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. 1 (2022) 85 - 91 ISSN Media Electronic: 2580-0760

Optimization Prediction of Big Five Personality in Twitter Users

Gita Safitri[1], Erwin Budi Setiawan[2]
[1,2]Informatics, School of Computing, Telkom University
[1]gitasftr@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id*

Abstract

Various kinds of information can be acquired from social media platforms; one of them is on Twitter. User biographical information and tweets are the essential assets for research that can describe the Big Five Personality, including openness, conscientiousness, extraversion, agreeableness, and neuroticism. Several previous studies have tried the prediction of Big Five Personality. However, the authors found problems in how to optimize the work of the personality prediction system. So, in this study, Big Five Personality predictions were carried out on users of Twitter and improved the performance of the personality prediction system. We implement optimization techniques such as sampling, feature selection, and hyperparameter tuning to enhance the performance. This study also applies linguistic feature extraction, such as LIWC and TF-IDF. By using 287 Twitter users that have permitted their data to be crawled acquired from an online survey using Big Five Inventory (BFI), and applying all optimization techniques, the average accuracy result is 84.22% which is a 74.44% gain over the specified baseline.

Keywords: Big Five Personality, SVM, TF-IDF, LIWC, Optimization

1. Introduction

Social media platforms are becoming a means of self-expression and social communication online. Therefore, social media platforms have become one of the biggest mines for obtaining someone's personal information. Ease of accessing social media allows users to build an online identity, share content (text, links, or images), and interact with others. Twitter is one of the most popular platforms nowadays, where tweets as a medium for writing and sharing posts with others. Then user's behavior and tweets can be easily obtained and analyze the personalities of the Twitter users [1][2][3].

A person's personality can influence life, career, and romance prospects. For companies, knowing a person's personality can be used for employee recruitment, career counseling, and health advice [4]. Several personality models can predict user personality, such as the Mayers-Briggs Type Indicator (MBTI), Big Five Personality, or Dominance Influence Steadiness Conscientiousness (DISC). However, the Big Five Personality is the most popular in psychology and appropriate for characterizing a person's personality [5][6].

Several studies have been attempted to identify and predict a person's Big Five Personality based on user biographical information and texts or tweets posted on social media accounts. Previous research by [5], using the Support Vector Regression method and obtained the smallest MAE value of 0.2739. The results were received by the combination of social behavior features with TF-IDF bigram. A study by [7] implement the SVM method and BERT as the semantic approach with the implementation of LIWC for the personality prediction system and achieved 80.07%. This research also presents that LIWC as their linguistic feature can improve the system performance. In research [8], four classification methods were compared, including Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR). The study also used social behavior features and feature extraction such as LIWC and MRC. The best results using the SVM classification method gain an 88% accuracy score. In comparison, NB 87.5%, LR 62.5%, and RF produces the lowest accuracy, 37.5%.

In the study [9], personality classification was carried out using the SVM and MNB methods as a comparison, and TF-IDF weighting was applied. By using 91 user accounts and 49,919 tweets, the highest accuracy reached 80.5% on SVM and 82% on MNB using scenario 3 with 300 tweets. The dataset used was small, but the labeling process was valid by a psychology expert. The study [10] implement decision tree C4.5 and

Accepted: 28-09-2021 | Received in revised: 18-02-2022 | Published: 27-02-2022

term weighting as linguistic approach. The total size of dataset is 145 Twitter users. The result obtained is 65.72% accuracy score for combination of social behavior features with linguistic approach. The author stated that the accuracy still small because the imbalanced number of class. Another study by [11], with 211 Twitter users data and 474,888 tweets using the Naïve Bayes classification method, implementing the LIWC extraction feature and two weighting methods, namely TF-RF and TF-IDF, resulted in an accuracy of 53.96%. The authors said that the low accuracy was due to the imbalanced data used and caused the model to predict the dominant class.

Based on several studies above, the prediction system on this research was also conducted using the SVM method because this method has been proven to provide exemplary performance in previous studies [8][9]. What distinguishes it from previous research is we use three kernels on SVM: Linear, Radial Basis Function (RBF), and Polynomial with social behavior features data to determine which kernel obtains the best performance then will be used as the baseline. On research [10] the author only uses social behavior features and term weighting as linguistic approach, hence this research will uses LIWC and TF-IDF as linguistic approaches because it's more able to identify the user's personality than only using social behavior features [5]. Due to imbalance data on study [10] and [11], this research will implement oversampling technique as solution. And also to further optimize the personality prediction system will implement other optimization techniques such as feature selection and hyperparameter tuning. The goal of this research is to see how linguistic features (LIWC and TF-IDF), and optimization techniques affect personality prediction on Twitter users that utilize the Big Five Personality model.

The structure of this research is as follows. Section 2 describes the methodology of the personality prediction system. Section 3 presents the results and discussion of the research conducted. Furthermore, in section 4, the conclusions and suggestions are based on our experiment result.

2. Research Methods

The system in this research consists of labeling, data crawling, preprocessing, implementing feature extraction (LIWC and TF-IDF), classification with SVM, hyperparameter tuning, and evaluating the performance. Figure 1 shows the system to predict the Big Five personality in Twitter users.

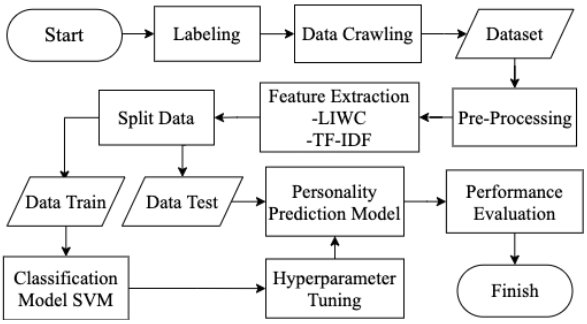

Figure 1 Personality Prediction System

2.1. Big Five Personality

One person's personality is different from another. Personality is one of the characteristics that can be considered to adapt to the environment. Having information about a person's personality can provide clues about how they will react to the current situation [12]. Several personality models can be used to predict the user's personality. However, the Big Five Personality model is the most widely used to describe personality traits [2].

There are five aspects of Big Five Personality: Openness (O), a person who has an active imagination, sensitivity to feelings, like differences, and high inquisitiveness. Conscientiousness (C) is a person who has conscientious, careful, and wide-awake. Extraversion (E) is a person who has an energetic personality, friendly, and conversational. Agreeableness (A) is a person who has a warm, caring, cooperative, and sympathetic personality. Neuroticism (N) is a person who has a personality full of anxiety, jealousy, loneliness, and tends to experience mood swings [13].

2.2. Data Labeling

The labeling is done by the result of the Big Five Inventory (BFI) questionnaire. BFI is used to determine the personality traits of each user. The questionnaire has been developed by previous research [14] that consists of 25 questions in which for each personality traits, there are five questions with a range of 1 to 5.

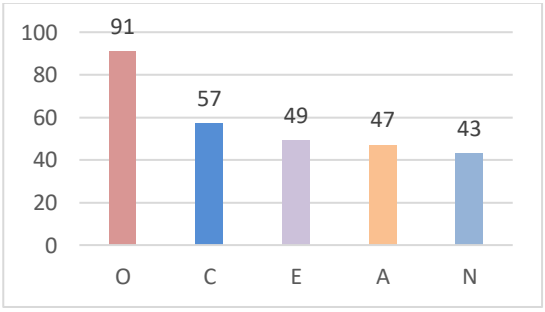

Figure 2 Distribution Data of Personality Users

DOI: https://doi.org/10.29207/resti.v6i1.3529

The questionnaire has been distributed online to Twitter users. Due to the difficulty in gathering respondents to fill out the BFI questionnaire, the results obtained only 287 Twitter users: 91 users for openness, 57 users for conscientiousness, 49 users for extraversion, 47 users for agreeableness, and 43 users for neuroticism. The distribution of the user personality traits can be seen in Figure 2.

2.3. Data Crawling

Data crawling is the process of collecting data from the Twitter website by utilizing the Twitter API [5]. The crawling process is carried out using a crawling data system that has been built by previous research [15]. The attributes we used are social behavior features and user tweets. We collected ten features of social features, including the number of followers, number of following, number of mentions, number of hashtags, number of URLs, number of media URLs, number of retweets, number of punctuations, number of uppercase, and number of tweets. The total amount of data collected for this study is 500.000 tweets from 287 Twitter users.

2.4. Preprocessing

Preprocessing is the first step in processing raw data. There are six preprocessing stages: data cleaning is a process of deleting symbols, numbers, retweets, and URLs from tweets; case folding is the stage to change capital letters to lowercase letters; tokenizing, the process of separating words and making a token; stop words is the step of eliminating all words that have no critical meaning; normalization is a process of normalizing words that have the same meaning but with different writings; and finally stemming, which is the process of changing words to become basic words by removing affixes to words. At the stop words and stemming stages, the author uses the python library "Sastrawi" [16][17].

2.5. Linguistic Inquiry Word Count (LIWC)

Linguistic Inquiry Word Count (LIWC) is one of the close-vocabulary methods to count text through language categories that have been determined and developed since 2007 by Pannebaker [5]. LIWC features have been tested, validated, and widely applied for psychological text analysis because LIWC counts words based on psychological word meaning categories. Table 1 is the correlation scores of the Big Five personality and the LIWC category that have been developed by previous research [11]. The stages carried out in the development of LIWC, namely: word collection, evaluation by experts, psychometric evaluation, as well as changes, and expansion. LIWC analyzes words to use based on more than 70 categories, including emotionality, function words (such as nouns, adverbs, etc.), personal issues (such as work, money, and religion), and social relationships [18][19][20].

Table 1 Correlation Score

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
1st Person 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
Preposition	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
Commnication	-0,06	-0,07	0,13	0,02	0
Friends	-0,01	0,06	0,15	0,11	-0,08
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	0,05	-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

DOI: https://doi.org/10.29207/resti.v6i1.3529

2.6. Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) is a word weighting method. The TF-IDF measures the frequency with which terms appear in a particular document and considers the significance of terms in the context of the entire document [21]. The greater the frequency of the words that appear, the higher the weight that will be given. TF-IDF can be calculated by equation (1).

$$TF \times IDF\ (d,t) = tf_{d,t} * log\frac{N}{df_{(t)}} \qquad (1)$$

Where $tf_{d,t}$ is the frequency of term t in the document, then N is the number of all documents, and lastly $df_{(t)}$ is the number of documents containing term t [11].

2.7. Classification with SVM

Support Vector Machine (SVM) is an algorithm for classification but can also be used for regression [22]. The basic principle of SVM is linear classification, but it has been developed to overcome non-linear problems by applying the concept of kernel trick [23]. By maximizing the margin between two classes, the SVM method maps the sample points into high dimensional feature space to find an ideal separating hyperplane [24]. The illustration shows in Figure 3.

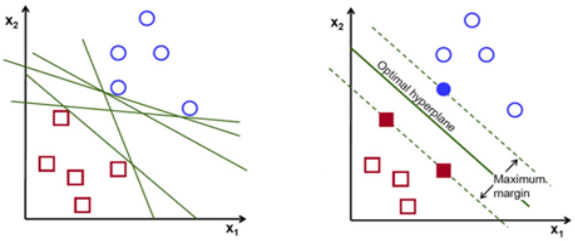

Figure 3 Optimal hyperplane illustration on SVM [25]

This study will test three types of kernels, namely Linear, Radial Basis Function (RBF), and Polynomials. The kernel that produces the best accuracy at a certain data ratio will be used for the second test scenario to completion. The following equation (2, 3, and 4) will calculate each kernel.

$$\text{Linear: } K(XiXj) = Xi^T Xj \qquad (2)$$

$$\text{RBF: } K(XiXj) = \exp\left(-\gamma \parallel Xi - Xj \parallel\right)^2 \qquad (3)$$

$$\text{Polynomials: } K(XiXj) = (X^T Xj + 1)^d \qquad (4)$$

This research implements three techniques to optimize the system performance: sampling, feature selection, and hyperparameter tuning. The sampling technique is one way to overcome the imbalance of the dataset. There are two sampling methods: under-sampling to remove several objects from the majority class and oversampling to add objects to the minority class [24]. In this study, the authors compare Random Under-Sampling (RUS) methods for under-sampling, Random Over-Sampling (ROS) for oversampling, and modified oversampling techniques SVM-SMOTE.

Feature selection is one of the critical roles. The most relevant feature has an impact on the accuracy system. There are five most conventional feature selection algorithms, namely Information Gain (IG), Chi-Squared method (CHI), Pearson Coefficient Correlation (PCC), Symmetrical Uncertainty Attribute Evaluation (SU), and CFS-Based Subset Evaluator (CFS) [26]. In this study, the authors used the chi-squared (CHI) method. The CHI (x^2) test was used for categorical features in the data set. The best CHI score will be used as the selected feature by calculating the CHI value between each feature and the target [26][27]. The calculation of CHI is shown in equation (5).

$$x^2 = \sum_{i=0}^{r}\sum_{j=1}^{n}\frac{(nij - eij)^2}{eij} \qquad (5)$$

Hyperparameter tuning is a technique for improving the performance of any learning algorithm. Some optimization techniques that can be used are grid search, random search, evolutionary algorithm, sequential model-based optimization [28]. In this research, the author uses grid search as a technique for algorithm optimization. The grid search algorithm works by trying all possible combinations of parameter values and returning the combination with the highest accuracy [29].

2.8. Performance Evaluation

This study uses a confusion matrix to testing the system. The accuracy values will be used to assess the performance of each model. The calculation of accuracy does not distinguish the correct label in different classes so that the number of true negative classes is included in the calculation [10]. calculation the accuracy value using equation 6.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (6)$$

3. Results and Discussions

This section will describe by showing the accuracy value of each scenarios. There are four experiment scenarios in this study. In the first scenario, we compare several kernel SVM, and the best result will be the baseline. The second scenario is tested using SVM with a sampling technique and then comparing them. The third scenario was conducted using social behavior features and LIWC features with or without chi-square. Finally, the test was carried out using a combination of social behavior data, LIWC, and TF-IDF. The outcome of each scenario (Table 2 to 6) is the average of the five trials.

DOI: https://doi.org/10.29207/resti.v6i1.3529

3.1. Result

After several testing using three SVM kernel functions with the total ratio of training data and test data based on research by [10] are 60:40, 70:30, 80:20, and 90:10. All kernel function configurations default from the python library. It was found that using the RBF kernel with 90:10 ratio data for social behavior features resulted in the highest accuracy of 48.28%. So, this is used as a baseline for this research. The score of the baseline is used as a point of comparison to evaluate all other methods in this research. The highest accuracy was obtained because the RBF kernel is better for data that is not linearly separated and has no prior knowledge of the data. Thus, the greater the amount of data used to train, the higher the accuracy can be obtained. The result is shown in Table 2.

Table 2 First Scenario Accuracy Result

Ratio Data	Average Kernel Accuracy (%)		
	Linear	RBF	Polynomial
60:10	33.72	37.58	33.22
70:30	34.24	40.44	37
80:20	37.94	45.18	37.22
90:10	38.6	**48.28**	37.92

In the second scenario, we balance the data using ROS, RUS, and SVM-SMOTE sampling techniques. This process aims to determine which sampling technique effectively balances the data and increases the accuracy score.

Table 3 Second Scenario Accuracy Result

Sampling	Average Accuracy (%)
RUS	32.74
ROS	50
SVM-SMOTE	**56.28 (+16.57)**

Table 3 shows that using SVM-SMOTE as a sampling technique gains the best accuracy than the others. The accuracy achieves 56.28%, it increases 16.57% from baseline. That happened because RUS and ROS balance the data by duplicating data randomly and causing overfitting [30]. While the SVM-SMOTE generate synthetic data randomly along the line that connects each minority class supporting vector to a few of its closest neighbors [31].

In the third scenario, we use a combination of social behavior features and LIWC features. This aims to increase the accuracy score that has been obtained previously. In this stage, we also implement no chi-square and with chi-square.

Table 4 Third Scenario Accuracy Result

Conditions	Average Accuracy (%)
SVM (Baseline)	48.28
Baseline + SMOTE + LIWC	65.18 (+35.00)
Baseline + SMOTE + Feature Selection LIWC	**71.74 (+48.6)**

Based on Table 4, SVM combined with SMOTE and LIWC features can increase the accuracy score by 65.18%. It shows that using linguistic features is effective in increasing the result of accuracy score. But using selecting features with chi-square was more effective in improving the accuracy score than only using the social behavior feature with all features in LIWC. By using chi-square obtained an accuracy of 71.74%, with an increase of 48.6% above the baseline.

From the results of previous experiments, the best feature is to add the LIWC features from the feature selection with chi-square. Therefore, in the fourth scenario, we use three features' combinations, namely social behavior, LIWC (Feature Selection), and TF-IDF. Then to optimize the accuracy results, we perform hyperparameter tuning with the grid search method. The results of this experiment show in Table 5.

Table 5 Fourth Scenario Accuracy Result

Conditions	Average Accuracy (%)
SVM (Baseline)	48.28
Baseline + SMOTE + Feature Selection LIWC + TF-IDF	75.92 (+57.25)
Baseline + SMOTE + Feature Selection LIWC + TF-IDF + Hyperparameter Tuning	**84.22 (+74.44)**

From the result in the fourth scenario, adding TF-IDF as a linguistic feature and implement hyperparameter tuning improved the accuracy score to 84.22%, with an increase of 74.44% from baseline.

Table 6 Comparison Personality Traits Accuracy Result

Personality Traits	Accuracy (%)	
	Baseline	Baseline + SMOTE + Feature Selection LIWC + TF-IDF + Hyperparameter Tuning
Openness	68,9	74,3
Conscientiousness	72,4	94,9
Extraversion	82,7	84,6
Agreeableness	75,8	87,2
Neuroticism	75,8	77

According to Table 5, the conditions that have the greatest influence on the accuracy score are baseline combined with linguistic features and optimization techniques. Furthermore, the accuracy score for each personality traits were calculated using equation (6). The comparison of accuracy result for each personality traits from baseline with the last scenario is shown in Table 6.

3.2. Discussion

Based on all test results, SVM with LIWC and TF-IDF linguistic approaches can increase the accuracy value. However, this is also assisted by implementing oversampling technique namely SVM-SMOTE due to the limitation in this research that the data was also imbalance. We also uses other optimization techniques,

DOI: https://doi.org/10.29207/resti.v6i1.3529

such as feature selection with CHI, and hyperparameter tuning with grid search to obtain more optimal results. Figure 4 shows the increase in the accuracy value against the baseline in each test scenario carried out.

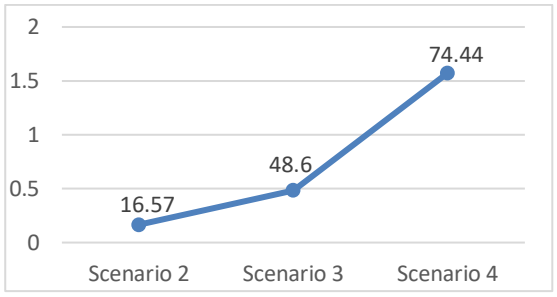

Figure 4 Accuracy Score Increase

4. Conclusion

This research presents an evaluation of the SVM method combined with linguistic approach and optimization techniques to improve the personality prediction system performance. The dataset size we used was as many as 287 Twitter users and 500,000 tweets data. Compared to utilizing only the SVM method, combining SVM with two linguistic approaches (LIWC and TF-IDF) and optimization techniques (SMOTE, chi-square, grid search) yielded better performance outcomes. The application of optimization techniques has proved to have a substantial influence on performance results. It's because SMOTE was able to deal with the dataset's imbalance, chi-square gives the feature more related to the classes, and grid search finds the best parameter to the RBF kernel. Suggestions for further research are expected to improve the performance of personality prediction systems by expanding the dataset because this research still used small size of data and also developing this research using various methods.

Reference

[1] D. Preoţiuc-Pietro, J. Carpenter, S. Giorgi, and L. Ungar, "Studying the dark triad of personality through twitter behavior," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 24-28-Octo, pp. 761–770, 2016, doi: 10.1145/2983323.2983822.

[2] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Inf.*, vol. 9, no. 5, pp. 1–20, 2018, doi: 10.3390/info9050127.

[3] C. Li, J. Wan, and B. Wang, "Personality Prediction of Social Network Users," *Proc. - 2017 16th Int. Symp. Distrib. Comput. Appl. to Business, Eng. Sci. DCABES 2017*, vol. 2018-Septe, pp. 84–87, 2017, doi: 10.1109/DCABES.2017.25.

[4] V. Varshney, A. Varshney, T. Ahmad, and A. M. Khan, "Recognising personality traits using social media," *IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017*, pp. 2876–2881, 2018, doi: 10.1109/ICPCSI.2017.8392248.

[5] A. T. Damanik and Masayu Leylia Khodra, "Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression," *J. Cybermatika*, vol. 3, no. 1, pp. 14–22, 2015.

[6] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality Prediction System from Facebook Users," *Procedia Comput. Sci.*, vol. 116, pp. 604–611, 2017, doi: 10.1016/j.procs.2017.10.016.

[7] G. D. Salsabila and E. B. Setiawan, "Semantic Approach for Big Five Personality Prediction on Twitter," *RESTI*, vol. 5, no. 4, pp. 680–687, 2021.

[8] S. Maloji, K. Mannepalli, N. S. J, K. B. Sri, and C. Sasidhar, "Big Five Personality Prediction from Social Media Data using Machine Learning Techniques," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 2412–2417, 2020, doi: 10.35940/ijeat.d7946.049420.

[9] D. E. Cahyani and A. F. Faishal, "Classification of Big Five Personality Behavior Tendencies Based on Study Field with Twitter Analysis Using Support Vector Machine," *7th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2020 - Proc.*, pp. 140–145, 2020, doi: 10.1109/ICITACEE50144.2020.9239130.

[10] 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," *2019 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Big Data Artif. Intell. IC3INA 2019*, pp. 114–119, 2019, doi: 10.1109/IC3INA48034.2019.8949601.

[11] F. Ilzam Nur Haq and E. Budi, "Implementasi Naive Bayes Classifier untuk Prediksi Kepribadian Big Five pada Twitter Menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dan Term Frequency-Relevance Frequency (TF-RF) Program Studi Sarjana Ilmu Komputasi Fakultas Informatik," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9785–9795, 2019.

[12] 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.

[13] Y. J. Nie, G. J. Gao, Y. X. Wang, D. X. Liu, and K. Gao, "Personality predicting model based on user's linguistic behavior," *Proc. 2017 9th Int. Conf. Model. Identif. Control. ICMIC 2017*, vol. 2018-March, no. Icmic, pp. 827–832, 2018, doi: 10.1109/ICMIC.2017.8321569.

[14] R. R. Mccrae *et al.*, "The NEO – PI – 3 : A More Readable Revised NEO Personality Inventory The NEO – PI – 3 : A More Readable Revised NEO Personality Inventory," *J. Pers. Assess.*, vol. 84, no. 3, pp. 261–270, 2016, doi: 10.1207/s15327752jpa8403.

[15] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter," no. October 2018, pp. 11–16, 2016, doi: 10.21108/indosc.2016.111.

[16] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, May 2003, doi: 10.1080/713827180.

[17] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, pp. 170–174, 2016, doi: 10.1109/ICODSE.2015.7436992.

[18] J. Golbeck, "Predicting Personality from Social Media Text," *AIS Trans. Replication Res.*, vol. 2, no. September, pp. 1–10, 2016, doi: 10.17705/1atrr.00009.

[19] I. Ergu, Z. Isik, and I. Yankayis, "Predicting Personality with Twitter Data and Machine Learning Models," Oct. 2019, doi: 10.1109/ASYU48272.2019.8946355.

[20] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.

[21] E. Tighe and C. Cheng, "Modeling Personality Traits of Filipino Twitter Users," pp. 112–122, 2018, doi: 10.18653/v1/w18-1115.

[22] C. Zoltan, "SVM and Kernel SVM | Towards Data Science." https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200 (accessed Aug. 26, 2021).

[23] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1," 2003, Accessed: Aug. 27, 2021. [Online]. Available:

DOI: https://doi.org/10.29207/resti.v6i1.3529

http://asnugroho.net.

[24] L. Demidova and I. Klyueva, "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem," *2017 6th Mediterr. Conf. Embed. Comput. MECO 2017 - Incl. ECYPS 2017, Proc.*, no. June, pp. 17–20, 2017, doi: 10.1109/MECO.2017.7977136.

[25] P. P. Ippolito, "SVM: Feature Selection and Kernels | Towards Data Science." https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c (accessed Sep. 20, 2021).

[26] A. Al Marouf, M. K. Hasan, and H. Mahmud, "Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction from Social Media," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 3, pp. 587–599, 2020, doi: 10.1109/TCSS.2020.2966910.

[27] A. Chugh, "ML | Chi-square Test for feature selection | GeeksforGeeks." https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/ (accessed Aug. 25, 2021).

[28] G. Y. N. N. Adi, M. H. Tandio, V. Ong, and D. Suhartono, "Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia," *Procedia Comput. Sci.*, vol. 135, pp. 473–480, 2018, doi: 10.1016/j.procs.2018.08.199.

[29] U. Malik, "Cross Validation and Grid Search for Model Selection in Python | Stack Abuse." https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/ (accessed Aug. 26, 2021).

[30] A. A. Arifiyanti and E. D. Wahyuni, "Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, pp. 34–39, 2020, doi: 10.33005/scan.v15i1.1850.

[31] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradig.*, vol. 3, no. 1, p. 4, 2011, doi: 10.1504/ijkesdp.2011.039875.

DOI: https://doi.org/10.29207/resti.v6i1.3529