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. 6 (2022) 1072-1078 ISSN Media Electronic: 2580-0760

Naïve Bayes-Support Vector Machine Combined BERT to Classified Big Five Personality on Twitter

Billy Anthony Christian Martani[1], Erwin Budi Setiawan[2]
[1,2]Informatics, School of Computing, Telkom University
[1]billymartani@student.telkomunversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstract

Twitter is one of the most popular social media used to interact online. Through Twitter, a person's personality can be determined based on that person's thoughts, feelings, and behavior patterns. A person has five main personalities likes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This study will make five personality predictions using the Naïve Bayes method – Support Vector Machine, Synthetic Minority Over Sampling Technique (SMOTE), Linguistic Inquiry Word Count (LIWC), and Bidirectional Encoder from Transformers Representations (BERT). A questionnaire was distributed to people who used Twitter to collect and become a dataset in this research. The dataset obtained will be processed into SMOTE to balance the data. Linguistic Inquiry Word Count is used as a linguistic feature and BERT will be used as a semantic approach. The Naïve Bayes method is used to perform the weighting and the Support Vector Machine is used to classify Big Five Personalities. To help improve accuracy, the Optuna Hyperparameter Tuning method will be added to the Naïve Bayes Support Vector Machine model. This study has an accuracy of 87.82% from the results of combining SMOTE, BERT, LIWC, and Tuning where the accuracy increases from the baseline.

Keywords: BERT, Big Five Personality, LIWC, Naïve Bayes-Support Vector Machine

1. Introduction

In the world of work, working in groups can make it easier to complete a job. Knowing the personalities of other group members will be able to help maximize the work of the group. Personality can be defined by how a person interacts with the surrounding environment [1]. Personality comes from the Greek word "*persona*" which means a symbol that represents a person's identity [2]. Personality traits can be seen through a person's thoughts, feelings, and behavior patterns to respond to certain circumstances (Roberts 2009 p 140) [3]. According to Lewis Goldberg [1], there are 5 main personalities: Openness, Continuousness, Agreeableness, Extraversion, and Neuroticism, commonly abbreviated as OCEAN. Each of these personalities has its advantages.

Today, communicating is easier because of the online platform to communicate called social media. Twitter, Facebook, Instagram, WhatsApp, and many more are social media that are used by people to communicate in online. According to we are social media [4], Twitter became one of the most popular social media used among users aged 16 to 64 years in Indonesia in January 2013. Twitter is one of the social media that provides microblogging services that allow users to send and read messages up to 140 characters in one message called tweets [5]. There is no limit if someone wants to write tweets so that someone can freely express what they want to share.

Several studies try to examine a person's personality through the classification method of the words of Twitter users' tweets. One of them is research that has been done by Willy et al [6]., who tried to use the Term Frequency Inverse Relevance Frequency method to convert the word tweets into vectors and then use the Decision Tree C.45 method to classify Big Five Personality. This study produces a model that can predict a person's personality by 65.72%. The author revealed that the data used in this study contained dominant data labels, so the model detected more dominant data labels than the others.

Another research was conducted by Salsabila et al [7]. used a dataset of 295 users and 511,617 tweets using the Synthetic Minority Over Sampling Technique (SMOTE), Linguistic Inquiry Word Count (LIWC), and Bidirectional Encoder from Transformers Representations (BERT) methods which resulted in an accuracy of 80.07%. The author reveals that the

Accepted: 01-08-2022 | Received in revised: 10-09-2022 | Published: 30-04-2022

semantic approach can produce better accuracy because the previously trained BERT model is more applicable to understanding words in sentences. The weakness of the research mentioned by the author is that the dataset is still small, namely 295 Twitter users.

Research conducted by Gita et al [8]. uses the SVM model combined with the TF-IDF, LIWC, and Hyperparameter tuning model to detect Big Five Personality. TF-IDF and LIWC function as feature extraction. Hyperparameter tuning is used to help the model combine various possible parameters to obtain the best parameters which will later be applied to the model. The results of this study resulted in a baseline accuracy of 74.44% and when the Hyperparameter Tuning is used to baseline, it gains accuracy to 84.22%. The weakness of this research is the small dataset used by the author, which is 287 Twitter user data.

Research conducted by Zain et al [9]. regarding the effectiveness of Naïve Bayes SVM weighting in classifying film reviews resulted in an accuracy of 88.8%. Naïve Bayes uses n-gram extraction weighting in its process. In that study, deletion of stop words did not improve classification performance. The author recommends the feature extraction process to use unigram and bigram simultaneously.

Based on previous research, the results used the Naïve Bayes Support Vector Machine produced better accuracy values than the Support Vector Machine method. So, in this study, we would try to predict the big 5 personalities using the Naïve Bayes Support Vector Machine method. Naïve Bayes (NB) was a classification method that used probability and statistical methods. Support Vector Machine (SVM) was a classification method that had the convenience of classifying labels using a hyperplane. These two methods had been combined with Naïve Bayes which would be played a role in weighting while Support Vector Machine would be played a role in classification based on the results of Naïve Bayes weighting. Synthetic Minority Overside Technique (SMOTE) prediction model to handle imbalance data, Bidirectional Encoder from Transformers Representations (BERT) as a semantic approach, and Linguistic Inquiry Word Count as a linguistic feature. To help improve accuracy, the hyperparameter method, namely Optuna, had been used. The advantage of using Optuna was that the parameters were built dynamically so that it was more likely to get the best parameters that other hyperparameter methods may not be able to obtain [19]. The dataset used in this study was from previous researched [8].

This paper will be divided into 4 parts. The first part is an introduction as described above. The second part is the method used to build the Big 5 Personality prediction system. The third part will explain the results of the experiments and the last part will explain the conclusions of this paper.

2. Research Methods

There are several methods incorporated in the system development which will be explained in this section. The personality prediction system that will be built can be seen in Figure 1, the system consists of data crawling, data labeling, pre-processing, training process where NBSVM will be implemented with SMOTE, LIWC, BERT, and Hyperparameter tuning. The system will continue with personality detection prediction, and the performance evaluation process to evaluate the result.

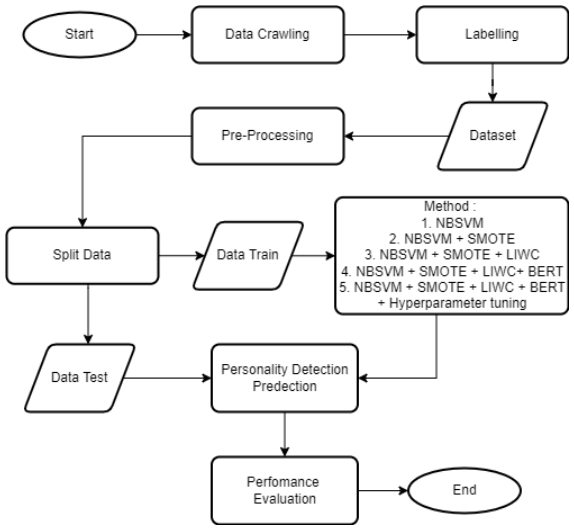

Figure 1 Personality System Prediction

2.1 Big Five Personality

Personality is how a person interacts with the surrounding environment [1]. Personality in today's humans can be categorized into 5 main personalities [10]. The first personality is Openness which has more value in terms of intelligence, imagination, open-minded, and sensitivity. The second personality is Conscientiousness, a personality that has more value in terms of being careful, responsible, and thorough. The third personality is Extraversion is a personality that has more value in terms of social, speaking, and activity. The fourth personality is Agreeableness which has more value in terms of being obedient, kind, simple, gentle, and cooperative. The last personality is neuroticism is a personality that has more value in terms of depression, anger, and insecurity.

2.2 Data Crawling

Data Crawling is a method of collecting and downloading data from a website [11]. Data will collect such as user followers, user following, the sum of tweets user, and the sum of some data such as tweets, URLs, media URLs, mentions, hashtags, retweets, and

DOI: https://doi.org/10.29207/resti.v6i6.4378

uppercase [7]. The data will become the Social Feature of the user which can be seen in Table 1 [7].

Table 1. Description of Social Feature Data

Social Feature	Descriptions
User Follower	The number of followers that the user has
User Following	The number of users following
Sum of Tweets	The number of users' tweets
Sum of URLs	The large number of URLs shared by users
Sum of media URLs	The large number of media URLs shared by users
Sum of Hashtags	The number of user hashtags
Sum of Retweets	The number of user retweets
Sum of Mentions	The number of user mentions
Sum of Uppercase	The number of uppercase letters used by Twitter users

2.3. Data Labelling

The data will label based on the results of the Big Five Inventory questionnaire [12]. This questionnaire has a total of 25 questions, each of the questions represents 1 to 5 scales, where 1 represents strongly disagree to 5 represents agree. The results of the questionnaire will determine the Big Five Personality label on Twitter users which is adjusted to OCEAN traits.

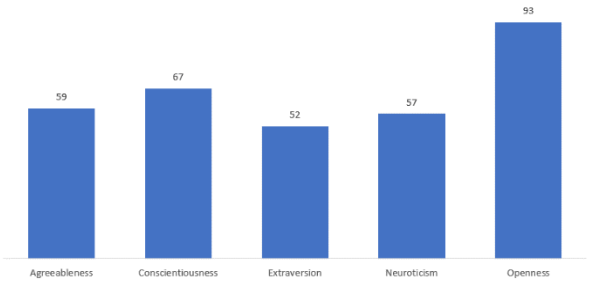

Figure 2 Big Five Personality Distribution on Twitter Users

In this research 295 user data have been collected from crawling. Figure 2 is the distribution of personality label data with agreeableness as many as 48 users, Conscientiousness as many as 59 users, Extraversion as many as 48 users, Neuroticism as many as 49 users, and Openness as many as 91 users.

2.4 Pre-processing

Pre-processing is one of the important factors in the classification algorithm to improve accuracy results [12]. Text Pre-processing will be divided into 6 stages in this study, namely cleansing, case folding, tokenizing, normalization, stop word, and stemming.

Cleansing aims to remove symbols, numbers, and URLs in sentences. Case folding aims to change uppercase letters to lowercase letters. Tokenizing is used to break sentences into words. Stop Word aims to eliminate words that have no meaning in a sentence. Normalization is used to normalize non-standard words into standard ones and stemming is the process of changing affixed words into their base words. The stemming process is carried out with the help of the "Sastrawi" library.

2.5 Synthetic Minority Over Sampling Technique (SMOTE)

Synthetic Minority Over Sampling Technique (SMOTE) is a data sampling technique for a minority class by connecting randomly selected data points from one of its k closest neighbors [13]. SMOTE is used to overcome data imbalances in the dataset. Imbalanced datasets can cause the model to be less accurate due to a lack of predictive accuracy with minority classes [14]. With the implementation of SMOTE, the dataset will be more balanced to increase the accuracy of the model.

2.6 Linguistic Inquiry Word Count (LIWC)

Linguistic Inquiry Word Count abbreviated as LIWC is a way to evaluate the emotional, cognitive, and structural components of a sentence based on the dictionary of words and their classification categories [15]. There are 2 vocabulary features in LIWC, namely open and closed vocabulary. In this study, a closed vocabulary feature will be used, namely calculating the correlation value of word categories based on the LIWC dictionary. Vocabulary was collected from the official website of the LIWC and translated into formal Indonesian. In previous studies [16], the correlation value of LIWC with the Big Five personality is shown in Table 2 [8].

Table 2. Correlation Score of LIWC

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
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

DOI: https://doi.org/10.29207/resti.v6i6.4378

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

2.7 Bidirectional Encoder from Transformers Representations (BERT)

The Bidirectional encoder from Transformers representations abbreviated as BERT is a method to extract features in the text in Nature Language Processing, such as sentiment classification, reading comprehension, and answering questions [17].

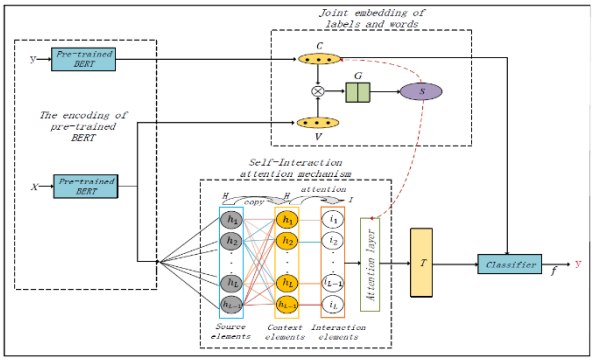

Figure 3 BERT Architecture [17]

In Figure 3, the dataset described by X will be processed into 2 stages, the first stage of the X dataset will be processed to become a sentence vector. The result of this process will produce a comprehensive value of semantics. Then the second X dataset will be created in the form of a token. The result of this process is to give a label to the token so that later it will be easier to implement into the classifier. Both processes will be included in the classifier for use in the next process.

2.8 Naïve Bayes Support Vector Machine (NBSVM)

Naïve Bayes is an easy algorithm to implement for classification because it has low complexity, which means that the training process, Naïve Bayes doesn't need too much data train [9]. Support Vector Machine or SVM is a method using hyperplanes that group data with maximum margins [18]. The kernel in SVM can maximize the results of grouping because each kernel has a different grouping calculation. An example of SVM can be seen in Figure 4.

For the above reasons, two methods can be combined with the Support Vector Machine tasked with calculating the ratio of the log and Naïve Bayes as a data retrieval that has the performance of taking values from data retrieval.

Naïve Bayes and Support Vector Machine can produce better performance by following equation (1)

$$w^t = (1-\beta)\omega + \beta w \quad (1)$$

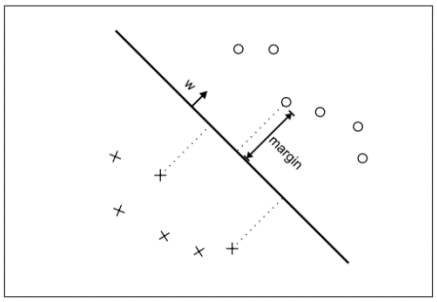

Figure 4 Support Vector Machine Illustration [18]

From equation (1) where ω= ‖w‖/|V| which is the result of the mean of w and β= [0,1] which is the interpolation parameter

To gain more accuracy, we will add Hyperparameter Tuning in the Naïve Bayes Support Vector Machine called Optuna. Optuna is a hyperparameter method that uses the define-by-run principle, which allows users to dynamically construct parameters [19]. In addition, Optuna has several advantages such as being able to increase the effectiveness of optimization by performing efficient searches and pruning algorithms.

2.9 Evaluation Performance

To measure the value of the performance evaluation of this study, the Confusion Matrix will be used. A confusion Matrix is a matrix that contains the actual classification and predictions made by the classification algorithm [20].

In Figure 5, there are 2 parts, namely predicted and actual. For example, N_{ij} is showing how many samples are identified in actual A_i but classified in A_j. By using the results in Figure 4, several measurements were obtained, namely Accuracy, Precision, Recall, and F1-Score.

		Predicted			
		A_1	... A_j ...		A_n
Actual	A_1	N_{11}	N_{1j}		N_{1n}
	A_i	N_{i1}	... N_{ij} ...		N_{in}
	A_n	N_{n1}	N_{nj}		N_{nn}

Figure 5 Confusion Matrix [20]

Precision is how many correct predictions from a class where the total number of predictions from that class. Precision can be calculated using the following equation (2)

$$Precision = Nii \,/ \sum_{k=1}^{n} N_{ki} \quad (2)$$

DOI: https://doi.org/10.29207/resti.v6i6.4378

The recall is how many correct predictions from a class where the total number of predictions from the actual number in that class. Recall can be calculated using the following equation (3)

$$Recall = Nii / \sum_{k=1}^{n} N_{ik} \quad (3)$$

F1-Score is the average of recall and precision. F1-Score can be calculated using the following equation (4)

$$F1 - Score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (4)$$

Due to imbalanced data, the accuracy will be calculated by looking at the F1-Score because F1-Score is obtained from calculating the average of precision and recall.

3. Results and Discussions

In this studied, we apply the naïve Bayes support vector machine method as a baseline. Next, we had built four scenarios. In the first scenario, we would tried to find the best ratio data that had been used in the next scenario. The second scenario would tried to compare the data that balanced with SMOTE and those that did not used SMOTE. The third scenario would tried to compare the data used feature expansion, LIWC, and BERT. And the last scenario would tried the model with hyperparameter tuning. The dataset used in this studied was 295 twitter users and 511.617 users' tweets.

3.1 *Result*

For the first scenario, we would look for the optimal data ratio that had been used in this study. The ratio data have been tested were 90:10, 80:20, 70:30, and 60:40. The test results in the first scenario could been seen in table 3. The table shows a data ratio of 80:20 resulting in the highest accuracy valued. With this, the 80:20 data ratio would continue have been used in the next scenario.

Table 3. First Scenario Result Comparison

Ratio	Accuracy (%)
90:10	38.62
80:10	42.71
70:10	29.09
60:10	31.77

In the second scenario, an experiment had been conducted to determine the effectiveness of data balancing. The results of the second scenario test could been seen in table 4. From the results of the second scenario test, it was found that using SMOTE could increase accuracy by 3.19%. This result proves that balancing the data could increase the accuracy of the model prediction so that smote had been applied to the next scenario.

In the third scenario, we would tried to apply LIWC as linguistic features and BERT as semantic approaches to improved accuracy results. The results of the third scenario experiment could been seen in table 5. From the results of the third scenario testing, the application of linguistic features could increase accuracy by 1.18% from the second scenario. This made LIWC would continue have been applied to improved accuracy in future tests. In addition, BERT as a semantic approach was also tested have been added as a feature expansion. The test results obtained an accuracy of 63.38%, an increase of 20.58% from the test if the model only used LIWC as a feature expansion. This made BERT proven to improved accuracy.

Table 4 Second Scenario Result Comparison

Condition	Accuracy (%)
NBSVM	42.71
NBSVM + SMOTE	50.83

Table 5. Third Scenario Result Comparison

Condition	Accuracy (%)
NBSVM + SMOTE	50.83
NBSVM + SMOTE + LIWC	52.01
NBSVM + SMOTE + LIWC +BERT	72.59

In the last scenario, the model had been applied used Optuna hyperparameter tuning. The results of the last scenario could been seen in table 6. From the test results, it was found that the accuracy increased to 87.82%. An increase of 10.53% occurred when Optuna hyperparameter tuning was applied to the NBSVM with SMOTE, LIWC, and BERT. Optuna was set to find the regularization valued and the maximum iteration valued that had been used to process the training model. The best valued had been applied to the model for further implementation into the prediction parameters.

Table 6. Fourth Scenario Result Comparison

Condition	Accuracy (%)
Baseline + SMOTE + LIWC + BERT	72.59
Baseline + SMOTE + LIWC + BERT + Hyperparameter Tuning	87.82

The comparison of four scenarios carried out by this studied could be seen in Table 7.

Table 7. All Scenario Result Comparison

Condition	Accuracy (%)
Baseline	44.67
Baseline + SMOTE	50.83 (+20.19)
Baseline + SMOTE + LIWC	52.01 (+33.98)
Baseline + SMOTE + LIWC + BERT	72.59 (+71.65)
Baseline + SMOTE + LIWC + BERT + Hyperparameter Tuning	87.82 (+107.66)

The results of testing the accuracy of the comparison of labeled personality traits based on the results of the NBSVM combined with BERT, LIWC, SMOTE, and Hyperparameter Tuning could been seen in Table 8.

Table 8. Comparison Personality Traits Accuracy Result

Personality Traits	Accuracy (%)
Openness	85.71
Conscientiousness	85.71
Extraversion	86.49
Agreeableness	97.30
Neuroticism	83.87
Average Accuracy	87.82

DOI: https://doi.org/10.29207/resti.v6i6.4378

3.2 Discussions

The dataset used in this studied was imbalanced. This could been seen in figure 2 where the openness label was the dominant label compared to other labels. This could lead to a decrease in accuracy due to the lack of predictive accuracy with the minority class. To overcome this problem, the oversampling method was used, namely SMOTE. SMOTE would create new data through data points randomly from one of its k nearest neighbors. The amount of data created by the label would followed the amount of the highest label so that the data had been balanced.

In the first scenario, it was found that a good data ratio to use a data ratio of 80:20. The ratio of 80:20 produces the highest accuracy compared to the other ratios, which was 44.67%. Because of that, the ratio of 80:20 had been used in the next scenario. The second scenario would tried to test the model with balanced data using SMOTE. From the test results, getting that balanced data could be improved accuracy. This could happen because by balancing the data, there had been no minority class so the model would not be dominant considering the majority class.

The third scenario was a test that used feature expansion, namely LIWC and BERT. From the test results, it was proven that LIWC could be improved accuracy. LIWC could group words into a category dictionary so that it could generate new word groups. This was what made the model predict more accurately because more data explains the label. Then tried to apply BERT and got an increase in accuracy. BERT could group words into certain groups which would then be labeled as groups so that word groups had been formed.

In the last scenario, Optuna hyperparameter tuning had been added to improve accuracy. The results obtained proved that adding Optuna hyperparameter tuning could increase accuracy. Optuna hyperparameter tuning would be looked for the best parameters had been applied to the model so that it was proven that the implementation of hyperparameter tuning could increase the accuracy of the prediction model. The results of the increased accuracy in the experiment could had been seen in figure 5 below.

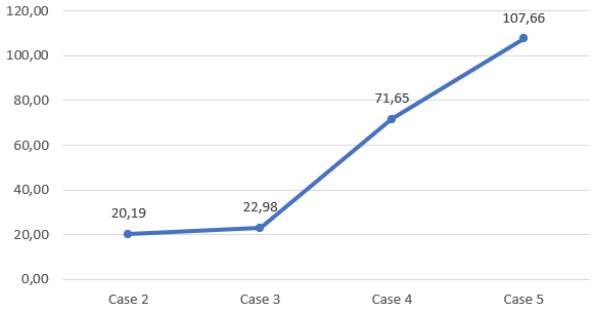

Figure 5 Accuracy Score Increase

4. Conclusion

In this researched, we tried to predict the Big Five Personalities of a twitter user by used the naïve bayes support vector machine method. The data we collected was 295 twitter users and 672.866 tweets from crawling results. The data could been said have been imbalanced because there was a label that dominated so data balancing must been done. To overcome the imbalanced data, we apply the SMOTE method. In addition, we also include LIWC as a linguistic feature, BERT as a semantic approached, and used Optuna hyperparameter tuning to improved model accuracy. In the results of this studied, the model managed to achieve an accuracy of 87.82%.

From the test results, it could been concluded that adding hyperparameter tuning to the model had been able to increase accuracy because when tuning was done, the model would looked for the best parameters have been used later in the prediction process. It was also recommended to tried other methods to produced better accuracy with the hyperparameter tuning method.

References

[1] Ipqi. 2016. “Teori Kepribadian Model Lima Besar (Big Five Personality)”, https://ipqi.org/category/management-article/hr-productivity/, (accessed Nov. 02, 2021)

[2] adm1. 2018. “Apa Beda Karakter, Kepribadian, Sifat, dan Temperamen?”, https://seputargk.id/apa-beda-karakter-kepribadian-sifat-dan-temperamen/, (accessed Nov. 02, 2021)

[3] Roige, S. S., Gray, J. C., Mackillop, J. K., Chen, C.-H., & Palmer, A. A. (2014). The genetics of human personality. *The Laryngoscope*, *2*, 2–31.

[4] Riyanto, Andi Dwi. “Hootsuite (We are Social): Indonesian Digital Report 2021”, https://andi.link/hootsuite-we-are-social-indonesian-digital-report-2021/ (accessed Nov. 02, 2021)

[5] D., Kustin Ayuwuragil. “Twitter”, https://m.merdeka.com/twitter/profil/, (accessed Nov. 02, 2021)

[6] Willy, Setiawan, E. B., & Nugraha, F. N. (2019). Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on social media Twitter. *2019 International Conference on Computer, Control, Informatics, and Its Applications: Emerging Trends in Big Data and Artificial Intelligence, IC3INA 2019*, *October 2019*, 114–119. https://doi.org/10.1109/IC3INA48034.2019.8949601

[7] Salsabila, G. D., & Setiawan, E. B. (2021). Semantic Approach for Big Five Personality Prediction on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 680–687. https://doi.org/10.29207/resti.v5i4.3197

[8] Gita Safitri, & Erwin Budi Setiawan. (2022). Optimization Prediction of Big Five Personality in Twitter Users. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(1), 85–91. https://doi.org/10.29207/resti.v6i1.3529

[9] Zain, F. F., & Sibaroni, Y. (2019). Effectiveness of SVM Method by Naïve Bayes Weighting in Movie Review Classification. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *5*(2), 108–114. https://doi.org/10.23917/khif.v5i2.7770

[10] Sonia, Roccas. Lilach, Sagiv. H., Schwartz Shalom. Ariel, Knafo. 2002. “The Big Five Personality Factors and Personal Values”. *SAGE.*

DOI: https://doi.org/10.29207/resti.v6i6.4378

[11] Eka Sembodo, J., Budi Setiawan, E., & Abdurahman Baizal, Z. (2016). *Data Crawling Otomatis pada Twitter*. *September* 11–16. https://doi.org/10.21108/indosc.2016.11

[12] Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*(3), 319–335. https://doi.org/10.1007/s10588-018-9266-8

[13] Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

[14] Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2011). Class imbalance, redux. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 754–763. https://doi.org/10.1109/ICDM.2011.33

[15] Goncalves, P., Fabrício, B., Matheus, A., & Meeyoung, C. (2013). Comparing and Combining Sentiment Analysis Methods Categories and Subject Descriptors. *Proceedings of the First ACM Conference on Online Social Networks*, 27–38.

[16] Ilzam Nur Haq, F., & Budi, E. (2019). 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 of Engineering*, *6*(2), 9785–9795.

[17] Dong, Y., Liu, P., Zhu, Z., Wang, Q., & Zhang, Q. (2020). A Fusion Model-Based Label Embedding and Self-Interaction Attention for Text Classification. *IEEE Access*, *8*, 30548–30559. https://doi.org/10.1109/ACCESS.2019.2954985

[18] Tong, S., & Koller, D. (2009). Support Vector Machine Active Learning with Applications to Text Classfication. *American Quarterly*, *61*(2), 417–421. https://doi.org/10.1353/aq.0.0077

[19] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

[20] Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

DOI: https://doi.org/10.29207/resti.v6i6.4378