KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 3, No 6, Juni 2023, Hal 1199-1207
DOI 10.30865/klik.v3i6.933
https://djournals.com/klik

Big Five Personality Detection on Twitter Users Using Gradient Boosted Decision Tree Method

Adhie Rachmatulloh Sugiono, Warih Maharani*

Fakultas Informatika, S1 Informatika, Telkom University, Bandung, Indonesia
Email: [1]adhiersugiono@student.telkomuniversity.ac.id, [2,*]wmaharani@telkomuniversity.ac.id
Email Penulis Korespondensi: [2]wmaharani@telkomuniversity.ac.id

Abstract− In 2020, the Covid-19 virus caused a pandemic that made most people more active on social media, such as Twitter. Twitter has a tweet feature allows its users to send short messages about how they feel and think at that moment. Based on someone's tweet, we know their mindset, and it allows us to know the personality of that person. One model of personality is the Big Five personality. Big Five divides personality into five classes: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Several ways can be done to determine personality, such as taking a psychological test. However, it can take a long time and total concentration. Therefore, this study conducted a Big Five personality detection on Twitter users using the Gradient Boosted Decision Tree (GBDT) method. This study aims to obtain a high accuracy value by weighting it through the TF-IDF method and using sentiment and emotion features. This study utilized an Indonesian dataset that was collected through Twitter API. This study consists of two scenario tests, with the first scenario test being carried out with an imbalanced dataset and the second scenario test being carried out by applying the oversampling technique with SMOTE method to handle the imbalanced dataset. By applying SMOTE method, this study obtained a high accuracy with a value of 60.36%.

Keywords: Twitter; Tweet; Personality; Big Five; Gradient Boosted Decision Tree;

1. INTRODUCTION

Social media has become a common thing in our daily lives. Especially after the pandemic in 2020, many people have become active on social media to interact and socialize with others [1]. One of the most frequently used social media is Twitter [2]. Twitter has a feature called a tweet, where users can express themselves about what they are feeling and share information that the user wants to convey [3]. Based on these tweets, we can get information and an overview of the feelings and mindset of the user who wrote the tweet. Twitter is one of the social media that makes it easy for many users to see the latest or trending information through the tweet feature. In addition, tweets can make it easier for researchers to get datasets because tweets can be categorized as big data [4]. Several studies have used tweets from Twitter users as datasets, such as research conducted by Balakrishnan et al. [5], [6], which detects cyberbullying by associating the personality of the user to create a pattern used to detect cyberbullying.

Personality is an important thing that is owned by everyone [7]. Personality influences the mindset, nature, and feelings of a person. However, everyone has a different personality, so personality can make someone unique. One well-known method in the world of psychology that can divide a person's personality into five classes is called the Big Five personality. The five classes are openness, conscientiousness, extraversion, agreeableness, and neuroticism [8]. The Big Five personality is one of the most researched personality structure measurements and is considered good by many experts [8].

Personality is one of the critical factors that allows one to know oneself better. Knowing individual personality allows the person to know their weaknesses or deficiencies that can be improved. You can discover your personality in various ways, such as by taking a psychological test. However, working on a psychological test takes a long time and requires total concentration. So, suppose someone is not focused on doing the work. In that case, that person can fill in answers that are not following their personality, which results in personality test results, not in accordance with their actual personality. Today's technology has developed far. So it is no stranger if many things in everyday life are done digitally.

Previously, there has been research carried out to detect personality on Twitter users using various methods, such as research by Shantika Valerin Therik, who used the Decision Tree C4.5 method [9] in 2021. This research was conducted with a total of 549,151 tweets from 287 Twitter users. After that, TF-IDF and LIWC data were added and applied to the SMOTE method, which was tested by applying the hyperparameter tuning technique using Grid Search with social behavior as a baseline. This study achieved an accuracy of 62.06%, with an increase in accuracy of 17.24% from the baseline. After using the SMOTE method, the accuracy value increased to 76.92% resulting in an accuracy value of 32.1% from the baseline.

In 2022 there was also research by Alvinda Amalia Rizkita, who uses three different methods, which are Support Vector Machine (SVM), Naïve Bayes (NB), and Naïve Bayes-Support Vector Machine (NBSVM) [10]. This research is pretty similar to research by Shantika Valerin Therik [9] who used TF-IDF and LIWC methods with SMOTE method to oversample the data and used hyperparameter tuning technique with Grid Search and social behavior. The accuracy results obtained from this study were 82.41% using the Support Vector Machine algorithm, 71.42% accuracy using the Naïve Bayes algorithm, and 53.86% accuracy using the Naïve Bayes-Support Vector Machine algorithm.

In 2019, there was also a study by Roji Ellandi, who used k-Nearest Neighbor (kNN) method [11]. This research succeeded in getting data from 331,439 tweets from 137 accounts. The results obtained using the k Nearest Neighbor method get the best performance with a value of k = 9, which is 60.97%. Moreover, the Naïve Bayes-Support Vector Machine (NBSVM) got an accuracy of 53.86%.

In this research, the method used is the Gradient Boosted Decision Tree (GBDT). GBDT is a DT-based supervised learning technique with a boosting approach for predicting a value [12]. The boosting algorithm will identify weak attributes through a looping process that begins by giving each observation the same weight to find the features that cause predictions to be less accurate. After being successfully identified, these features will be weighted higher in the next iteration [13]. Finally, the results will be combined using voting for classification or mean for regression to produce the final model. GBDT is one of the most popular algorithmic methods in recent years, with many awards in machine learning and data mining competitions [14]. which uses the GBDT method to expand features with fastText on topic classification in 2022. This research obtains the highest accuracy results, namely 91.39%. The same year, Dhuhita Trias Maulidia [16], conducted feature expansion research with Word2vec for topic classification using the GBDT method. The highest accuracy result obtained in this study was 85.44%. Based on these studies, the GBDT method can potentially obtain high-accuracy results. Therefore, this study uses the GBDT method.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The design of the big five personality detection system for Twitter users using the Gradient Boosted Decision Tree (GBDT) method, which will be made for research, can be seen in Figure 1.

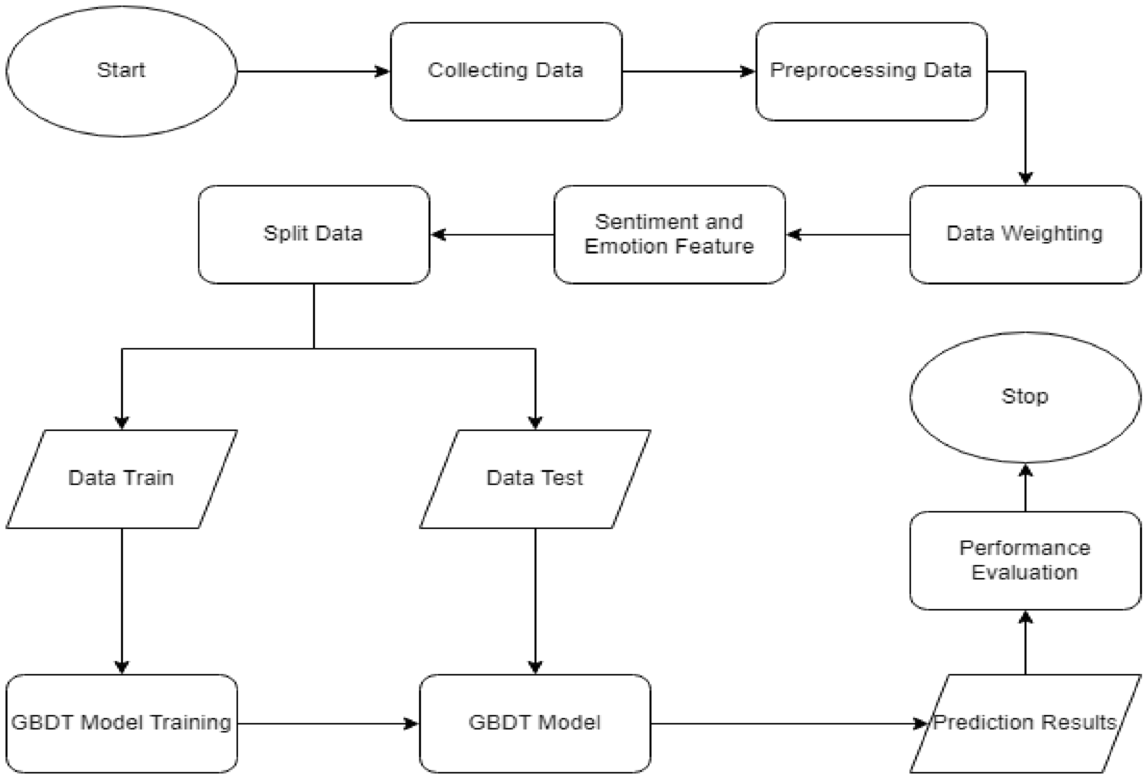

Figure 1. Flowchart

2.2 Data Collection

In this stage, the dataset used results from the data crawling process on tweets from Twitter users who have filled out the questionnaire. The questionnaire aims to determine the usernames and personalities of Twitter users based on the Big Five personalities. Crawling data is done using the Twitter API provided by Twitter. This crawling data connects the Twitter API with a Python script to get data results in CSV format.

2.3 Preprocessing

After the data labeling stage, the data will be preprocessed, which aims to transform unstructured data into structured and following modeling needs. The data preprocessing stage in this study can be seen in Table 1 below.

Table 1. Data Preprocessing

Process	Description
Raw Data	Raw data that has been crawled from the data collecting process

Data cleaning	Removing characters from the data and converting all the slang words to proper word
Case folding	Changing all the words to be lowercase
Tokenization	Slicing the data into string token
Stopwords removal	Removing words that are considered to be unimportant
Stemming	Removing prefixes and or suffixes from the data

2.4 Term Weighting

In this study, term weighting is carried out using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This process is carried out to determine the weight of the word from the data that has been obtained [17]. TF-IDF can be calculated using equation (**1**.

$$tfidf_{i,j} = tf_{i,j} \times idf_j \tag{1}$$

2.5 Sentiment Feature and Emotion Feature

This section aims to identify the feelings and moods corresponding to specific words in the user's tweet. The dictionary used is the NRC Word-Emotion Association Lexicon, created by Saif Mohammad [18]. The NRC Emotion Lexicon contains a collection of words and their connections with eight fundamental emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), which will be used in this study, and two sentiments (negative and positive).

2.6 Gradient Boosted Decision Tree Algorithm

As one of the ensemble models through a boosting approach, Gradient Boosted Decision Tree (GBDT) integrates several DTs into a robust classification. DT has a weakness, which is easy to overfit [19]. After carrying out the integration process on DT, GBDT will experience an increase in predictive performance. Thus, the GBDT learns from the mistakes made by the previous DT so that the next DT is combined with increasing accuracy and reducing errors in classification. The calculation formula is obtained from the sum of the previous DT models. The GBDT method can be calculated by the following equation (**2**.

$$F(x_i) = h_0(x_i) + h_1(x_i) + h_2(x_i) + \cdots + h_n(x_i) \tag{2}$$

In equation (**2**, there is a sum of weak to N DT models where *h* is the DT model from the weakest to the *n*th DT model in the boosting method, which aims to minimize the number of errors. At the same time, N is a parameter that determines how many boosting stages to perform.

2.7 Synthetic Minority Over-sampling Technique

Oversampling is a strategy utilized to overcome class imbalance in a dataset. Class imbalance occurs when the number of instances in different classes is significantly skewed, which can lead to biased model performance. In such cases, oversampling techniques focus on boosting the representation of the minority class by duplicating existing data instances from that class [20]. By generating additional samples, oversampling aims to balance the distribution and provide the model with a more comprehensive and representative training set. Synthetic Minority Over-sampling Technique (SMOTE) is one of the oversampling methods. SMOTE will be using the k nearest neighbors from the minority class to generate additional training data by selecting sample from each minority class and create new instances [21].

2.8 Evaluation

In this last section, an evaluation of the performance of the GBDT method was carried out using the Confusion Matrix method. The Confusion Matrix can be seen in Table 2 below.

Table 2. Confusion Matrix

Confusion Matrix		**Actual Result**	
		Positive	**Negative**
Predicted Result	**Positive**	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

After obtaining the predicted value of the confusion matrix, the model's performance will be assessed based on the following factors that can be seen from equation (3), **Error! Reference source not found.**, (5), and (6).

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3}$$

$$precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$f1 - score = \frac{2 \times (precision \times recall)}{(recall + Tprecision)} \tag{6}$$

Accuracy refers to the proportion of accurate predictions out of all predictions made. However, accuracy may not effectively assess the classifier's performance when dealing with imbalanced and biased data. Hence, additional metrics like precision and recall become necessary for evaluation purposes. A classification is considered superior if both precision and recall values are closer to one. The F1-Score is also similar to precision and recall, F1-Score has an excellent value of 1 and a worst value of 0, providing a comprehensive evaluation measure [22].

3. RESULT AND DISCUSSION

This study utilizes an Indonesian dataset, which successfully gathered 275 Twitter users and approximately 233,750 collected tweets. The dataset was employed to collect data and assign labels to the Big Five personalities based on the questionnaire results. The distribution of the Big Five personality labeling data among Twitter users is presented below in Figure 2, providing insights into the representation of each personality trait within the dataset.

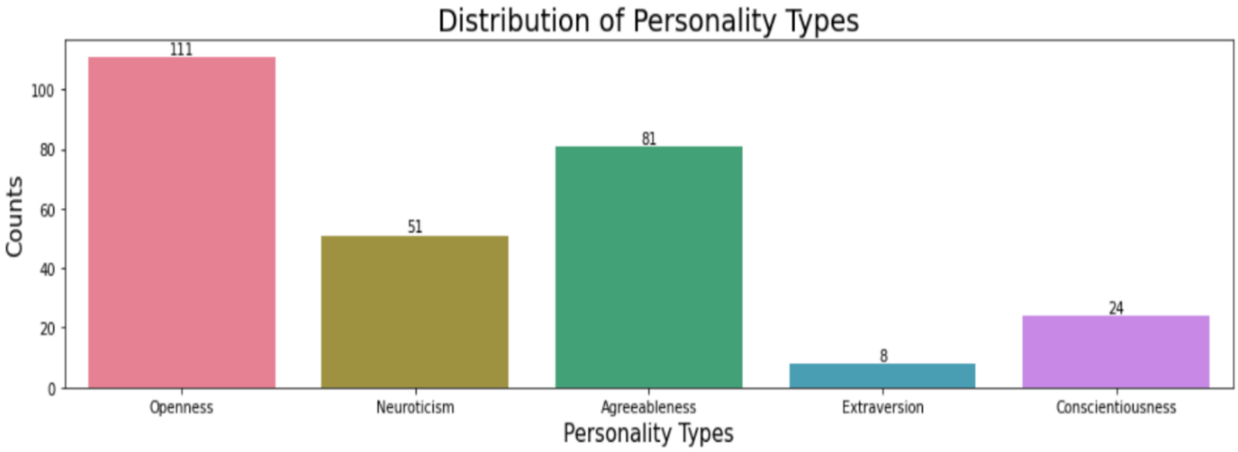

Figure 2. Data Distribution

In Figure 2, upon observing the collected dataset from the data collection process, it can be seen that the dataset is imbalanced. Out of the 275 Twitter users, there are 111 individuals with Openness personality, 24 with Conscientiousness personality, 8 with Extraversion personality, 81 with Agreeableness personality, and 51 with Neuroticism personality. Notably, there is a significant disparity between the number of individuals with Openness and Extraversion personalities, with a difference of 103 individuals favoring Openness. This imbalance in the dataset distribution highlights the need to address potential challenges associated with imbalanced classes during the subsequent analysis and classification tasks.

Table 3. Data preprocessing result

Process	Output
Raw Data	Gw yakin hari ini pasti membawa kebahagiaan!!
Data cleaning	Saya yakin hari ini pasti membawa kebahagiaan
Case folding	saya yakin hari ini pasti membawa kebahagiaan
Tokenization	[‘saya’, ‘yakin’, ‘hari’, ‘ini’, ‘pasti’, ‘membawa’, ‘kebahagiaan’]
Stopwords removal	[‘saya’, ‘yakin’, ‘hari’, ‘pasti’, ‘membawa’, ‘kebahagiaan’]
Stemming	[‘saya’, ‘yakin’, ‘hari’, ‘pasti’, ‘bawa’, ‘bahagia’]

Following the preprocessing phase, the dataset undergoes a weighting process utilizing the TF-IDF method. After that, sentiment and emotion calculations are performed on the TF-IDF weighted results, capturing each word's nuanced sentiment and emotional aspects. These calculations are conducted to extract meaningful features from the data. Table 4 illustrates the outcomes of the feature extraction process conducted on each Twitter user, showcasing the corresponding labels assigned to the Big Five personalities of the Twitter users. The table provides insights into the extracted features and their alignment with the individuals' personality traits.

Table 4. Feature Extraction Results

Username	P	N	A1	A2	D	F	J	S1	S2	T	Label
a2lir	3.92	1.95	1.11	0.86	0.92	1.59	3.38	1.61	0.86	2.59	1

abcdenjiiii	3.68	1.58	1.38	1.93	0.61	1.02	2.49	1.34	1.04	1.56	1
achadianrani	3.92	1.86	1.90	1.86	1.02	1.76	2.92	1.66	2.40	2.12	5
adamumemo	2.40	0.96	0.68	1.49	0.28	0.82	1.73	0.58	0.54	1.47	5
…	…	…	…	…	…	…	…	…	…	…	…
zulfaanf	6.43	6.31	2.17	5.95	1.58	3.32	6.29	3.46	3.06	5.02	4

In Table 4, column "P" stands for positive and column "N" stands for negative, these two columns are the result of extracting sentiment feature., On the other hand, column "A1" stands for anger emotion, column "A2" stands for anticipation emotion, column "D" stands for disgust emotion , column "F" stands for fear emotion, column "J" stands for joy emotion, column "S1" stands for sadness emotion, column "S2" stands for surprise emotion, and column "T" stands for trust emotion. All these columns are the result of extracting emotion feature. Column "Label" is the user's Big Five personality label. In column "Label", it can be seen that the value varies from 1 - 5 in order to help with the classification process. Value number 1 stands for openness personality, value number 2 stands for conscientiousness personality, value number 3 stands for extraversion personality, value number 4 stands for agreeableness personality, and value 5 stands for neuroticism personality.

In this study, classification will be separated into three parts: classification for sentiment feature, emotion feature, and all existing features. Training on the classification model is carried out with a ratio of data train and data test of 80:20.

3.1 First Scenario

As stated in the previous section in the first scenario, the dataset that has undergone the weighting process with the TF-IDF method and feature extraction with sentiment and emotion features was put to the classification process with the Gradient Boosted Decision Tree model. The parameters used in this test are learning rate = 0.05, n_estimators = 100, and max_depth = 5. The confusion matrix result of all of the test can be seen in below.

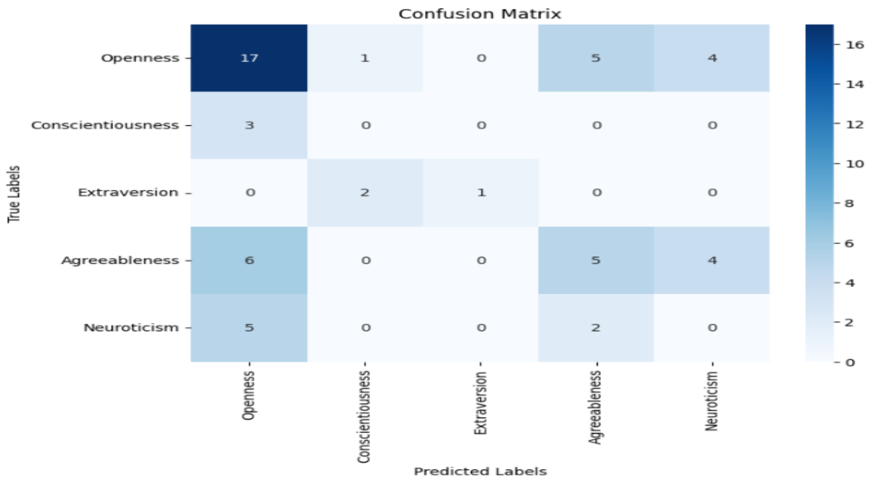

Figure 3. Confusion matrix of sentiment feature test

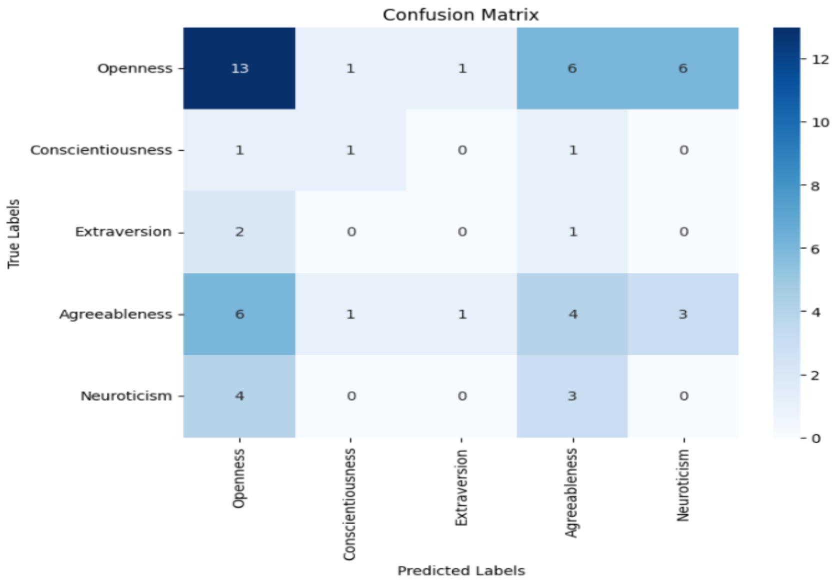

Figure 4. Confusion matrix of emotion feature test

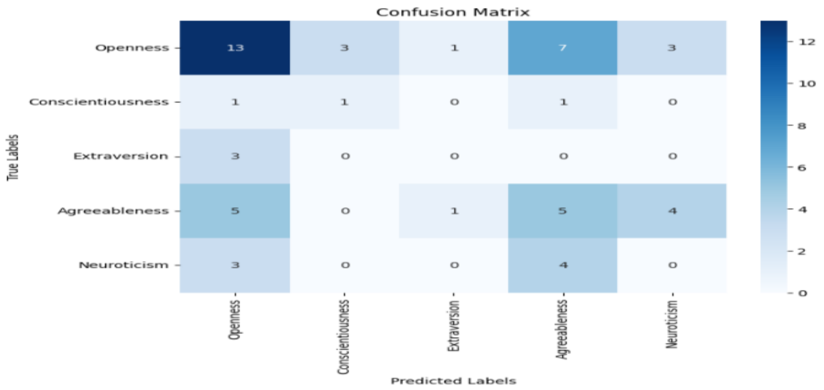

Figure 5. Confusion matrix of all feature test

Figure 3, Figure 4, and Figure 5 provides a visual representation of the performance and predictions of the classification model. It can be observed that from all three of the confusion matrixes, the openness label has the highest number of detections of true positives from the classification method. Based on the result of the confusion matrix, predicting the openness label is more accessible due to a more balanced data distribution and high amount of data which offers the model ample information, enabling the model to make a more accurate prediction. Whereas predicting extraversion labels proves challenging due to the low amount of data on label extraversion and imbalances in the dataset. Table 5 shows the result of accuracy, precision, recall, and f1-score of the first scenario.

Table 5. First Scenario Result

Test	Accuracy	Precision	Recall	F1-Score
Sentiment Feature	**41.82%**	**43.74%**	**41.82%**	**41.61%**
Emotion Feature	32.73%	33.64%	32.73%	33.17%
All feature	34.55%	34.91%	34.55%	34.63%

As mentioned earlier, when faced with imbalanced data, the accuracy metric might not be enough to evaluate the classifier's performance. Hence, the result of precision, recall, and f1-score can provide the classifier's overall performance and, in this case, the Gradient Boosted Decision Tree method. Based on the test results in Table 5, the sentiment feature has overall the best results out of the three with accuracy at 41.82%, precision at 43.74%, recall is the same as accuracy at 41.82%, and f1-score at 41.61%. On the other hand, the emotion feature and all feature tests have similar overall performance with emotion feature test managed to obtain the accuracy of 32.73%, precision of 33.64%, recall of 41.82%, and f1-score of 33.17%. And lastly, all feature test which obtained the accuracy of 34.55%, precision of 34.91%, recall of 34.55%, and f1-score of 34.63%.

3.2 Second Scenario

In the following test scenario, the imbalanced dataset was effectively handled by applying the oversampling technique, with the SMOTE (Synthetic Minority Over-sampling Technique) method. Synthetic samples were generated by implementing SMOTE to address the class imbalance issue and create a more balanced distribution within the dataset. The primary objective of employing SMOTE was to mitigate the adverse effects caused by the class imbalance and subsequently enhance the performance of the classification model. This approach provided the model with a more representative training set, enabling it to learn and generalize better, resulting in improved classification accuracy and overall model effectiveness.

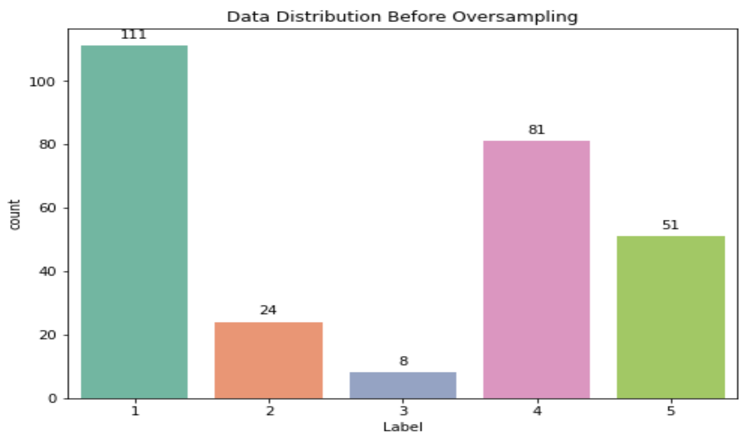

Figure 6. Data Distribution before Oversampling

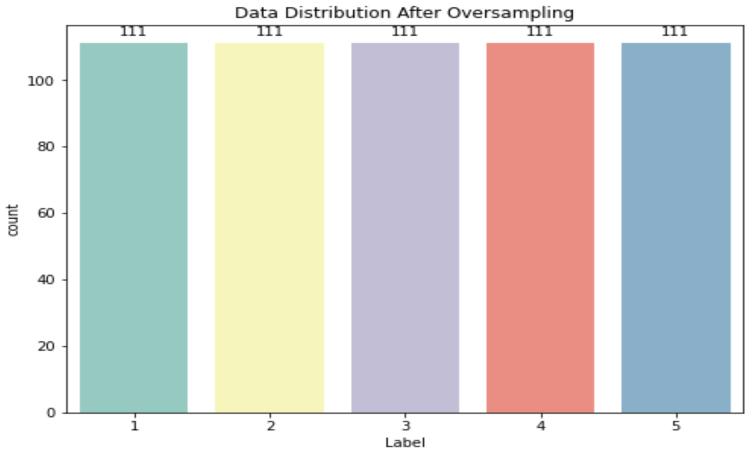

Figure 7. Data Distribution after Oversampling

Figure 6 and Figure 7 show the difference in the data distribution on the dataset after and before oversampling techniques were implemented. In result, the total amount of data significantly increased from the total data of 275 to 555. The equalization among all labels can be seen, with each label now having the same amount of data. For instance, label 3 representing extraversion, which initially had only eight instances in the dataset, has now been augmented to 111 instances. Similarly, label 2 indicating Conscientiousness, initially comprising 24 instances, has also been amplified to 111 instances. This balanced data distribution across all labels ensures fair representation and avoids bias towards any particular personality trait. Notably, the parameters that were used in the second scenario test remain the same in the first scenario, with learning rate = 0.05, n_estimators = 100, and max_depth = 5.

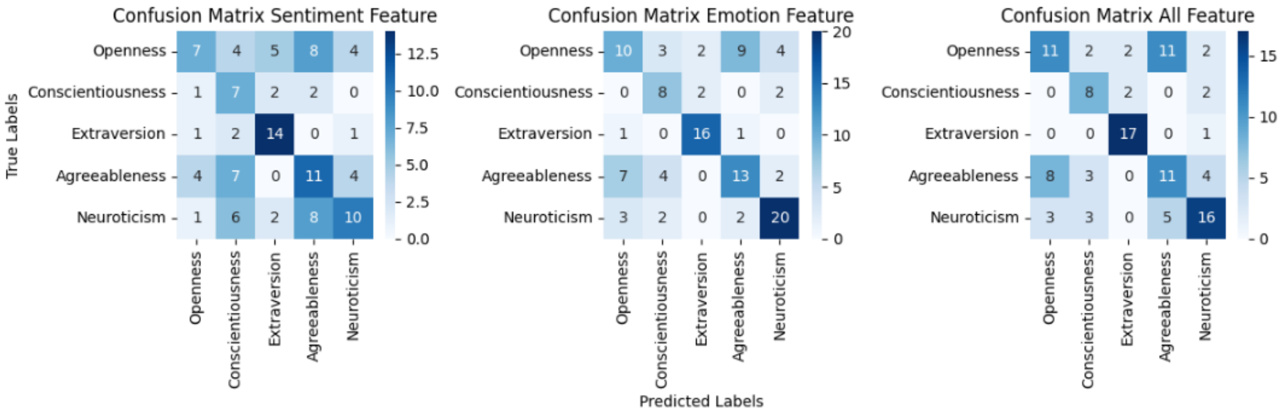

Figure 8. Confusion matrix of the second scenario

It can be observed from Figure 8 that the result of all three of the confusion matrices varies compared to the first scenario, where the openness label has the highest number of detections of true positives of the classification method. In this second scenario, the detection of the true positives is spread out with extraversion, agreeableness, and neuroticism, showing improvement in numbers compared to the first scenario's confusion matrixes.

Table 6. Second Scenario Result

Test	Accuracy	Precision	Recall	F1-Score
Sentiment Feature	44.14%	47.08%	44.14%	43.41%
Emotion Feature	**60.36%**	**59.63%**	**60.36%**	**60.78%**
All feature	56.76%	56.26%	56.76%	56.11%

According to Table 6, the performance of all tests of the classification process on the dataset that has undergone the oversampling process with the SMOTE method has increased in accuracy, precision, recall, and f1-score. This shows that the inclusion of SMOTE method has played a substantial role in enhancing the model's performance. The test that has the most significant improvement out of all the performance tests is the emotion feature test. In the emotion feature test, the accuracy has increased by 27.63% from the initial accuracy of 32.73% to 60.36%, precision has increased by 25.99% from the initial precision of 33.64% to 59.63%, the recall has the same improvement as accuracy, and f1-score that has improved by 27.61% from the initial f1-score of 33.17% to 60.78%. Furthermore, all feature test managed to gain an increase of accuracy by 22.21% from the initial accuracy of 34.44% to 56.76%, precision has increased by 21.35% from the initial precision of 34.91% to 56.26%, the recall has the same improvement as accuracy, and f1-score has increased by 21.48% from the initial f1-score of 34.63% to 56.11%. Lastly, the test that has the lowest improvement out of the three despite having the highest overall performance in the first scenario is the sentiment feature test. The sentiment feature test has an improvement in accuracy by 2.32% from the initial accuracy of 41.82% to 44.14%, precision increased by 3.34% from the initial precision of 43.74% to 47.08%, the recall has the same improvement as accuracy, and f1-score increased by 1.8% from the initial f1-score of 41.61% to 44.14%. The comparison of the overall performance from the first scenario and the second scenario can be seen in the next table below.

Table 7. Comparison of the first scenario and second scenario

Test	First Scenario				Second Scenario (+ SMOTE)			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Sentiment Feature	41.82%	43.74%	41.82%	41.61%	44.14%	47.08%	44.14%	43.41%
Emotion Feature	32.73%	33.64%	32.73%	33.17%	60.36%	59.63%	60.36%	60.78%
All Feature	34.55%	34.91%	34.55%	34.63%	56.76%	56.26%	56.76%	56.11%

4. CONCLUSION

This study used the Gradient Boosted Decision Tree method to classify Big Five personalities among Twitter users. An Indonesian dataset was managed to get collected. However, the dataset was proved imbalanced, with the distance between the highest and lowest data being 103 out of 275. To overcome an imbalanced dataset, the oversampling method SMOTE was employed to improve the overall performance of the classification model. The classification process is divided into three parts that involve only a specific feature to get classified: the sentiment feature, the emotion feature, and all features. This study consists of two scenarios, the first one is the classification process without handling the imbalanced dataset, and the second one is the classification with oversampling method SMOTE to handle the imbalanced dataset. It can be seen from the previous section that in the first scenario, the highest accuracy is on the sentiment feature tests with accuracy at 41.82%, precision at 43.74%; recall is the same as accuracy at 41.82%, and f1-score at 41.61%. On the other hand, in the second scenario test that involves oversampling by applying SMOTE method, all tests show improvement in accuracy, precision, recall, and f1-score. The highest accuracy is on the emotion feature which the accuracy has increased to 60.36% (+27.63%), precision has increased to 59.63% (+33.64%), the recall has increased to 60.36% (+27.63%), and f1-score that has improved to 60.78% (+27.61%). This shows that balancing on a large dataset can impact the classification process's performance. Future research should focus on refining the classification model using hyperparameter tuning techniques such as the Grid Search method. Future research should also look for alternative preprocessing that do not involve stemming since the process might alter the meaning of a certain word.

REFERENCES

[1] D. A. González-Padilla and L. Tortolero-Blanco, "Social media influence in the COVID-19 Pandemic," International braz j urol, vol. 46, no. 1, pp. 120–124, Jul. 2020, doi: 10.1590/S1677-5538.IBJU.2020.S121.

[2] P. E. Walck and E. W. Scripps, "Twitter: Social Communication in the Twitter Age," International Journal of Interactive Communication Systems and Technologies, vol. 3, no. 2, pp. 66–69.

[3] A. Y. O'Glasser, R. C. Jaffe, and M. Brooks, "To Tweet or Not to Tweet, That Is the Question," Semin Nephrol, vol. 40, no. 3, pp. 249–263, May 2020, doi: 10.1016/J.SEMNEPHROL.2020.04.003.

[4] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science, Proceeding, pp. 386–390, Oct. 2019, doi: 10.1109/ICECOS47637.2019.8984523.

[5] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," Comput Secur, vol. 90, p. 101710, Mar. 2020, doi: 10.1016/J.COSE.2019.101710.

[6] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using Big Five and Dark Triad features," Pers Individ Dif, vol. 141, pp. 252–257, Apr. 2019, doi: 10.1016/J.PAID.2019.01.024.

[7] M. Piotrowska, "The importance of personality characteristics and behavioral constraints for retirement saving," Econ Anal Policy, vol. 64, pp. 194–220, Dec. 2019, doi: 10.1016/J.EAP.2019.09.001.

[8] A. Oshio, K. Taku, M. Hirano, and G. Saeed, "Resilience and Big Five personality traits: A meta-analysis," Pers Individ Dif, vol. 127, pp. 54–60, Jun. 2018, doi: 10.1016/J.PAID.2018.01.048.

[9] S. V. THERIK, "Deteksi Kepribadian Big Five Pengguna Twitter Dengan Metode C4.5," 2021, Accessed: Nov. 27, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/172327/slug/deteksi-kepribadian-big-five-pengguna-twitter-dengan-metode-c4-5.html

[10] A. A. RIZKITA, "Prediksi Kepribadian Big Five Pengguna Media Sosial Twitter Dengan Metode Naive Bayes-Support Vector Machine (NBSVM)," 2022, Accessed: Nov. 27, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/176993/slug/prediksi-kepribadian-big-five-pengguna-media-sosial-twitter-dengan-metode-naive-bayes-support-vector-machine-nbsvm-.html

[11] R. ELLANDI, "Prediksi kepribadian Big Five dengan Term-Frequency Inverse Document Frequency Menggunakan Metode k-Nearest Neighbor pada Twitter," 2019, Accessed: Nov. 27, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/152263/slug/prediksi-kepribadian-big-five-dengan-term-frequency-inverse-document-frequency-menggunakan-metode-k-nearest-neighbor-pada-twitter.html

[12] H. Rao et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," Appl Soft Comput, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/J.ASOC.2018.10.036.

[13] M. M. Nishat et al., “Performance Investigation of Different Boosting Algorithms in Predicting Chronic Kidney Disease,” 2020 2nd International Conference on Sustainable Technologies for Industry 4.0, STI 2020, Dec. 2020, doi: 10.1109/STI50764.2020.9350440.
[14] Q. Li, Z. Wen, and B. He, “Practical Federated Gradient Boosting Decision Trees,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 4642–4649, Apr. 2020, doi: 10.1609/AAAI.V34I04.5895.
[15] R. A. YAHYA, “Ekspansi Fitur Dengan fastText pada Klasifikasi Topik di Twitter Menggunakan Metode Gradient Boosted Decision Tree,” 2022, Accessed: Dec. 05, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/179208/slug/ekspansi-fitur-dengan-fasttext-pada-klasifikasi-topik-di-twitter-menggunakan-metode-gradient-boosted-decision-tree.html
[16] D. T. MAULIDIA, “Ekspansi Fitur dengan Word2vec untuk Klasifikasi Topik dengan Gradient Boosted Decision Tree di Twitter,” 2022, Accessed: Dec. 05, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/179194/slug/ekspansi-fitur-dengan-word2vec-untuk-klasifikasi-topik-dengan-gradient-boosted-decision-tree-di-twitter.html
[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] S. M. Mohammad, “Practical and Ethical Considerations in the Effective use of Emotion and Sentiment Lexicons,” Nov. 2020, Accessed: Jun. 16, 2023. [Online]. Available: https://arxiv.org/abs/2011.03492v2
[19] W. Liang, S. Luo, G. Zhao, and H. Wu, “Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms,” Mathematics 2020, Vol. 8, Page 765, vol. 8, no. 5, p. 765, May 2020, doi: 10.3390/MATH8050765.
[20] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” 2020 11th International Conference on Information and Communication Systems, ICICS 2020, pp. 243–248, Apr. 2020, doi: 10.1109/ICICS49469.2020.239556.
[21] J. Brandt and E. Lanzén, “A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification,” 2021, Accessed: Jun. 18, 2023. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-432162
[22] A. ANGSAWENI, “Identifikasi Kepribadian Big Five pada Pengguna Twitter menggunakan Metode AdaBoost,” 2022, Accessed: Nov. 27, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/183749/slug/identifikasi-kepribadian-big-five-pada-pengguna-twitter-menggunakan-metode-adaboost.html