Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 5 (2022) 818 - 823 ISSN Media Electronic: 2580-0760

Big Five Personality Assessment Using KNN method with RoBERTA

Athirah Rifdha Aryani[1], Erwin Budi Setiawan[2]
[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]athirahrifdha@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstract

Personality is the general way a person responds to and interacts with others. Personality is also often defined as the quality that distinguishes individuals. Social media was created to help people communicate remotely and easily. These personalities fall into five categories known as the Big Five personality traits, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). The use of K-Nearest Neighbour (KNN) is a method of classifying objects based on the training data closest to them. To overcome the data imbalance during training data, we use K-Means SMOTE (Synthetic Minority Oversampling Technique). Other features such as LIWC (Linguistic Inquiry Word Count), Information Gain, Robustly Optimized BERT Approach (RoBERTa), and hyperparameter tuning can improve the performance of the systems we build. The focus of this study is to present an analysis of Twitter user behavior that can be used to predict the personality of the Big Five Personality using the KNN method. The Important aspect to consider when using this method, namely accuracy in classifying the Big Five Personalities. The experimental results show that the accuracy of the KNN method is 72.09%, which is 95.28% gain above the specified baseline.

Keywords: *Big Five Personality, K-Nearest Neighbours (KNN), RoBERTa, LIWC, Information Gain*

1. Introduction

In this modern era, almost every society has a social network as a means of communication and expressing each user's personal views on different aspects of life. Twitter is a social media that is widely used by several countries to express feelings and activities written in one or two sentences[1]. Internet-based media allow users to interact and express themselves, directly or indirectly, with large audiences or with user-generated content and interactions with others (Caleb T. Carr dan Rebecca A. Hayes, 2015)[2]. Language-based predictions are made by analyzing word choice and word position in a defined category based on the language used. Language analysis was carried out on several social media profiles, the use of everyday language, and short messages[3].

Personality is understood as an individual's state of mind that depends on behavior, emotions, and attitudes such as the differences in the characteristics of each person. The Big Five personality traits are considered an effective way to determine a person's personality because they are more informative.[4]. The Big Five personality traits are often abbreviated as the "OCEAN" model, openness, conscientiousness, extroversion, agreeableness, and neuroticism[1].

Several studies try to measure a person's personality through the Twitter user word classification method. One of them is research conducted[4], The author tries to use the LIWC method to count words automatically based on the category, then use the Support Vector Machine (SVM) method to classify the Big Five Personalities. This study produces a model that can predict a person's personality by 80.07%. The authors show that research can improve the performance of personality prediction systems by collecting more data from respondents and experimenting with different methods such as combining BERT with deep learning to improve the performance of personality prediction systems.

In other research [3] the prediction of the Big Five personality using TF-IDF and with the K-Nearest Neighbour (KNN) method, it can be concluded that the higher accuracy of the components of social behavior and language and by measuring performance in testing the k value = 9 of 60.97%, while the social and linguistic behavior component with a performance of k=1 has a low value with a value of 39.02%.

Other Research [5] conducted research using a semantic approach of the type RoBERTa (Robustly Optimized BERT Approach) Obtaining an accuracy value of

Accepted: 04-08-2022 | Received in revised: 13-09-2022 | Published: 31-10-2022

83.2%, the highest value of 81.3%, and the median value of 86.5%.

In this study, the aim is to conduct a test to find a way with a complete formula to obtain a model that can improve the classification accuracy and personality prediction of the Big Five using K-Nearest Neighbours. To improve the accuracy achieved in K-Nearest Neighbours by combining LIWC, Information Gain, RoBERTa, K-Means SMOTE, and hyperparameter tuning Tested.

We build a personality detection to predict a person's 5 big personality traits using the K-Nearest Neighbours (KNN) method. The K-Nearest Neighbours method classifies the subjects based on the training data closest to the subject to be used as the Big Five Personality classifier. K-Means SMOTE (Synthetic Minority Synthetic Engineering) is a predictive model for dealing with imbalanced data, Information Gain is used as a feature selection method with the highest feature rating is the most relevant feature and has closely related to the linked dataset, Robustly BERT Approach (RoBERTA) as semantic approach and Linguistic Inquiry Word Count as linguistic feature word counts can improve performance based on correlations between speech and psychologically relevant text. Hyperparameter settings will be added to find the best setting for KNN and help improve accuracy.

This research is structured as follows. Section 2 describes how Twitter's personality prediction system was investigated. Section 3 presents the results of the experiment and discussion and Section present conclusions.

2. Research Methods

The system consists of labelling, data crawling, pre-processing, implementation of feature extraction (LIWC and information retrieval), classification by K-Nearest Neighbours, hyperparameter tuning, and performance evaluation. Figure 1 shows a system that predicts a Twitter user's Big Five personality.

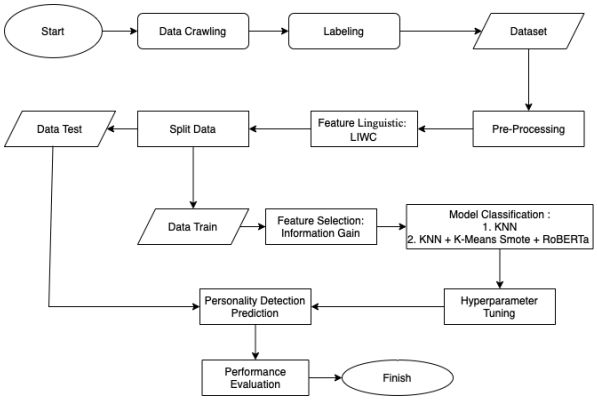

Figure 1. Personality Prediction System

2.1 Big Five Personality

Personality is one of the dominant characteristics or behaviors of a person. The basic personality traits of individuals are interconnected with the main basic traits, such as Extraversion, Conscientiousness, Agreeableness, Neuroticism, and Openness to Experience. Our personality can be how people make decisions, interact with others, and the way they think[6].

Openness personality traits are people who have high curiosity, active imagination, attention, and caring. Conscientiousness personality traits are people who have the character of being careful, careful, wide open, and thorough. Extraversion personality traits are friendly, easy-going, talkative, cheerful, passionate, and enthusiastic. Agreeableness personality traits have traits such as high sympathy, caring, and most like to work in a team. The personality trait of neuroticism is someone anxious, nervous, indecisive, and frustrated. Linguistic aspects tend to have significant personalities [7].

2.2 Crawling

Data crawling is a method of collecting and downloading data from websites[8]. Crawled data for Twitter users are usernames, tweets, social characteristics, followers, followers, tweets, URLs, media URLs, hashtags, retweets, mentions, and capitalizations, which can be seen in Table 1.

Table 1. Description of Social Feature Data[4]

Social Feature	Descriptions
Number of Followers	The number of followers that the user has
Number of Following	The number of users following
Number of Tweets	The number of users tweets
Number of URL	The huge number of URLs shared by users
Number of Media URL	The huge number of media URLs shared by users
Number of Hashtags	The wide variety of user hashtags
Number of Retweets	The number of user retweets
Number of Mentions	The number of user mentions
Number of Capital Letters	The number of Capital letters used by Twitter users

2.3 Labelling

Labelling is based on the results of the Big Five Test Web survey[9]. The questionnaire consists of 120 questions, which are shown on five scales namely, very inaccurate, moderately inaccurate, neither correct nor inaccurate, moderately accurate, and very accurate.

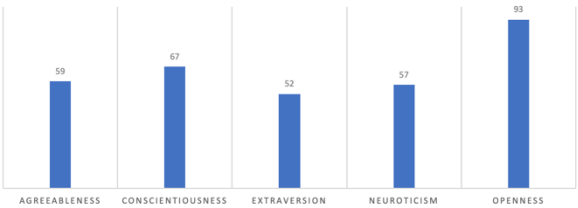

Figure 2. Big Five Personality Distribution on Twitter Users

DOI: https://doi.org/10.29207/resti.v6i5.4394

In this study, 328 user data was collected during the crawl. Figure 2 shows the distribution of personality labelling data with 59 users agreeable, 67 users conscientious, 52 users extraversion, 57 users neurotic, and 93 users openness.

2.4 K-Nearest Neighbour (KNN)

K-Nearest Neighbour (KNN) is an algorithm that classifies it based on the closest distance[3], [10]. K-Nearest Neighbour classifiers are usually based on the Euclidean distance between the test sample and the specified training sample[11]. The nearest neighbour k is specified below,

$$D(p,q) = \sqrt{\sum_{i}^{n}(q_i - p_i)} \quad (1)$$

In formula 1, q_i is data the attributes of which have been normalized and P_i is the new test data on top of the training data[10].

2.5 Robustly Optimized BERT Approach (RoBERTa)

RoBERTa relies on the BERT language masking strategy, to predict the part of the text that is intentionally hidden in a sample of unclassified Languages. RoBERTa gets a significant performance increase by using the BERT architecture and training process[12].

RoBERTa aims to improve the BERT training model. RoBERTa is comparable to the performance of all the previous BERT methods[5].

2.6 Information Gain (IG)

Information Gain (IG) is an associate entropy-based practicality analysis methodology that is widely used in machine learning. As a result of data, the acquisition is used in feature selection, it's defined because of the amount of data provided by the feature for the text category. The collection of data is calculated in step with the number of terms that will be accustomed to classify the knowledge, to measure the importance of lexical parts for classification[13]. The information gain equation is presented below,

$$IG(c,t) = S(c) + \sum_{j \in Value(t)}^{n} \frac{|C_j|}{|C|} S(C_j) \quad (2)$$

In Formula 2, C is a collection of documents, in which there is no functionality t. S(c) is the entropy of all functions c (before splitting), S (c_j) is the entropy function c for class t = j (post-splitting), value (t) is the set of possible values for class t, n is the number of potential values for class t, it is. Most helpful for classification for C. $|c_j|$ is the number of sample classes the value of which = j, |c| is the number of samples for each class. The case of S(c) is formulated as follows,

$$S(c) = -\sum_{i=1}^{m} P(c_i) \log P(c_i) \quad (3)$$

In Formula 3, $p(c_i)$ is the probability for the i characteristic and m is the maximum amount of characteristics.

2.7 K-Means SMOTE

Another technique within the class of techniques accentuation sure category regions uses k-means to cluster the minority class before applying SMOTE among the found clusters. The expressed goal of this technique is to boost category regions by making samples among present clusters of the minority category[14].

The K-Means SMOTE sampling method is used to balance the positive and negative class cases. As a comparison, the initial sampling methods, SMOTE and K-Means SMOTE were applied in the pre-processing step[15].

2.8 Linguistic Inquiry Word Count (LIWC)

Linguistic Inquiry Word Count (LIWC) is an associate in the method for investigating words in line with their classes Pennebaker has been within the development of LIWC since 2007. LIWC has two characteristics, which are open and closed vocabulary. The closed vocabulary performance is ready to research the correlation between language and psychological variables. Table 2 shows the correlation scores between the LIWC class and Big Five Personality developed in previous analysis. The closed vocabulary performance is defined by the collection classes of words using LIWC, which has a significant correlation value. The vocabulary is collected on the official website of LIWC by translating the vocabulary into a proper Indonesian language[4].

TABLE 2. . LIWC Correlation Scores [4]

LIWC Category	O	C	E	A	N
1st person	-0.19	0.02	0.03	0.08	0.10
2nd person	-0.16	0	0.16	0.08	-0.15
3rd person	-0.06	-0.08	0.04	0.08	0.02
plural	-0.10	0.03	0.11	0.18	-0.07
Pronouns	-0.21	-0.02	0.06	0.11	0.06
Negations	-0.13	-0.17	-0.05	-0.03	0.11
Assent	-0.11	-0.09	0.07	0.02	0.05
Prepositions	0.17	0.06	-0.04	0.07	-0.04
Numbers	0.08	0.04	-0.12	0.11	-0.07
Affect	-0.12	-0.06	0.09	0.06	-0.12
Positive Emotion	-0.11	-0.02	0.11	0.14	0.01
Negative Emotion	0	-0.18	0.04	-0.15	0.16
Anxiety	-0.2	-0.05	-0.03	-0.03	0.17
Anger	0.3	-0.19	0.03	-0.23	0.13
Sadness	-0.3	-0.11	0.02	0.01	0.10
Discrepancy	-0.12	-0.13	-0.07	-0.04	0.13
Tentative	-0.06	-0.10	-0.11	-0.07	-0.12
Certainty	-0.06	-0.10	0.10	0.05	0.13
Seeing	-0.04	-0.01	-0.03	0.09	-0.01
Hearing	-0.08	-0.12	0.12	0.01	0.02
Feeling	-0.01	-0.05	0.06	0.10	0.10
Communication	-0.06	-0.07	0.13	0.02	0
Friends	-0.01	0.06	0.15	0.11	-0.08

DOI: https://doi.org/10.29207/resti.v6i5.4394

Family	-0.17	0.05	0.09	0.19	-0.07
Humans	-0.09	-0.12	0.13	0.07	-0.05
Time	-0.22	0.09	0.02	-0.12	0.01
School	0.02	0.04	-0.07	-0.01	0.06
Job/work	0.04	0.07	-0.08	-0.07	0.07
Achievement	-0.05	0.14	-0.09	0.05	0.01
Home	-0.20	0.50	0.03	0.19	0
Sports	-0.14	0	0.05	0.06	-0.01
Tv/movies	0.05	0.06	0.05	-0.05	-0.02
Music	0.04	-0.11	0.13	0.08	-0.02
Money/finance	-0.04	-0.08	-0.04	-0.11	0.04
Metaphysical	0.07	-0.08	0.08	-0.01	-0.01
Death	0.15	-0.12	0.01	-0.13	0.03
Religion	0.05	-0.04	0.11	0.06	-0.03
Sexuality	0	-0.06	0.17	0.08	0.03
Eating/drinking	-0.15	-0.04	0.18	0.03	-0.01
Sleep	-0.14	-0.03	0.02	0.11	0.10
Grooming	-0.20	-0.05	-0.01	0.07	0.05
Swear words	0.06	-0.14	0.06	-0.21	0.11

2.9 Hyperparameter

The traditional way to perform hyperparameter optimization has been grid search. It only exhaustively explores a specific subset of the hyperparameter space Λ of learning algorithm A. The trellis search algorithm typically needs to be guided by performance metrics measured by mutual validation of training sets[16].

2.10 Evaluation score

In this last process, this score consists of accuracy, F1 score, precision, and recall. Accuracy is used to measure the ratio of false positives to the accuracy value and used to measure the overall correction ratio of the model. The F1 score is a combination of precision and recall for an overall measure of model accuracy. Precision is used to measure the ratio of false positives to the accuracy value. A recall is used to measure the ratio of true positives to precision values[17].

Performance evaluation was performed by an associate evaluation metric employing a confusion matrix. True positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) area unit the four classes of confusion matrices[18].

3. Result and Discussion

This section describes the accuracy values for each scenario. This investigation includes his four experimental scenarios. In the first scenario, we tested and compared using K-Nearest Neighbours and K-Means SMOTE. Second scenario by adding LIWC functions. The third scenario is a comparison of Baseline, LIWC, Information Gain, and RoBERTa. The final scenario was baseline data, LIWC, Information Gain, RoBERTa, and hyperparameter comparison. The purpose of the above scenario is to find a comparison result.

3.1 Results

Before starting the experiment, run some tests to split the data into baseline K-nearest neighbours consisting of 90:10, 80:20, 70:30, 60:40, and 50:50 found the optimal ratio for The K-Nearest Neighbours classification and prediction method used for each scenario with the results of the tests shown in Table 3. The method for each scenario was run five times to ensure consistent program results. As a result, we found that the ratio of 80:20 scored the highest, with an accuracy of 37.29%.

Table 3. The ration Comparison

Ratio	Accuracy (%)
90:10	36,67
80:20	**37,29**
70:30	33,71
60:40	34,75
50:50	33,11

The first scenario uses the K-Nearest Neighbours (KNN) method as a benchmark and compares it with the combination of the K-Nearest Neighbours (KNN) with the K-Means SMOTE method. Comparisons are made to optimize the model and are used to process data about imbalances.

Table 4. Accuracy Result from First Scenario

Condition	Accuracy (%)
K-Nearest Neighbours	37,29
K-Nearest Neighbours+ K-Means SMOTE	51,36

The second scenario determines the most effective feature to use for the model, therefore we tend to compare social features with social features combined with language features (LIWC). As declared within the previous scenario, KNN with K-Means SMOTE has the same performance, hence in the second scenario, we tend to use KNN with K-Means SMOTE. The second scenario was conducted to find out the most effective feature to implement during this model. This scenario compared social features and the combination of social features with linguistic features (LIWC). The results of the second scenario are shown in Table 5.

Table 5. Accuracy Result from Second Scenario

Condition	Accuracy (%)
K-Nearest Neighbours + LIWC	33,90
K-Nearest Neighbours + K-Means SMOTE + LIWC	53,80

The implementation of K-Nearest Neighbours using the social feature K-Means SMOTE victimization scored 51,36%. On the opposite hand, achieved an accuracy score of 53,80% by implementing the linguistic feature, LIWC, combined with social functions. It showed that the addition of the LIWC feature contains an important impact on the linguistic aspects of the model. It tested that a combination of the LIWC feature with a social feature is better than using a social feature for the personality prediction system.

DOI: https://doi.org/10.29207/resti.v6i5.4394

According to previous research, the best feature to implement is the combination of linguistic features (LIWC) and social features. Therefore, in the third scenario, we implement the combination of linguistic feature (LIWC) and Information Gain as a selection feature. To optimize the accuracy results, we perform hyperparameter tuning using a grid search method. The results of this experiment are shown in Table 6.

Table 6. Accuracy Result from Third Scenario

Condition	Accuracy (%)
Baseline + K-Means SMOTE + LIWC + Information Gain + RoBERTa	60,08
Baseline + K-Means SMOTE + LIWC + Information Gain +RoBERTa + Hyperparameter Tuning	72,82

The final scenario compares K-Nearest Neighbours plus K-Means SMOTE with K-Nearest Neighbour plus K-Means SMOTE and RoBERTa using the best characteristics we identified in the previous scenario.

Table 7. Scenario Result using comparison

Condition	Accuracy (%)
Baseline	37,29
Baseline + K-Means SMOTE	51.36 (+37.73)
Baseline + K-Means SMOTE + LIWC	53.80 (+44.27)
Baseline + K-Means SMOTE + LIWC + Information Gain + RoBERTa	60.08 (+61.12)
Baseline + K-Means SMOTE + LIWC + RoBERTa + Hyperparameter Tuning	72.82 (+95.28)

The final scenario applies hyperparameter tuning to the method and compares all scenario results and improvement values from the baseline to the latest scenario. The LIWC result using K-Nearest Neighbors, K-Means SMOTE, and RoBERTa without hyperparameter tuning was 60.08%. On the other hand, using hyperparameter tuning for the combination method achieved an accuracy of 72.82%. The results showed that hyperparameter tuning can improve the performance of combinatorial methods, as hyperparameters can find the optimal parameters used for the method. A comparison of test results for each scenario is shown in Table 8.

The parameters used in setting the hyperparameter are leaf_size with a value of 20, the metrics using Minkowski, n_neighbours with a value of 1, the value of p is 2, and weights using uniform.

Table 8. Comparison Personality Traits Accuracy Result

Personality Traits	Accuracy (%)
Neuroticism	69.80
Extraversion	81.90
Agreeableness	82.90
Conscientiousness	58.55
Openness	70.90
Average Accuracy	72.82

3.2 Discussion

In this research, we apply the K-Nearest Neighbors method combined with K-Means SMOTE, LIWC, and RoBERTa. This accuracy can be improved if tested using hyperparameter tuning.

When comparing data ratios, we recommend using an 80:20 data ratio. This is because the 80:10 ratio gives the best accuracy compared to the other ratios (37.29%). This data report will be used for further testing. The first scenario will try to test the model with balanced data using K-Means SMOTE. From the test results, obtaining balanced data can improve accuracy. Balance data is resulted by equalizing the data, there will be no minority class, prevailing over the majority class in the model. With the above results, the next test will apply the data set that has been balanced by K-Means SMOTE.

The second scenario is testing using feature extensions, namely LIWC. From the test results, it has been proven that LIWC can improve accuracy. LIWC can group words into a dictionary of categories so that new groups of words can be created. This is what allows the model to predict more accurately because more data explains the label.

The third scenario implements RoBERTa on top of the previous scenario method. The RoBERTa as an add-on to this method has an accuracy of 60.08%. The results showed that adding RoBERTa to the method had a significant impact on personality prediction performance.

The final scenario is a test to test the model with hyperparameter tuning. This test will use K-Means SMOTE, LIWC, Information Gain, and RoBERTa to help improve the accuracy of the K-Nearest Neighbours model. In this study, Information Gain functions as feature selection, and 100 data items are randomly selected. This test shows increased accuracy when hyperparameter tuning is applied. Hyperparameter tuning will search for the best parameters to apply to the model to demonstrate that performing hyperparameter tuning can increase the accuracy of the predictive model. The results of increasing the precision of the experiment can be seen in Figure 4 below.

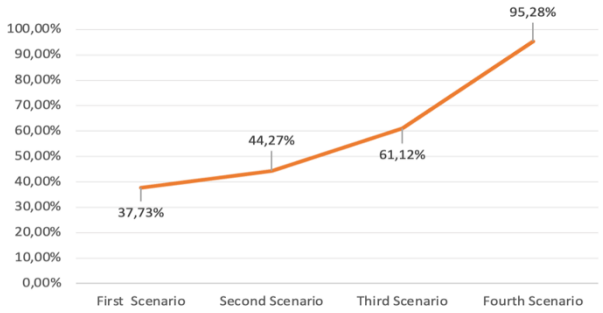

Figure 4. The Improvement of Accuracy Scores.

DOI: https://doi.org/10.29207/resti.v6i5.4394

4. Conclusion

In this study, we combine K-Means SMOTE, LIWC, and RoBERTa as a semantic approach to predict the Big Five personality of Twitter users using the K-nearest neighbor method. System performance results are suitable for datasets of as many as 328 Twitter users and 672,866 tweet data. Implementing a semantic approach is the key to improving system performance.

The data can be said to be imbalanced because there is a dominant label, so data balancing must be done. To overcome data imbalance, we apply the K-Means SMOTE method. In addition, we also include LIWC as a language feature, RoBERTa as a semantic approach, and use hyperparameter tuning to improve model accuracy. As a result of this study, the accuracy value obtained was 72.82% with an increase of 95.28% from Baseline Accuracy.

In the first and second scenarios use the linguistic features of LIWC, and K-Means SMOTE to process imbalanced data, but note that the accuracy of this method is very low. However, in the third scenario, the Roberts model is an improvisation of his BERT model, so the semantic approach seems to have a greater impact on performance. So, the result is stronger, and the system performance is better. Hyperparameter tuning also improves accuracy, as accuracy can be improved by providing optimal parameters to implement in the method. The dataset we collected included 320 Twitter users. With this number of users, the dataset size is small enough to make predictions, which is the limitation of this study.

Reference

[1] N. Febrianto, I. Prasetia, and A. Wijaya, "Pembuatan Sistem Prediksi Kepribadian 'The Big Five Traits' dari Media Sosial Twitter." [Online]. Available: http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_

[2] M. G. Tambunan[1] and E. B. Setiawan, "Prediksi Kepribadian DISC Pada Twitter Menggunakan Metode Decision Tree C4.5 dengan Pembobotan TF-IDF dan TF-RF."

[3] R. Ellandi, E. Budi, S. S. Si, N. Fida, S. Nugraha, and M. P. Psi, "Prediksi kepribadian Big Five dengan Term-Frequency Inverse Document Frequency Menggunakan Metode k-Nearest Neighbor pada Twitter."

[4] G. D. Salsabila and E. B. Setiawan, "Semantic Approach for Big Five Personality Prediction on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 680–687, Aug. 2021, doi: 10.29207/resti.v5i4.3197.

[5] Y. Liu *et al*., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[6] F. Celli and B. Lepri, "Is Big Five better than MBTI? A personality computing challenge using Twitter data." [Online]. Available: https://twitter.com/search-advanced

[7] C. Yuan, J. Wu, H. Li, and L. Wang, *Personality Recognition based on User Generated Content*. IEEE, 2018.

[8] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter," Sep. 2016, pp. 11–16. doi: 10.21108/indosc.2016.111.

[9] "Big Five Personality Test." https://bigfive-test.com/ (accessed Jul. 09, 2022).

[10] B. Yudha Pratama NRP, A. Ec Ir Riyanarto Sarno, and R. A. Nur Esti, "Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM."

[11] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.

[12] D. Faraj and M. Abdullah, "SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pre-trained Model."

[13] S. Lei, "A feature selection method based on information gain and genetic algorithm," in *Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012*, 2012, vol. 2, pp. 355–358. doi: 10.1109/ICCSEE.2012.97.

[14] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," Nov. 2017, doi: 10.1016/j.ins.2018.06.056.

[15] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[16] R. Ghawi and J. Pfeffer, "Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity," *Open Computer Science*, vol. 9, no. 1, pp. 160–180, Jan. 2019, doi: 10.1515/comp-2019-0011.

[17] Willy, E. B. Setiawan, and F. N. Nugraha, "Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter," in *2019 International Conference on Computer, Control, Informatics, and its Applications: Emerging Trends in Big Data and Artificial Intelligence, IC3INA 2019*, Oct. 2019, pp. 114–119. doi:10.1109/IC3INA48034.2019.8949601.

[18] K. Prameswari and E. B. Setiawan, "Analisis Kepribadian Melalui Twitter Menggunakan Metode Logistic Regression dengan Pembobotan TF-IDF dan AHP."

DOI: https://doi.org/10.29207/resti.v6i5.4394