parameter 1 and Parameter 2. As for Parameter 3, its value will be determined by the parameter attribute's optimal value. This experiment compares learning_rate, max_depth, and n_estimators, the three XGBoost parameters. And in this experiment, we will use balanced data and 90:10 split form. Based on the outcomes of this experiment, Parameter 3 with learning_rate = 4, max_depth = 8, and n_estimators = 400 produces the best results. The experiment for this scenario is summarized in

Table 9 and the precision, recall, F1-Score of the optimal value as shown in Table 10.

Table 9. Experiment Result of Third Test Scenario

Label Personality	Accuracy
Param 1: learning_rate = 0.2, max_depth = 5, n_estimators = 200	71.15%
Param 2: learning_rate = 0.3, max_depth = 7, n_estimators = 100	73.08%
Param 3: learning_rate = 0.4, max_depth = 8, n_estimators = 400	**75.00%**

Table 10. Evaluation Result of Third Test Scenario

Label Personality	Precision	Recall	F1-Score
Agreeableness	0.64	0.70	0.67
Conscientiousness	**0.90**	0.90	0.90
Extroversion	0.85	1.00	0.92
Neuroticsm	0.78	0.64	0.70
Openness	**0.56**	0.50	0.53

4. DISCUSSION

Using the Emotion feature and GridSearchCV to identify hyperparameters for the XGBoost Algorithm, the model can produce optimal results, based on the results of three scenario tests. As observed in the first test scenario involving the 90:10 split form, the Openness Label has a greater precision value than the other four labels. While the Conscientiousness and Extraversion labels yield a precision value of 0, indicating that the model is unable to predict these two labels. This is influenced by the limited quantity of data as well as the unbalanced distribution of data. The quantity of data associated with the Openness label is bigger, resulting in an enormous amount of data that enables the model to make accurate predictions. Compared to the Extroversion and Conscientiousness labels, the quantity of data is small, so the model receives limited information and it is challenging to predict these two labels.

Concerning the second test scenario, balanced data proves to be a factor that enables the model to generate optimal predictions. It is presented in Table 8 That the Extroversion label, which experiences oversampling, generates more data than the Openness label, which experiences under sampling and therefore generates less data than the other four labels, is the label with the highest precision at this time. The high precision value for the Extroversion label indicates that the model receives a great deal of information about this label in order to make accurate predictions. Unlike the Openness label, which has limited data, the model has less information to correctly predict this label.

For the final scenario, the experiment was conducted by comparing and adjusting the parameters based on the outcomes of the preceding two scenarios. Using balanced data and a 90:10 division, it was discovered that the third parameter value produced the best outcomes. The quantities of the third parameter's learning_rate, max_depth, and

n_estimators are greater than those of the previous two scenarios' learning_rate, max_depth, and n_estimators. Moreover, as can be seen in Table 10, as is the case. The precision value for the Extroversion label has decreased marginally, while the Openness label's value has increased. Therefore, if the model employs a high parameter value, its value will be more optimal.

From the results of the three scenario tests, it is possible to create an effective model to identify personality by choosing appropriate parameters and a balanced distribution of data.

Furthermore, it is evident from a comparison of the outcomes of the three test scenarios that the first test scenario offers significantly superior results than the research done by Angsaweni [12]. Only 28.57% accuracy was achieved by Angsaweni utilizing the AdaBoost technique and three extraction features, compared to 38.46% accuracy for the first test scenario. Additionally, Angsaweni's research did not perform experiments by balancing data distribution and determining the optimal parameters, which is still a weakness in investigating studies in his research.

By analyzing the results of the test scenarios that were run, particularly on the Openness, Extroversion, and Conscientiousness labels based on the emotional value of each Twitter user, it can be proven in this study that the XGBoost Algorithm can be used to detect the Big Five personalities. The results show that the XGBoost Algorithm generates accurate predictions due to its capacity to handle complicated data and the speed through which huge Twitter datasets were processed.

When dealing with conditions of an uneven data distribution, the XGBoost Algorithm has some limitations. The model's ability to detect personality labels might be impaired when some personality labels have less data samples. In addition, the quality of emotional data also affects the performance of the algorithm, where incomplete or unstructured emotional data can reduce prediction accuracy.

Therefore, to ensure the XGBoost Algorithm performs at its best in predicting the Big Five personalities in Twitter Users, extra attention must be paid to imbalanced data distribution as well as to strong emotional data quality.

5. CONCLUSION

Based on the findings of this study's first, second, and third scenario tests, it can be concluded that the XGBoost Algorithm can be used to build a model capable of detecting the personality of Twitter users, specifically by using a 90:10 split form, adjusting XGBoost parameters based on the needs of the model, and distributing data evenly. Because these factors contribute to the development of the optimal detection model. There are also suggestions for future research, such as using balanced distribution data and removing the stemming process through preprocessing. Because the true meaning of words can be altered by stemming, which also leads to an inaccurate prediction of the information content of words.

REFERENCES

[1] N. Fatwikiningsih, *Teori Psikologi Kepribadian Manusia*. Yogyakarta: CV. ANDI OFFSET, 2020. [Online]. Available: https://books.google.co.id/books?hl=en&lr=&id=UCn-DwAAQBAJ&oi=fnd&pg=PP1&dq=kepribadian&ots=Xu3giO5q_z&sig=QibB7ZyqLRO4E226VLKxPGwjGVQ&redir_esc=y#v=onepage&q=kepribadian&f=false

[2] T. A. Widiger and C. Crego, "The Five Factor Model of personality structure: an update," *World Psychiatry*, vol. 18, no. 3, pp. 271–272, 2019, doi: 10.1002/wps.20658.

[3] S. A. Utami, N. Grasiaswaty, and S. Z. Akmal, "Hubungan Tipe Kepribadian Berdasarkan Big Five Theory Personality dengan Kebimbangan Karier pada Siswa SMA Relationship between Types of Personality Based on Big Five Theory Personality with Career Indecision among High School Students," vol. 6, no. 1, 2018.

[4] V. Febriyanti, N. Eva, and S. Andayani, "Tingkat kesejahteraan psikologis ditinjau dari tipe kepribadian big five psychological well-being level based on big five personality type," *Psycho Idea*, vol. 20, no. 2, pp. 141–152, 2022.

[5] F. Celli and B. Lepri, "Is Big Five better than MBTI?," *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, vol. 2018, no. December 2018, pp. 93–98, 2019, doi: 10.4000/books.aaccademia.3147.

[6] G. Alderotti, C. Rapallini, and S. Traverso, "The Big Five personality traits and earnings: A meta-analysis," *J Econ Psychol*, no. October, p. 102570, 2022, doi: 10.1016/j.joep.2022.102570.

[7] S. Berkovsky *et al.*, "Detecting personality traits using eye-tracking data," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12, 2019, doi: 10.1145/3290605.3300451.

[8] S. V. Therik, E. B. Setiawan, and U. Telkom, "Deteksi Kepribadian Big Five Pengguna Twitter," *eProceedings of Engineering*, vol. 8, no. 5, pp. 10277–10287, 2021.

[9] N. Han *et al.*, "How social media expression can reveal personality," *Front Psychiatry*, vol. 14, no. March, pp. 1–12, Mar. 2023, doi: 10.3389/fpsyt.2023.1052844.

[10] D. Ruby, "62 Twitter Statistics In 2023 — (Users, Revenue & Trends)," *Demand Sage*,

2023. https://www.demandsage.com/twitter-statistics/

[11] S. Kemp, "DIGITAL 2022: INDONESIA," *Data Reportal*, 2022. https://datareportal.com/reports/digital-2022-indonesia

[12] A. Angsaweni and W. Maharani, "Identification of Big Five Personality on Twitter Users using the AdaBoost Method," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, pp. 377–383, 2022, doi: 10.47065/bits.v4i2.1853.

[13] K. F. Lydia and E. B. Setiawan, "Sistem Prediksi Kepribadian DISC Pengguna Twitter Dengan Algoritma Support Vector Machine (SVM) Menggunakan Metode Pembobotan TF-IDF Dan ANP," 2019.

[14] R. P. Pratama and W. Maharani, "Predicting Big Five Personality Traits Based on Twitter User Using Random Forest Method," in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021, pp. 110–117.

[15] M. K. Nasution, Rd. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Naïve Bayes Dan Algoritma Xgboost Pada Klasifikasi Penyakit Diabetes," *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 9765–9772, 2021, [Online]. Available: https://journal.ubpkarawang.ac.id/mahasiswa/index.php/ssj/article/view/424/338%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15759

[16] M. R. Kurniawanda and F. A. T. Tobing, "Analysis Sentiment Cyberbullying In Instagram Comments with XGBoost Method," *IJNMT (International Journal of New Media Technology)*, vol. 9, no. 1, pp. 28–34, 2022, doi: 10.31937/ijnmt.v9i1.2670.

[17] Z. Qi, "The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model," *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, pp. 1241–1246, 2020, doi: 10.1109/ICAICA50127.2020.9182555.

[18] B. Quinto, *Next-Generation Machine Learning with Spark*. 2020. doi: 10.1007/978-1-4842-5669-5.

[19] Xgb. Developer, "Release 1.7.3 xgboost developers." 2023. [Online]. Available: https://xgboost.readthedocs.io/en/stable/get_started.html

[20] Y. Wang and X. S. Ni, "A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization," *International Journal of Database Management Systems (IJDMS)*, vol. 11, no. 1, pp. 243–250, Jan. 2019, doi: https://doi.org/10.48550/.

[21] H. Jahanshahi *et al.*, "Text Classification for Predicting Multi-level Product Categories," in *31st Annual International Conference on Computer Science and Software Engineering*, 2021, pp. 33–42. [Online]. Available: http://arxiv.org/abs/2109.01084

[22] W. Nugraha and A. Sasongko, "Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search," *SISTEMASI : Jurnal Sistem Informasi*, vol. 11, no. 2, pp. 391–401, 2022.

[23] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.

[24] G. Shobha and S. Rangaswamy, *Chapter 8 Machine Learning*, 1st ed., vol. 38. Elsevier B.V., 2018. doi: 10.1016/bs.host.2018.07.004.