Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

Personality Detection on Twitter Social Media Using IndoBERT Method

Tri Ayu Syifa'ur Rohmah[1], Warih Maharani[2,*]

Informatics Department, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]triayusyifa@student.telkomuniversity.ac.id, [2,*]wmaharani@telkomuniversity.ac.id
Email Penulis Korespondensi: wmaharani@telkomuniversity.ac.id
Submitted:**20/07/2022**; Accepted:**18/08/2022**; Published**: 30/09/2022**

Abstract−Personality is the fundamental characteristic of human beings that makes humans unique. Because of these differences in human characteristics, personality becomes a benchmark for consideration in various recruitment processes. One way to predict personality is to apply an interview system or fill out questionnaires which often experience problems due to ineffectiveness in terms of time and cost. Results become inaccurate if prospective employees do not know themselves well. The big five personality method, divided into openness, conscientiousness, extraversion, agreeableness, and neuroticism, is widely used to predict personality. This study uses a deep learning method, IndoBERT, to detect personality based on five dimensions according to the big five personalities whose data is taken from Twitter tweets with crawling data. From the results of these studies, it is known that personality research using the IndoBERT method without a stemming process has a higher accuracy rate of 0.46.

Keywords: Personality; Big Five Personality; Tweet; IndoBERT

1. INTRODUCTION

Personality is one of the absolute requirements for humans to radiate their existence in the world, both in themselves and in their social life. Personality makes a person unique and different from one person to another, or so-called individual difference [1]. Because personality is a unique thing every human has, it is not uncommon for personality to be used as one of the requirements in various recruitments such as school registration and job registration through personality tests. This is because there is a relationship between personality and employee performance, as explained by Vianny Typeton Delima in his 2020 research on the impact of personality on employee performance at Batticaloa Teaching Hospital that personality is positively correlated with work performance. [2]. One way to find out the personality of prospective employees in a company is to use the services of the HRD or Human Resource Department. However, using manual systems such as interviews and filling out questionnaires related to personality becomes ineffective in terms of time and cost. In addition, some problems arise when the test is carried out using the system to answer questions provided on paper or other media becomes inaccurate due to the lack of ability of prospective employees to recognize themselves. Therefore, we need a system that can be used to detect personality to be more efficient in terms of time and cost.

The Big Five Personality method is a method for determining a person's personality, which is divided into five dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience, as stated by Raad and Mlacic in their book entitled The Big Five Personality Factors: The psycholexical approach. to personality in 2020 [3]. Nasyroh also conducted research in 2017 to determine the influence of personality using the big five methods on employee performance. From this research, it can be concluded that two dimensions have a significant relationship with employee performance, namely agreeableness, and neuroticism, while the other three dimensions have an insignificant relationship. [4]

As previously explained, there are several drawbacks to using the questionnaire system, so we need another method that can be used to detect a prospective employee, as has been done by Rahma Indira in 2021 by using the Long Short-Term Memory (LSTM) method and the Word2Vec feature. [5]. This study shows that personality detection can be seen from Twitter tweets and requires several steps to complete, including data collection, data preprocessing, feature extraction, classification, and evaluation. From the results of this study, it is known that the model with ten combined tweets produces a higher level of accuracy. This study requires two more significant amounts of data because it is mentioned that the accuracy is low. There is overfitting because the data is small. In other studies, to detect the Big Five Personality, the C4.5 method can be used as has been done by Shantika Valerin Therik in 2021 in her journal entitled Big Five Personality Detection of Twitter Users Using the C4.5 Method. [6]. In this study, the baseline from the accuracy of social behavior has an accuracy rate of 44.82%. Then it is compared with the addition of TF-IDF and LIWC data with SMOTE, both of which are carried out by applying the hyperparameter tuning technique. The addition of TF-IDF and LIWC has an accuracy of 62.06%, an increase of 17.24% from the baseline. Meanwhile, with the addition of TF-IDF, LIWC, and SMOTE data, the accuracy is 76.92%, with an increase of 32.1% from the baseline.

Research in the field of NLP or Natural Language Processing that uses Indonesian is not available quite a lot due to the lack of available data sources even though Indonesian is the language used by around 199 million people with the 11th most frequently used language in the world in 2022 [7]. Therefore, a pre-trained model was created called IndoBERT [8]. IndoBERT is a variation of BERT (Bidirectional Encoder Representations from Transformers) by following BERT-Base (uncased), which has 12 tasks with trains of more than 220M words which have primary resources on the Indonesian Wikipedia (74M words), Kompas, Tempo and Liputan6 articles (total 55M) and Indonesia Web Corpus (90M words). IndoBERT just published its paper in September 2020 with data collection from

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

Indo4B, which is used for pre-trained contextual monolingual training available to the public such as texts, blogs, news, and websites [8]. The use of the IndoBERT method in the study of detecting the use of abusive sentences in Indonesian texts conducted by Hadiyan Putra in 2021 also succeeded in obtaining the highest level of accuracy compared to the use of the KNN, SVM, and Naïve Bayes methods. [9]. In the first experiment, KNN and SVM were known only to be able to classify data to the main classes, and Naive Bayes was able to classify minority classes but had lower accuracy than IndoBERT, namely 0.3391 in class 1, 0.0863 in class 2, and 0.2368 in the third class.

This study will use the IndoBERT method to detect personality based on the Big Five Personality by considering features in the form of tweets and data labels from respondents, which will produce the types of dimensions of the respondent's personality. The selective use of Twitter as a social media chosen to obtain the Twitter dataset is a social media that has the second most users in the world after Facebook, and Indonesia is the fifth country with Twitter users in the world after Brazil [10]. The study results are expected to find out how to detect the personality of a user's tweet with the big five personalities and the IndoBERT method.

2. RESEARCH METHODOLOGY

2.1 Research Flow

This study has five main stages to detect personality from tweets: data collection, data preprocessing, classification and evaluation. Figure 1 is a system design flowchart used in the research.

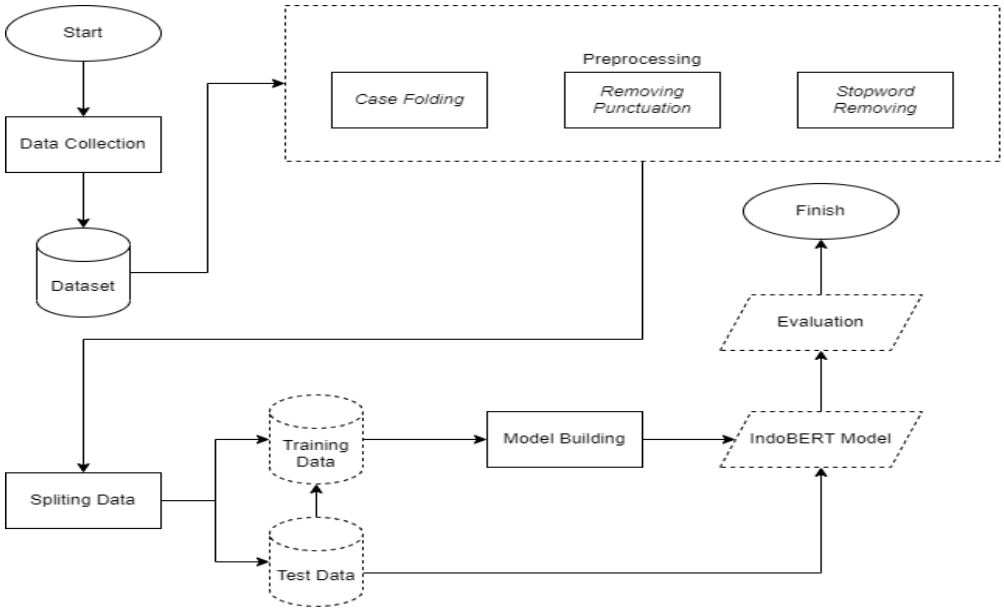

Figure 1. Flowchart

2.2 Data Collection

Data collection is done by asking respondents to fill out a questionnaire with questions about names, usernames, and personality related to the Big Five Inventory [11]. Then the username data from the questionnaire will be used for data crawling by utilizing the API [12] and personality questions will be used as the respondent's personality label. After crawling the data and weighing the questions from the questionnaire, the data is combined into one CSV file, a dataset from data collection consisting of names, usernames, tweets, and labels. From this data collection process, 258 data were obtained, with the number of each label being shown in table 1.

Table 1. Total account per label

Label	Total
Neuroticism	46
Agreeableness	82
Openness	101
Conscientiousness	24
Extraversion	5

2.3 Data Preprocessing

After crawling the data, the next step is to clean the crawled data so that the data is more structured and avoids noise. This study uses four preprocessing stages: case folding, removing punctuation, removing stopwords, and stemming. At this stage, case folding is used to change all data into lower case, removing punctuation is used to remove all

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

punctuation marks, removing stopwords is used to remove words that have no meaning, and stemming is used to change words into essential words. Examples of preprocessing data can be seen in table 2.

Table 2. Preprocessing Flow

Process Name	Before	After
Case folding	Operasi Perburuan Teroris di Kawasan TNLL Berakibat Objek Wisata Ditutup dan Hambat Program Konservasi.	operasi perburuan teroris di kawasan tnll berakibat objek wisata ditutup dan hambat program konservasi.
Removing punctuation	operasi perburuan teroris di kawasan tnll berakibat objek wisata ditutup dan hambat program konservasi.	operasi perburuan teroris di kawasan tnll berakibat objek wisata ditutup dan hambat program konservasi
Removing stopword	operasi perburuan teroris di kawasan tnll berakibat objek wisata ditutup dan hambat program konservasi	operasi perburuan teroris kawasan tnll berakibat objek wisata ditutup hambat program konservasi
Stemming	operasi perburuan teroris di kawasan tnll berakibat objek wisata ditutup dan hambat program konservasi	operasi buru teroris kawasan tnll akibat objek wisata tutup hambat program konservasi

2.4 IndoBERT Model Development

BERT (Bidirectional Encoder Representations from Transformers) is an open-source machine learning framework for natural language processing (NLP) developed by Google in 2018. Previously, language models could only read text input sequentially from left to right or vice versa. With the BERT method, the language model can read in both directions at once [13]. BERT is designed to help computers understand the meaning of ambiguous language in the text by using the surrounding text to establish context using two pre-trained BERT tasks, Mask ML and NSP [13]. IndoBERT is a transformed-based model using the BERT style following the BERT-Base (uncased) configuration. IndoBERT uses two mechanisms: an encoder used to read input and a decoder to generate predictions. IndoBERT has 12 hidden layers and has been trained with more than 220M words taken from three primary sources, namely the Indonesian language Wikipedia (74M), news articles from Kompas, Tempo, and Liputan6 (55M), and the Indonesian web corpus (90M) [14].

2.5 Evaluation

To evaluate the results of a classification study, an evaluation method called the confusion matrix is needed. In the confusion matrix, there are four values used, namely TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). The following is an explanation of the four values [15].

a. TP is used to store values whose predictions and actual values are correct.
b. TN is used to store values for which the predicted and actual values are wrong.
c. FP stores values that are predicted to be accurate but are false.
d. FN is used to store values that are predicted to be false but are true.

There are four indicators generated by the confusion matrix, including Accuracy, Precision, Recall, and F1-Score or F-Measure. In the calculation of this evaluation, still consider the values that have been described previously. The formula for the confusion matrix equation can be seen in formulas (1) – (4).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1 Score} = \frac{2 \text{ x Recall x Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

3. RESULT AND DISCUSSION

The next step is to create a classification model for the big five personalities using the indoBERT method. The IndoBERT method itself is a method that can have two mechanisms, namely encoder and decoder. The encoder is in charge of reading the incoming input, while the decoder is at the cost of generating predictions. In addition, because indoBERT can read text from two directions at once, it allows the model to learn the text from the surrounding environment.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

This research was conducted using 258 datasets from crawling Twitter. The scenario that will be tested lies in the preprocessing difference dataset. In the first scenario, the dataset used is a dataset that only goes through the case folding process and removes punctuation marks. Then in the second scenario, the dataset used is the dataset in the first scenario plus the process of removing punctuation marks. Then, in the third scenario, the dataset used is the dataset used in the second scenario plus the stemming process. In the third scenario, the training data and test data are separated by a 0.1 splitting data with 26 training data and 232 test data. Due to the same amount of data and split data used, the distribution of labels used in each scenario is also the same. The division labels in all datasets used in each scenario can be seen in figure 2.

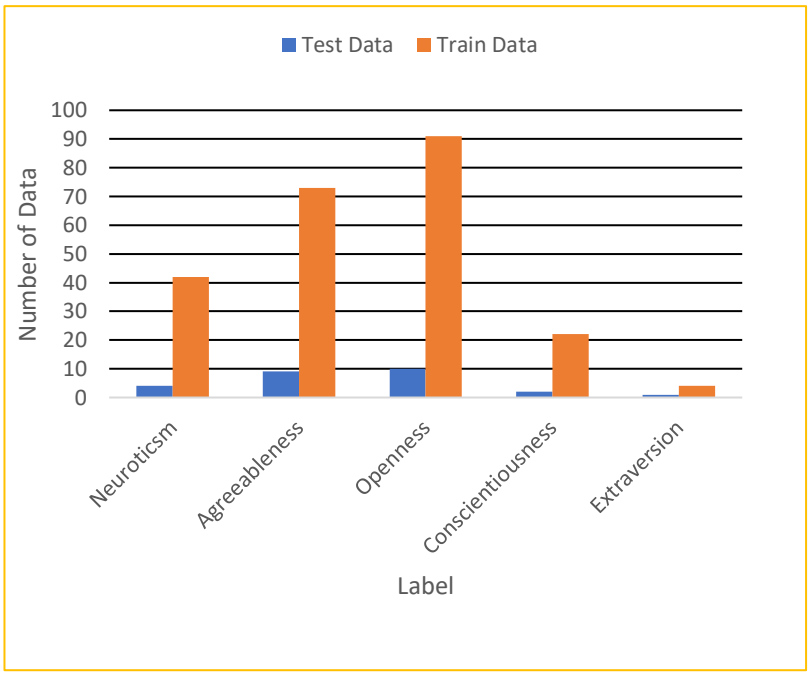

Figure 2. Number of Test Data and Train Data

From the results of the first scenario research, using a dataset that has gone through the case folding process and removing punctuation marks, an accuracy of 0.46 is obtained, as shown in Figure 3. In the first scenario, it has precision, recall, and F1-score results, as seen in table 3.

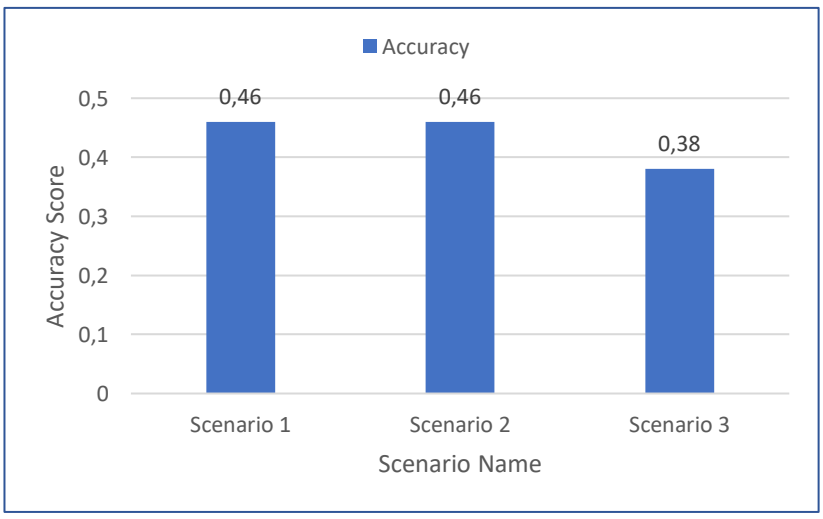

Figure 3. IndoBERT's Accuracy Result

Table 3. Precision, Recall, F1-Score Scenario 1

Label	Precision	Recall	F1-Score
Neuroticism	0	0	0
Agreeableness	0.44	0.44	0.44
Openness	0.80	0.47	0.59
Conscientiousness	0	0	0
Extraversion	0	0	0

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

Then in the second scenario, using a dataset that has gone through the process of case folding, removing punctuation, and removing stopwords, the accuracy is the same as in the first scenario, which is 0.46 as shown in Figure 3. In this second scenario, the results of precision, recall, and F1-score can be seen in table 4.

Table 4. Precision, Recall, F1-Score Scenario 2

Label	Precision	Recall	F1-Score
Neuroticism	0	0	0
Agreeableness	0.67	0.50	0.57
Openness	0.60	0.43	0.50
Conscientiousness	0	0	0
Extraversion	0	0	0

Finally, in the third test scenario using a dataset that has gone through case folding, removing punctuation, removing stopwords, and stemming, the accuracy is 0.38, as shown in Figure 3. This accuracy has a difference of 0.08 lower than the accuracy of the first and second test scenarios. This is also supported by precision, recall, and F1-score, which can be seen in table 5.

Table 5. Precision, Recall, F1-Score Scenario 3

Label	Precision	Recall	F1-Score
Neuroticism	0	0	0
Agreeableness	0.56	0.42	0.48
Openness	0.50	0.36	0.42
Conscientiousness	0	0	0
Extraversion	0	0	0

In addition to the three scenarios tested using the IndoBERT method, the study also tries to apply machine learning algorithms to compare the models used. In this comparison experiment, the first and second scenarios are used. From this comparison, the IndoBERT method has a higher accuracy level than the Logistic Regression, AdaBoost, and Decision Tree Classifier methods in the first and second test scenarios. The results of the accuracy of the comparison method can be seen in figure 4.

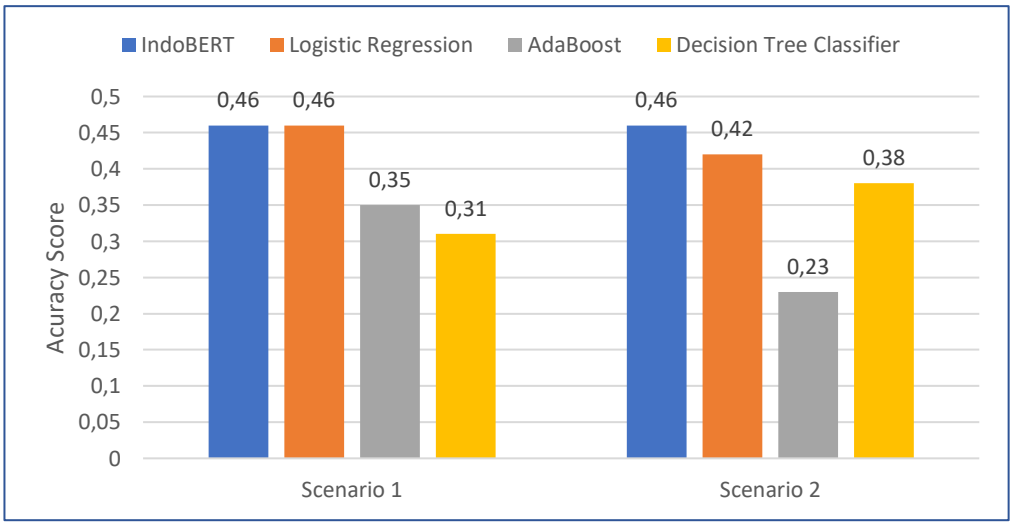

Figure 4. IndoBERT Comparison With Machine Learning Methods

4. CONCLUSION

In the research that has been done, it is known that the highest accuracy results are obtained in the first and second scenarios, namely 0.46. In contrast, the accuracy of the third scenario has a difference of 0.8. It can be analyzed that the stemming carried out in the third scenario dataset preprocessing has an important influence because the affixes can change the meaning of words or sentences. Then the number of precision, recall, and F1-Score results in all test scenarios on labels neuroticism, conscientiousness, extraversion are due to the lack of training data that fits the label, and after testing the test data, it turns out that the predicted label does not match the actual label. Increasing the number of datasets and balancing the amount of data on each label are needed for future research. Then related to the use of the deep learning method, namely IndoBERT to predict personality using the big five personalities, it can be considered for future research because it has been compared with other methods, and it is proven that IndoBERT has higher accuracy in the first and second scenarios.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 448−453
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1895

REFERENCES

[1] C. George Boeree, Personality Theories: Melacak Kepribadian Anda Bersama Psikolog Dunia. prismasophie, 2017.
[2] V. Delima, "Impact of Personality Traits on Employees' Job Performance in Batticaloa Teaching Hospital," Jun. 2020.
[3] B. Raad, The Big Five Personality Factors: The psycholexical approach to personality. 2000.
[4] M. NASYROH and R. Wikansari, "Hubungan antara Kepribadian (Big Five Personality Model) dengan Kinerja Karyawan," Ecopsy, vol. 4, no. 1, pp. 10–16, 2017.
[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] S. V. Therik and E. B. Setiawan, "DETEKSI KEPRIBADIAN BIG FIVE PENGGUNA TWITTER DENGAN METODE C4.5."
[7] "• Most spoken languages in the world | Statista." https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/ (accessed Jun. 17, 2022).
[8] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Dec. 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.85
[9] H. K. Putra, "Deteksi Penggunaan Kalimat Abusive Pada Teks Bahasa Indonesia Menggunakan Metode IndoBERT," e-Proceeding of Engineering, vol. 8, pp. 3028–3038, 2021.
[10] "Twitter: most users by country | Statista." https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed Jul. 16, 2022).
[11] O. P. John and S. Srivastava, "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives."
[12] "Twitter Developer Platform overview | Docs | Twitter Developer Platform." https://developer.twitter.com/en/docs/platform-overview (accessed Jul. 16, 2022).
[13] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Naacl-Hlt 2019, no. Mlm, 2018, Accessed: Jul. 07, 2022. [Online]. Available: https://github.com/tensorflow/tensor2tensor
[14] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[15] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," Aug. 2020.