Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

Identification of Big Five Personality on Twitter User using the AdaBoost Method

Ajeung Angsaweni[1], Warih Maharani[2,*]

Informatics Department, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]ajeung@student.telkomuniversity.ac.id, [2,*]wmaharani@telkomuniversity.ac.id
Correspondence Author Email: wmaharani@telkomuniversity.ac.id
Submitted: **12/07/2022**; Accepted: **14/08/2022**; Published**: 30/09/2022**

Abstract−Social media is one of the lifestyles in the modern era that uses web-based technology for social interaction. As one of the most popular social media platforms, Twitter allows users to express themselves through tweets that can show their personality. The Big Five theory states that a person's personality is divided into five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Several methods have been used to conduct user personality research based on activity on social media. The AdaBoost method is used in this study to identify the personality of Twitter users using sentiment, emotion, social, PCA, and POS-tag features. There are two test scenarios in this study. The first is testing the AdaBoost model with all features, and the second is testing the AdaBoost model with a combination of three features. The research indicates that the data preprocessing method can affect the model. The results showed that the AdaBoost model with all the features and without the stemming process had the highest accuracy value of 53.57%.

Keywords: Twitter; Big Five; AdaBoost

1. INTRODUCTION

Social media has an impact on changes in human social life. Using internet-connected devices, users can communicate, interact, and share content on social media platforms. According to DataReportal's data in Digital 2022: Indonesia [1], Indonesia has 191.4 million active social media users out of a population of 277.7 million. As one of the most popular networking sites, Twitter allows users to post messages to the internet through text, pictures, and videos, known as tweets. In addition, social media user activities can reveal information about their personalities. Personality is a unique combination of behaviors, emotions, and thought patterns that influence cognition, motivation, behavior, and life choices [2], [3]. Twitter is considered an accurate application for identifying personality because Twitter users tend not to worry about the words used in tweets [4].

To identify a person's personality, research can be done using the Five-Factor Model approach, also known as The Big Five. This research method is widely used and has been used successfully in several previous studies [4]–[6]. This theory divides personality into five dimensions known as OCEAN, which stands for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [2], [7]. Personality prediction research can also be done using the Myer-Briggs Type Indicator (MBTI) approach [8], [9] or DISC [10], [11] which classifies personality into dominance, inducement, submission, and compliance.

The flow of personality prediction research generally involves the processes of data extraction, preprocessing, feature extraction, and classification [12]. Pratama and Maharani [6] conducted a study in which personality prediction was performed using a Random Forest classification with a feature extraction process consisting of emotional, sentiment, and social analysis based on statistical data from user Twitter accounts. The highest accuracy is 69.23%, obtained by combining features and comparing test data 20:80 [6]. Classification with Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes is also used in personality prediction research by Maharani and Effendy [4]. Research [4] has found openness to be the simplest trait to measure, while neuroticism is considered quite difficult. In predicting personality traits, the SVM approach produces a good performance, with a score of 59.45% [4]. Larger sample size and more features will yield better results [4]. Aditi and Suja [13] compared the Multinomial Naïve Bayes (MNB), Latent Dirichlet Allocation (LDA), and AdaBoost methods to predict personality using an English dataset. Research [13] used a multi-label classification and found that MNB had the highest accuracy, while AdaBoost and LDA had nearly the same accuracy on all labels except openness.

In ensemble learning, the boosting algorithm iteratively trains the weak classifier and then adds it to the strong classifier by minimizing bias or variance due to training records [14]. Adaptive Boost, or AdaBoost, is one of the boosting algorithms that can improve accuracy by combining several weak learners and correcting previous weak classifier errors [15]. However, this method is noise-sensitive, which means that if there is more noise data, the AdaBoost algorithm will spend more time and be less efficient [16]. AdaBoost is considered to be adaptive because it assigns weights to the base model based on model accuracy and changes the weights of the training data based on prediction accuracy [14]. Thus, the background of this study uses AdaBoost as a method of identifying personality.

Due to limited resources, personality research is conducted using the Big Five approach. The Indonesian dataset is derived from user tweets and used as sentiment, emotion, and POS-tagging features. Twitter user account statistics are also used as a social feature consisting of the number of tweets, following, followers, and favorites. This paper is structured as follows. The research methods will discuss in section 2, the results will discuss in section 3, and conclusions and future work will discuss in section 4.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

2. RESEARCH METHODOLOGY

2.1 Research Flow

The personality identification process of this study will consist of five major stages: data crawling, data preprocessing, feature extraction, classification, and evaluation. The flowchart in Figure 1 illustrates the process, which will then be explained in each sub-chapter.

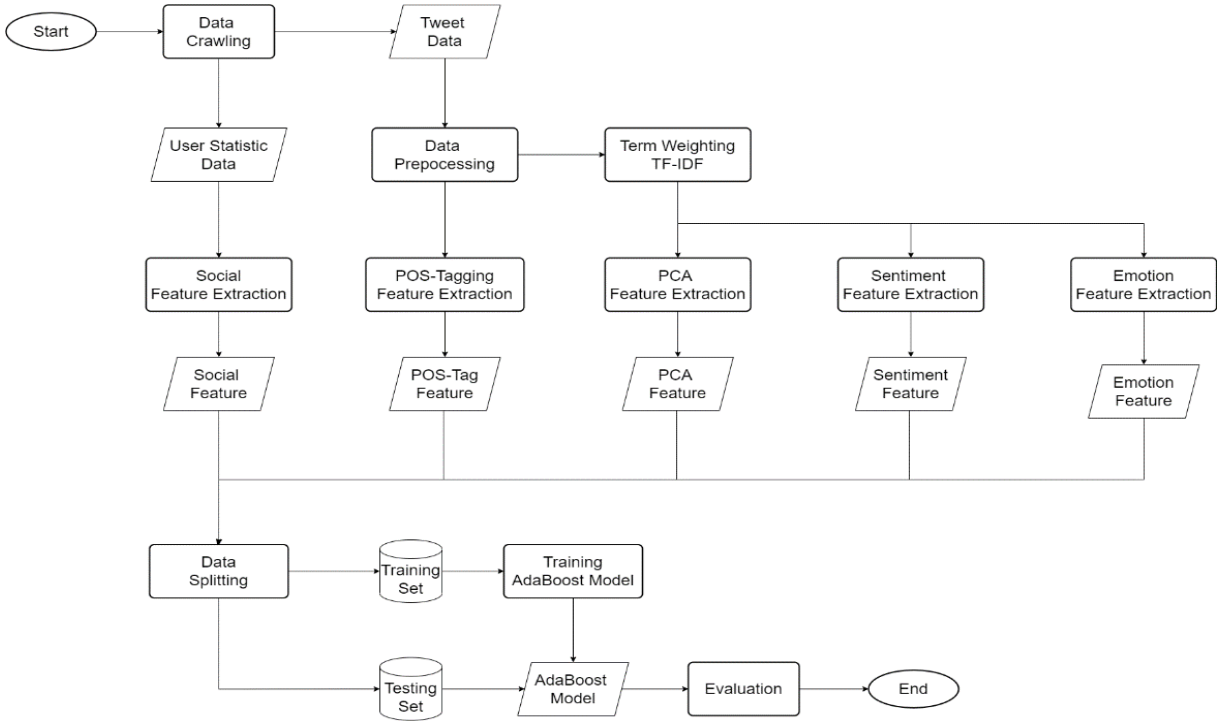

Figure 1. Flowchart

2.2 Data Crawling

In this study, the dataset used was an Indonesian dataset of Twitter users who had filled out a questionnaire in the form of personality questions based on the Big Five Inventory (BFI). The dataset is collected using the Twitter API, which consists of 1000 tweets per account and user account statistics such as the number of tweets, followers, following, and favorites, all of which will be used in the social feature extraction. After filling out the questionnaire, each respondent was labeled according to the BFI assessment.

2.3 Data Preprocessing

Data preprocessing is done to process data into a more efficient format for use in the next process. At the initial stage, case folding is done to change all letters in the tweet to lowercase. Retweets, usernames, hashtags, links, and newlines will be removed through the removing punctuation process. The data is then separated into tokens, which are then stemmed by removing the prefix and suffix for each word. Stop word removal is also carried out to remove words that often appear, such as *'yang', 'di', 'dari', 'aku', 'dan', 'ke',* so that only words that are considered important will be produced. The final step is text normalization, which reduces the unique tokens and makes them into standard form. Table 1 contains examples of preprocessing data.

Table 1. Data preprocessing example

Process	Output
Raw Tweet	Alhamdulillah tugas hari ini selesai semua, bikin 2 video sama 1 rekaman audio, mantap sekali ngasih tugasnya bapak 🙂
Case folding	alhamdulillah tugas hari ini selesai semua, bikin 2 video sama 1 rekaman audio, mantap sekali ngasih tugasnya bapak 🙂
Removing punctuation	alhamdulillah tugas hari ini selesai semua bikin 2 video sama 1 rekaman audio mantap sekali ngasih tugasnya bapak
Stemming	alhamdulillah tugas hari ini selesai semua bikin 2 video sama 1 rekam audio mantap sekali ngasih tugas bapak
Tokenizing	'alhamdulillah', 'tugas', 'hari', 'ini', 'selesai', 'semua', 'bikin', '2', 'video', 'sama', '1', 'rekam', 'audio', 'mantap', 'sekali', 'ngasih', 'tugas', 'bapak'

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

Removing stop word	'alhamdulillah', 'tugas', 'selesai', 'bikin', '2', 'video', '1', 'rekam', 'audio', 'mantap', 'ngasih', 'tugas'
Text normalization	'alhamdulillah', 'tugas', 'selesai', 'buat', '2', 'video', '1', 'rekam', 'audio', 'mantap', 'memberi', 'tugas'

2.4 Term Weighting TF-IDF and PCA

In this study, Term Frequency-Inverse Document Frequency (TF-IDF) is used to calculate the weight of each word in the document. The ratio of word occurrences in a document is measured by TF, while the informativeness of terms is measured by IDF [17]. TF-IDF can be calculated using equation (1).

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

Due to the enormous data dimensions produced by the TF-IDF method, this study applies Principle Component Analysis (PCA) to compress the dimensions of big TF-IDF data sets into smaller data sets. By discarding minor components, PCA effectively reduces the number of features.

2.5 Sentiment Feature and Emotion Feature

SentiStrength Indonesia is used to analyze word sentiment for sentiment feature extraction. SentiStrength uses a positive-negative scale. The strength of positive sentiment is worth 1 (not positive) to 5 (very positive), while negative sentiment is worth -1 (not negative) to -5 (very negative). Emotion feature extraction was used to analyze the emotions in words using the EmoLex dictionary, which contains eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) and two sentiments (positive and negative) [18].

2.6 Part-of-speech Tag

Part-of-speech (POS) tagging is used to classify words in a text, such as verbs, adjectives, nouns, and others. The Indonesian language corpus was used for POS tagging in this study. Table 2 describes the tagset used for POS Tagging.

Table 2. POS tag description [19]

No	Tag	Detail
1	ADV	Adverbs
2	CC	Coordinating conjunction
3	DT	Determiner
4	FW	Foreign word
5	IN	Preposition
6	JJ	Adjective
7	NEG	Negation
8	NN	Noun
9	NNP	Proper noun
10	NUM	Number
11	PR	Pronoun
12	RP	Particle
13	SC	Subordinating conjunction
14	SYM	Symbols and punctuation
15	UH	Interjection
16	VB	Verb
17	ADJP	Adjective phrase
18	DP	Date phrase
19	NP	Noun phrase
20	NUMP	Number phrase
21	VP	Verb phrase

2.7 Classification

In this study, the multi-class classification was carried out using the AdaBoost method. This method combines several weak learners into a strong classifier, improving the model's performance [20]. The base learner in the AdaBoost classification can use any method [20], but this study will use the default base learner, the Decision Tree. In the first iteration, each sample is given the same weight, 1/N, and the entire classification error is evaluated. The correct classification will be given a lower weight in the next iteration, while the incorrect classification will be given a higher weight. When the iteration is finished, the weight will be calculated automatically for each classification in each iteration based on the error rate, resulting in a strong classifier [20], [21]. The AdaBoost algorithm works by assigning weights to the training data and selecting the classification with the lowest error weight, resulting in a strong classifier built as a linear combination of weak learners [20]. AdaBoost, in general, can be calculated using equation (2), where F(x) is a strong classifier, α is a weight, and f(x) is a weak classifier.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \cdots + \alpha_n f_n(x) \quad (2)$$

AdaBoost is considered to have good predictive performance in several problems, and it is easy to determine the hyper-parameters [22]. The AdaBoost classification has three important parameters: base_estimator, n_estimator, and learning_rate [20]. Table 3 shows an explanation for each parameter as well as the default values.

Table 3. AdaBoost parameters [20], [21]

Parameter	Detail	Default
base_estimator	Base learner to train the model	Decision Tree Classifier (max_depth=1)
n_estimator	Number of models to be trained iteratively	50
learning_rate	Weights applied to each classification in iteration	1.0

2.8 Evaluation

A confusion matrix is used to evaluate the performance of the classification results. The truth table in Table 4 shows four terms in the representation of the confusion matrix results: TN (True Negative), TP (True Positive), FN (False Negative), and FP (False Positive).

Table 4. Confusion matrix [14]

Actual Result	Predicted Result	
	True	False
True	TP (correct result)	FP (unexpected result)
False	FN (missing result)	TN (correct absence of result)

The results of the confusion matrix representation are used in the following equation (3)-(6) to calculate the accuracy, precision, recall, and F1 scores.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\ Score = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (6)$$

Accuracy is defined as the ratio of correct predictions to total predictions. However, if the data is unbalanced and biased, the evaluation metrics fail to capture the classifier's effectiveness. Therefore, other metrics, such as precision and recall, are required to calculate the evaluation. The classification is better if the precision and recall values are close to one. Therefore, the F1-Score value is required to account for precision and recall. The F1-Score represents the average harmonic precision and recall. F1-Score has the best value of 1 and the worst value of 0 [23].

3. RESULT AND DISCUSSION

In this study, an Indonesian dataset was used, which was derived from Twitter account statistic information and tweets from 277 Twitter user accounts that completed the BFI questionnaire. Figure 2 shows the distribution of respondents' personalities. According to the table, openness is the most common trait in this dataset, while extraversion is the least common. This test's scenario is based on preprocessed data and a combination of features. In the first scenario, all features are tested, including sentiment, emotion, social, PCA, and POS-tag features. Furthermore, the second scenario is being tested with three different features combination. All experiment was carried out with and without the stemming process.

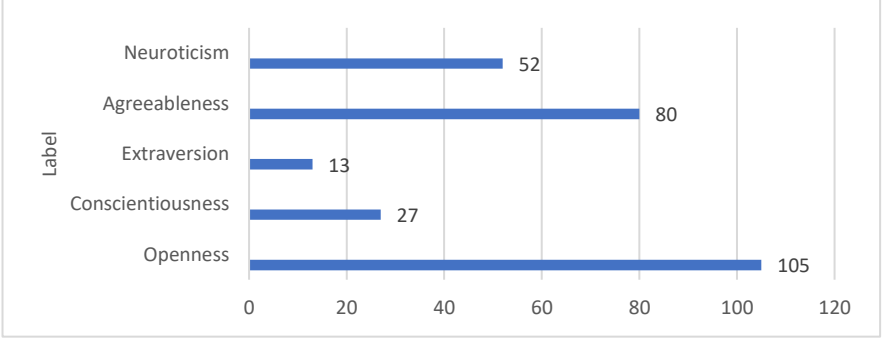

Figure 2. Data distribution label

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

In the first test scenario, all existing features are combined for testing. This test uses a comparison of 90:10, 80:20, 70:30, and 60:40 test data. Based on these tests, it was found that a 90:10 test data comparison produced the best accuracy. Figure 3 shows that with n_estimator 50, the test with stemming has the highest accuracy of 46.43%. At the same time, the test without stemming has a 7.14% higher accuracy than with the stemming process, which has a 53.57% accuracy with n_estimator 250. This is due to the stemming process of removing prefixes and suffixes from each word, resulting in fewer word variants. For example, "seminggu" and "Minggu" share the same root, "Minggu". However, the two words have distinct meanings. "seminggu" refers to 7 days, whereas "Minggu" refers to the day's name.

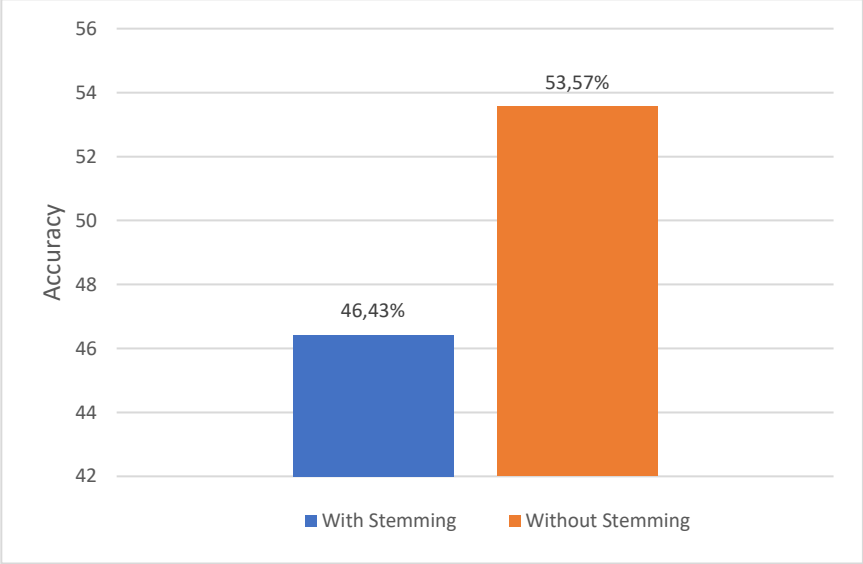

Figure 3. Accuracy results of all feature

Table 5 is a confusion matrix table showing the prediction and actual results from the highest accuracy, 53.57%. The table shows that there are still some incorrect predictions. Due to a lack of data, the extraversion trait is difficult to predict, whereas the openness trait is well predicted. As a result, the extraversion label has a low-performance value. Table 6 shows precision, recall, and f1-score values for each label.

Table 5. Confusion matrix of prediction results from the first scenario

Actual Result	Predicted Result				
	Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	8	1	0	2	1
Conscientiousness	0	1	0	0	1
Extraversion	0	0	0	0	1
Agreeableness	1	2	1	3	2
Neuroticism	1	0	0	0	3

Table 6. Precision, Recall, and F1-Score results

Label	Precision	Recall	F1-Score
Openness	80%	67%	73%
Conscientiousness	25%	50%	33%
Extraversion	0%	0%	0%
Agreeableness	60%	33%	43%
Neuroticism	38%	75%	50%

The combination of the three features in the second scenario was tested with a n_estimator of 100 and a comparison 90:10 data test. The combinations of these three features are emotion, PCA, and POS-tag; emotion, PCA, and sentiment; emotion, PCA, and social; emotion, sentiment, and social; PCA, POS-tag, and sentiment; PCA, POS-tag, and social; PCA, sentiment, and social; and POS-tag, sentiment, and social. As in the first scenario, scenario testing was conducted with and without the stemming process. According to the test results shown in Figure 4, the eight tests achieved high accuracy when not stemmed. The combination of emotion, PCA, and sentiment features produces the most significant difference in accuracy. The accuracy of the test with stemming is 21.43%, while the accuracy of the test without stemming is 50%, which is 28.57% higher. Based on all the tests carried out in the second scenario, the average accuracy without stemming is 30.80%, while with stemming is 22.32%.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

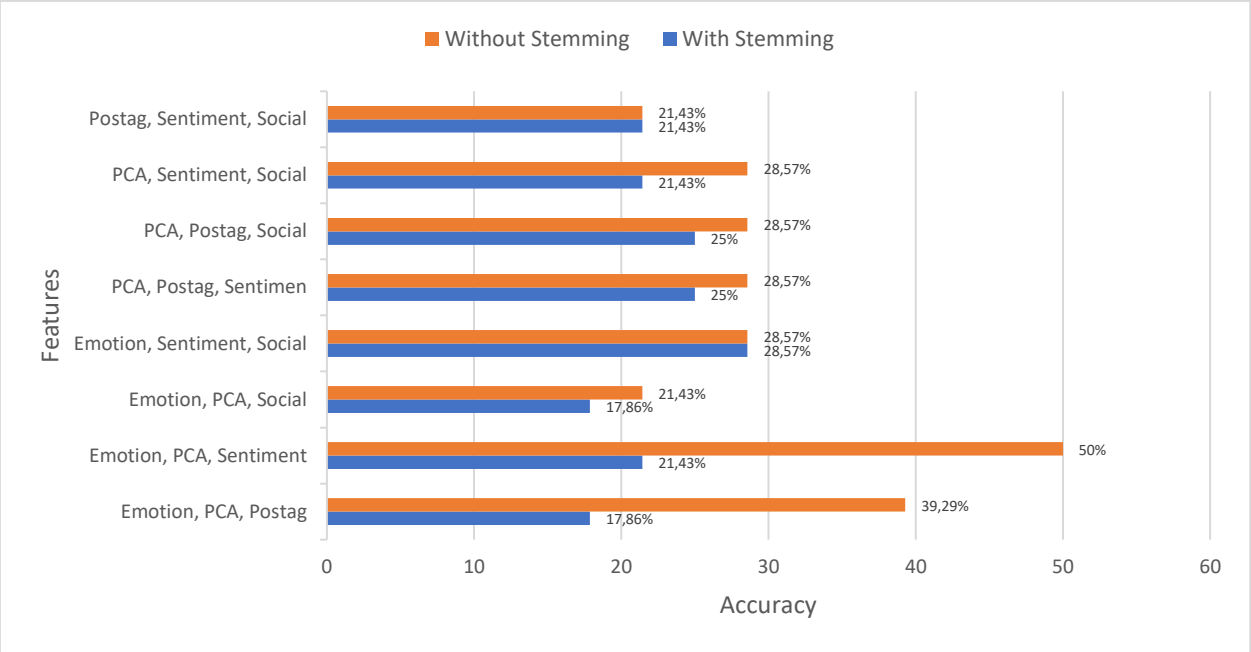

Figure 4. Accuracy results from a combination of 3 features

Table 7 shows the predicted and actual results with the highest accuracy of 50%. According to the table, the agreeableness trait is predicted well. From nine actual data, correctly predicted six personalities. Meanwhile, the conscientiousness trait incorrectly predicts personality. Table 8 shows the performance value of each label.

Table 7. Confusion matrix of prediction results from the second scenario

Actual Result	Predicted Result				
	Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	6	0	0	5	1
Conscientiousness	0	0	0	1	1
Extraversion	1	0	0	0	0
Agreeableness	2	1	0	6	0
Neuroticism	2	0	0	0	2

Table 8. Precision, Recall, and F1-Score results from second scenario

Label	Precision	Recall	F1-Score
Openness	55%	50%	52%
Conscientiousness	0%	0%	0%
Extraversion	100%	0%	0%
Agreeableness	50%	67%	57%
Neuroticism	50%	50%	50%

4. CONCLUSION

According to the study, the AdaBoost method uses several features such as sentiment, emotion, social, PCA, and POS-tag. Two test scenarios are carried out in this study, namely testing on all features and testing with a combination of three features. Both test scenarios were carried out with and without the stemming process. The best model built from the experiments is the first scenario, the AdaBoost model using all features without a stemming process, with an accuracy of 53.57%. Meanwhile, the best accuracy in the second scenario is the combination of emotion, PCA, and sentiment features without stemming with an accuracy of 50%. We also discovered that stemming affects the built model's performance when preprocessing data. Tests without stemming performed better than tests with stemming because the stemming process produced fewer word variants. More and better-balanced data will result in better performance. In the future, other feature extraction techniques, such as Word2Vec or Doc2Vec, might be used. In addition, more advanced models are expected to be used.

REFERENCES

[1] S. Kemp, "Digital 2022: Indonesia — DataReportal – Global Digital Insights," Feb. 15, 2022. https://datareportal.com/reports/digital-2022-indonesia (accessed May 07, 2022).

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 377−383
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1853

[2] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," IEEE Intelligent Systems, vol. 32, no. 2, pp. 74–79, Mar. 2017, doi: 10.1109/MIS.2017.23.
[3] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," Artificial Intelligence Review, vol. 53, no. 4, pp. 2313–2339, Apr. 2020, doi: 10.1007/s10462-019-09770-z.
[4] W. Maharani and V. Effendy, "Big five personality prediction based in Indonesian tweets using machine learning methods," International Journal of Electrical and Computer Engineering, vol. 12, no. 2, pp. 1973–1981, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1973-1981.
[5] R. Indira and W. Maharani, "Personality Detection on Social Media Twitter Using Long Short-Term Memory with Word2Vec," in 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), 2021, pp. 64–69. doi: 10.1109/COMNETSAT53002.2021.9530820.
[6] 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. doi: 10.1109/ICoDSA53588.2021.9617501.
[7] N. Abood, "Big Five Traits: A Critical Review," Gadjah Mada International Journal of Business, vol. 21, no. 2, pp. 159–186, 2019, doi: 10.22146/gamaijb.34931.
[8] Z. Mushtaq, S. Ashraf, and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," in Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020, Nov. 2020, pp. 1–5. doi: 10.1109/INMIC50486.2020.9318078.
[9] R. Moraes, L. L. Pinto, M. Pilankar, and P. Rane, "Personality Assessment Using Social Media for Hiring Candidates," in 2020 3rd International Conference on Communication Systems, Computing and IT Applications (CSCITA), 2020, pp. 192–197. doi: 10.1109/CSCITA47329.2020.9137818.
[10] A. Dwi Hartanto, E. Utami, S. Adi, S. Raharjo, M. Yusa, and A. Kamaludin, "Classifying User Personality Based on Media Social Posts Using Support Vector Machine Algorithm Based on DISC Approach," in 2020 2nd International Conference on Cybernetics and Intelligent System, ICORIS 2020, Oct. 2020, pp. 1–4. doi: 10.1109/ICORIS50180.2020.9320760.
[11] H. Setiawan and A. A. Wafi, "Classification of Personality Type Based on Twitter Data Using Machine Learning Techniques," in 2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020, Nov. 2020, pp. 94–98. doi: 10.1109/ICOIACT50329.2020.9332152.
[12] P. S. Dandannavar, S. R. Mangalwede, and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 62–65. doi: 10.1109/CTEMS.2018.8769304.
[13] A. v. Kunte and S. Panicker, "Using textual data for Personality Prediction: A Machine Learning Approach," in 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 529–533. doi: 10.1109/ISCON47742.2019.9036220.
[14] V. Kotu and B. Deshpande, "Classification," Data Science, pp. 65–163, Jan. 2019, doi: 10.1016/B978-0-12-814761-0.00004-6.
[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Ensemble learning," Data Mining, pp. 479–501, Jan. 2017, doi: 10.1016/B978-0-12-804291-5.00012-X.
[16] Y. Zhang et al., Research and Application of AdaBoost Algorithm Based on SVM. 2019.
[17] V. N. Gudivada, D. L. Rao, and A. R. Gudivada, "Information Retrieval: Concepts, Models, and Systems," Handbook of Statistics, vol. 38, pp. 331–401, Jan. 2018, doi: 10.1016/BS.HOST.2018.07.009.
[18] "Sentiment and emotion lexicons - National Research Council Canada." https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/sentiment-emotion-lexicons (accessed Jun. 23, 2022).
[19] "nlp-id · PyPI." https://pypi.org/project/nlp-id/ (accessed Jul. 05, 2022).
[20] D. Sarkar and V. Natarajan, "Ensemble Machine Learning Cookbook," 2019.
[21] "1.11. Ensemble methods — scikit-learn 1.1.1 documentation." https://scikit-learn.org/stable/modules/ensemble.html#adaboost (accessed Jun. 27, 2022).
[22] J. Moreira, A. C. P. de L. F. Carvalho, and T. Horváth, A general introduction to data analytics. John Wiley & Sons, Inc., 2018.
[23] G. Shobha and S. Rangaswamy, "Machine Learning," Handbook of Statistics, vol. 38, pp. 197–228, Jan. 2018, doi: 10.1016/BS.HOST.2018.07.004.