

# Fine-tuning bidirectional encoder representations from transformers for the X social media personality detection

Selvi Fitria Khoerunnisa<sup>1</sup>, Bayu Surarso<sup>2</sup>, Retno Kusumaningrum<sup>3</sup>

<sup>1</sup>Master of Information System, School of Postgraduate Studies, Universitas Diponegoro, Semarang, Indonesia

<sup>2</sup>Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia

<sup>3</sup>Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia

## Article Info

### Article history:

Received May 18, 2024

Revised Apr 15, 2025

Accepted Jun 8, 2025

### Keywords:

BERT

Fine tuning

Myers-Briggs type indicator

Personality detection

Sequence learning

Twitter (X)

## ABSTRACT

Understanding personality traits can help individuals reach their full potential and has applications in various fields such as recruitment, advertising, and marketing. A widely used tool for assessing personality is Myers-Briggs type indicator (MBTI). Recent advancements in technology have allowed for research on how personalities can change based on social media use. Previous research used machine learning methods, deep learning methods, until transformers-based method. However, these previous approaches must be revised to require extensive data and a high computational load. Although transformer-based methods like bidirectional encoder representations from transformers (BERT) excel at understanding context, it still has limitations in capturing word order and stylistic variations. Therefore, this study proposed integrating fine-tuning BERT with recurrent neural networks (RNNs) consisting of vanilla RNN, long short-term memory (LSTM), and gated recurrent unit (GRU). This study also uses a BERT base fully connected layer as a comparison. The results show that the BERT base fully connected layer approach strategy has the best evaluation results in class extraversion/introversion (E/I) of 0.562 and class feeling/thinking (F/T) of 0.538. then, the BERT+LSTM approach strategy has the highest accuracy for the intuition/sensing (N/S) class of 0.543 and judging/perceiving (J/P) of 0.532.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Retno Kusumaningrum

Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro

St. Prof. Jacub Rais, Universitas Diponegoro, Tembalang, Semarang 50275, Indonesia

Email: retno@live.undip.ac.id

## 1. INTRODUCTION

Personality refers to an individual's typical behavioral, emotional, and cognitive patterns that are mostly displayed when engaging with others. Analyzing personality has attracted a lot of interest. This has resulted in various areas such as recruitment, advertising, and marketing. These fields study how personality influences different aspects to enhance the effectiveness of strategies. Understanding an individual's personality provides insights into their general characteristics or attitudes, there by maximizing their potential [1]. This knowledge also holds significance in the career field. Researchers in [2], [3] suggests that individuals perform better when their personality aligns with their job, making it easier to adapt to the work environment without requiring extensive physical abilities. Herr *et al.* [4] supports the idea that individuals work best in roles with lower physical demands.

The method to discover personality is by completing the questionnaire for the personality test. A widely used tool for assessing personality is Myers-Briggs type indicator (MBTI). MBTI is founded on Carl

Jung's personality theory [5] and employs four-factor model including: extraversion (E) or introversion (I), intuition (N) or sensing (S), feeling (F) or thinking (T), and judging (J) or perceiving (P). However, traditional personality tests have limitations as responses may not always be consistent due to random or haphazard answering, leading to variable predicted results. As a result, this study suggests using social media, particularly X (commonly known as Twitter), to detect personality traits. X provides a platform for users to interact, express thoughts and feelings through tweets, which indirectly reveal aspects of their personality [6].

Several studies have examined the use of X data for MBTI personality detection using natural language processing (NLP). The approaches for MBTI personality classification mostly involve binary classification. Frkovic *et al.* [7] demonstrated that the binary classification approach is more effective than the multiclass classification approach. The researchers initially used syntactic analysis features and n-gram characteristics with classical machine learning methods. The research in this field also utilized classical machine learning methods [8], [9], [10] in MBTI personality detection. However, classical machine learning methods challenges in feature extraction, as missing or incomplete features can lead to suboptimal outputs [11]. This limitation can be overcome by using deep learning methods, particularly sequential-based architectures like recurrent neural networks (RNNs). Since RNNs can automatically extract features and consider semantic dependencies, making it superior to classical machine learning methods. This has been proven in research by [12], [13].

Following the advancements in deep learning, the bidirectional encoder representations from transformers (BERT) is a noteworthy breakthrough in NLP. BERT employs the transformers framework, similarly to the generative pre-trained transformer (GPT) and large language model meta-AI (LLaMA). The architecture relies on self-attention mechanism, enabling every token to be evaluated alongside all other tokens at the same time [14]. As a result, attention weights between tokens are calculated, enabling the model to access information from all inputs. BERT employs the encoder structure of transformers, capturing context from both directions to comprehend text thoroughly, making it suitable for classification. Several previous studies have employed BERT for MBTI personality detection. For instance, research by [15], [16] utilized BERT as a word embedding approach. Additionally, researchers by [17], [18] conducted research by fine-tuning BERT to classify the four different MBTI dimensions. In contrast to employing BERT exclusively as a word embedding, which depend on vectors generated from pre-trained models, fine-tuning BERT involves retraining or transferring the knowledge from the BERT model utilizing a specific task dataset [19], and subsequently incorporating a fully connected layer that maps the representation results into the intended output. However, although BERT excels in understanding context, it still has limitations in capturing word order and stylistic variations. Therefore, this study proposed integrating fine-tuning BERT with RNNs to examine the impact of enhanced context modeling.

The proposed method aims to improve model performance by understanding the global context of words and capturing deeper meaning related to MBTI personality. The remainder of the paper is organized as follows: section 2 outlines the methodology, providing a comprehensive overview of the data, fine-tuning BERT for personality detection, and it suggest the integration of RNNs. Section 3 explain experiment setup and a discussion of those results, this study implication and future research. Finally, section 4 contains the conclusion of this study.

## 2. METHODOLOGY

The methodology starts with collecting data by scraping tweets from the X social media platform. Once the data is collected, it is preprocessed before being input into the model. The performance of the model is evaluated by an accuracy metric. Figure 1 depicts the phases of the proposed study, and the details of each phase will be explained in the subsequent sub-sections.

### 2.1. Dataset

The dataset contains tweet data from Indonesian X users who have shared their results of personality tests from various personality test services 16personalities.com. The dataset consists of tweet data from Indonesian X users who have shared their personality test results from the website 16personalities.com. The 50 most recent tweets from these users will be retrieved using the X application programming interface (API) library. At the same time, the shared personality test results will be used as data labels. After that, data labels will be divided into four categories: E/I, N/S, F/T, and J/P. 5120 data were successfully collected with balanced classes for all MBTI personality types.

Figure 2 shown the length of the data. It can be seen that some users have a token length of less than 200, which means that those users have less than or equal to 4 tokens in each of their tweets. On the other hand, some users have long tweets with more than 1,000 tokens, so these users have more than 20 tokens or

words per tweet. Meanwhile, most data distribution is around the 300 range. Thus, on average, one tweet contains 5 to 7 words. From Figure 2, it can be seen that the data is long and quite complex.

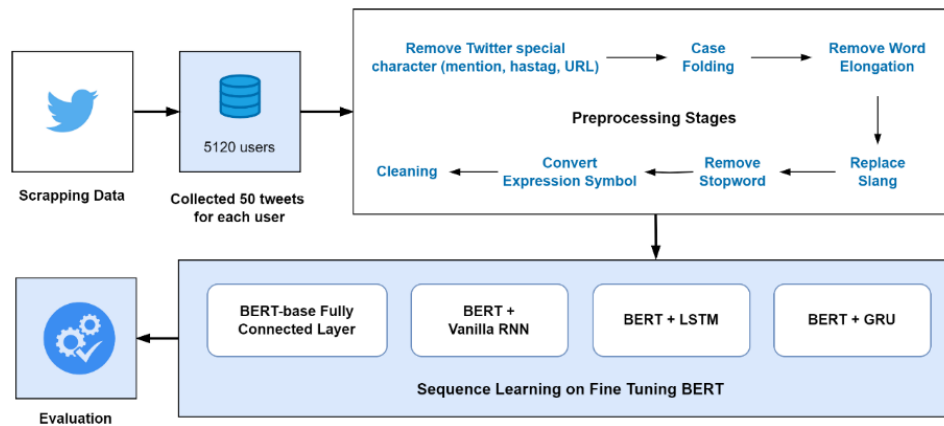


Figure 1. Research methodology

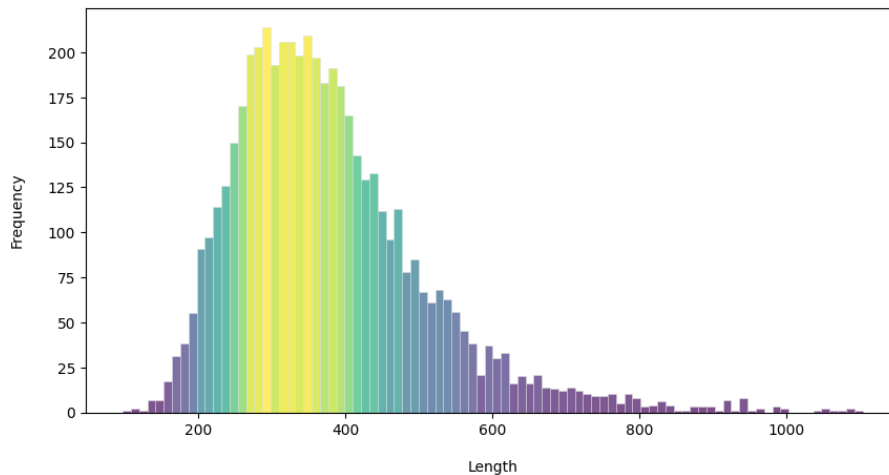


Figure 2. Distribution of length token data

## 2.2. Preprocessing

The X data resulting from crawling is data that has much noise and is unstructured, so it requires preprocessing to reduce the dimensions of the data during training [20], [21]. All the preprocessing steps carried out using the IndoNLP library. Preprocessing is essential because it allows the model identify distinct patterns in the data, enabling the analysis and classification of personalities. The preprocessing steps conducted in order are as follows:

- i) Remove X special characters, such as mentions, hashtags, and URLs, because they are meaningless [22].
- ii) Case folding, converted from all letters to lowercase. Additionally, all letters are converted to lowercase to ensure consistency and prevent significant variations in word vectors.
- iii) Convert emoji or emoticon to strings.
- iv) Data cleaning, which includes remove word elongation, slang words, and stop words. Stop words are frequently meaningless, so the model can focus only on essential words that contribute more to the text's meaning [23].

## 2.3. Fine tuning BERT

BERT is a versatile model that trains bidirectional representations of unlabeled text [14]. BERT has greatly improved NLP by effectively understanding context and semantics in text, analyzing information from both directions [24], [25]. BERT can be implemented using two approaches: feature-based and fine-

tuning. The feature-based model is preserved, and the output is a feature vector for the subsequent classification model [26], this process also known as word embedding process. The resulting vector will be sent through the specified classifier. In contrast, fine-tuning retrains the model to solve a more specific problem by modifying or adjusting the model architecture, showcasing BERT's adaptability to different tasks.

As explained before, this study implemented fine tuning BERT. First at all, the dataset required to fit the BERT input format, necessitating a tokenization process to align with the pre-trained model. It involved adding unique tokens to each sentence and converting the data into vectors. Fine-tuning was crucial to adjust all parameters precisely. Special symbols such as [CLS] and [SEP] were added at the beginning and end of each input, with padding used to ensure uniform data length [14]. The [CLS] token was included in the downstream task as an aggregate representation summarizing the input sequence information. To fine-tune the vector classification model related to [CLS], it was input into the encoder before adding a neural network layer above the output layer. Figure 3 shows the output layer can be integrated with other architectures. As illustrated in Figures 3(a) to 3(d), the integration of fine-tuning BERT with the RNNs utilized in this study. The selected RNNs include vanilla RNN, long short-term memory (LSTM), and gated recurrent unit (GRU). Apart from that, experiments will also be carried out with the base layer of BERT, with only a fully connected layer added as a comparison. A description of each RNN method is explained in the following subsection.

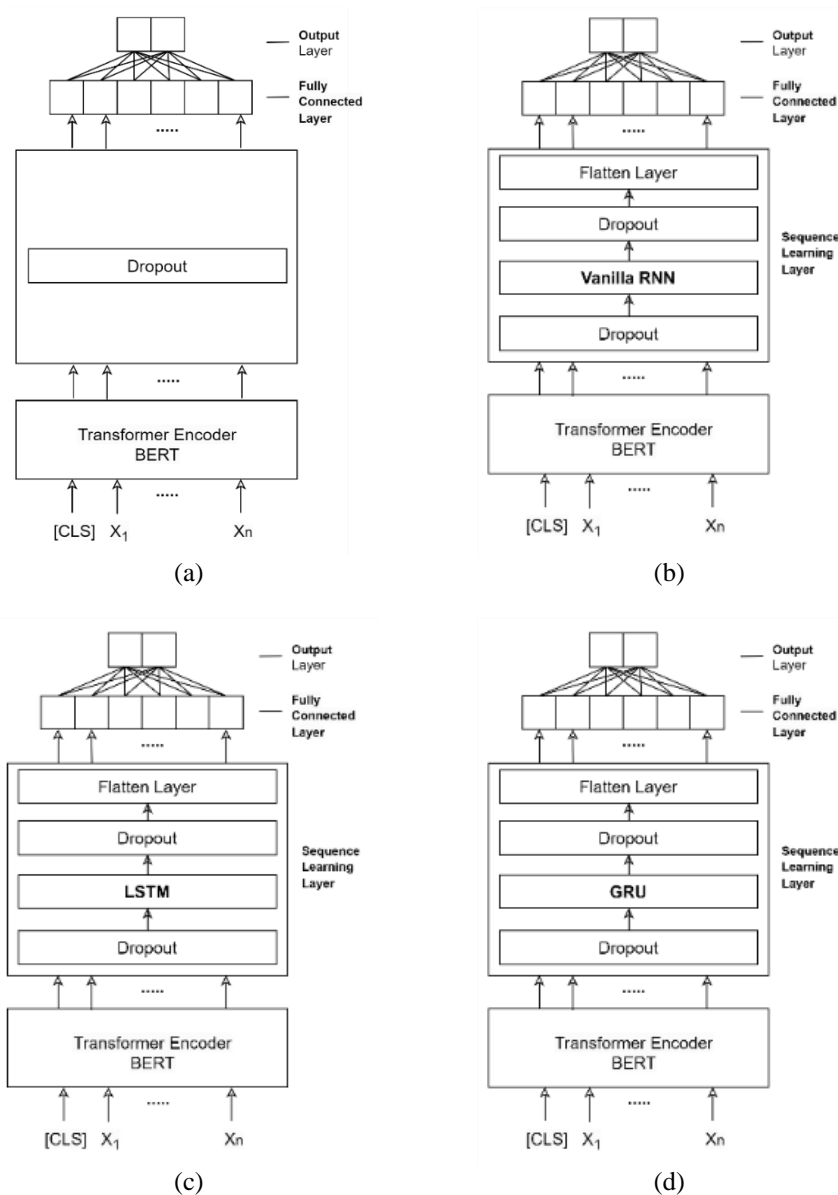


Figure 3. Architecture of fine tuning (a) BERT base fully connected layer, (b) BERT+vanilla RNN, (c) BERT+LSTM, and (d) BERT+GRU

The lower layers of BERT contain more generalized information, whereas the upper layers focus on specific tasks information. In fine-tuning, the classification task's model is initialized with pre-trained parameters, and then modified to fit the labeled data; several of the adjustments mentioned in detail are explained as follows. As depicted in Figures 3(a) to 3(d), the model architecture approach is BERT-based fully connected layer, this architecture uses a representation of  $X$  words provided by BERT (last hidden layer), which is directly connected to the fully connected layer without any hidden layers. The activation function is applied to the final layer for classification. BERT+(vanilla RNN, LSTM, or GRU) layer. The architecture uses word representations generated from the last hidden layer of BERT, which is connected to the RNN layer. It performs RNNs before connecting into the fully connected layer and output with an activation function. The architecture of this approach also uses dropout as a regulation technique.

#### 2.4. Vanilla recurrent neural network

RNNs is a type of neural network with loops. RNNs has memory and allows it to store existing information [27]. Vanilla RNN is a type of RNNs with only one iteration, meaning that vanilla RNN can only store information from one previous state.

#### 2.5. Long short-term memory

LSTM is a form of RNNs that addresses the limitation of vanilla RNN. If vanilla RNN can only store information from one previous state, LSTM can store information from all previous states and overcome long-term text dependency [28], [29]. Features of LSTM consist of memory cells and three gate units (input gate, forget gate, and output gate) to read, store, and update information.

#### 2.6. Gated recurrent unit

GRU is also an improved architecture over vanilla RNN and can handle long-term dependencies of text. The distinction between GRU and LSTM is found in the type of gate they possess. If LSTM has three gates, GRU only has update gates and reset gates [30], [31]. The update gate is a merging input gate and forget gate, while the reset gate sets the value from the previous state to continue to the next state.

### 3. RESULTS AND DISCUSSION

#### 3.1. Implementation

The implementation of this study used the pre-trained IndoBERTweet-base-uncased [32] form IndoLEM, a pre-trained language model for Indonesian language which have 409 M tokens. IndoBERTweet has been trained based on BERT-base-uncased by utilizing 12 attention heads, 12 hidden layers, feed-forward hidden layers, and 180 epochs [14]. After processing, the data is tokenized using the same approach as the pre-trained model. Tokenization not only separates punctuation and removes invalid characters but also prepares the data for analysis. The upper limit for sentence length is determined to be 512 tokens according to the distribution of token lengths in the dataset. If an input is shorter than this length, zeros are added to pad it; if it exceeds this limit, it is truncated to fit. The tokenized dataset is separated into three segments: training data, testing data, and validation data, following an 80:20 division for training and testing. The validation data consists of 10% of the training data. The model was trained with hyperparameters: batch size 16 and epoch 25. The dropout probability was set for all layers at 0.5. The AdamW optimizer utilizes a learning rate of  $1e-5$ . For evaluation, the study employs a confusion matrix along with accuracy metrics since the dataset is balanced. Subsequently, all experiments were conducted using the T4 GPU, a Turing architecture GPU intended to enhance the inference process of deep learning models.

#### 3.2. Result and discussion

This study used a binary approach to recognize MBTI personality, with four categories consist of E/I, N/S, F/T, and J/P. The method using fine-tuning BERT integrated with RNNs. Table 1 and Figure 4 compare the average of the results. Figure 4 demonstrates that fine-tuning BERT with a fully connected layer achieves the highest average accuracy with a value of 0.533. Then, it was followed by BERT+vanilla RNN of 0.523 and BERT+LSTM with a value of 0.518. Moreover, the last is BERT+GRU, with a value of 0.504. From Figure 4, shown that addition of RNNs architecture after BERT fine-tuning affects the model prediction results. BERT with a transformers base is used to understand the relationship between words in the text, so adding RNNs that aim to capture sequences will likely diminish the incremental value that LSTM might contribute. The possibility of ambiguous and overly short tweets also makes more difficult. As a result, integration with RNNs often fails to improve performance. Meanwhile, with fully connected layer, the BERT representation results are directly entered into a simple matrix and mapped to the desired label. Although, as seen in Table 1 BERT+LSTM fine-tuning has the best accuracy for the N/S class and the J/P class,

with values of 0.543 and 0.532. For the other two classes, the best accuracy is using the BERT base fully connected layer method, with values of 0.562 and 0.538.

Table 1. Experiment result

Fine tuning strategies	Label			
	E/I	N/S	F/T	J/P
BERT BASE FULLY CONNECTED LAYER	0.562	0.524	0.538	0.510
BERT+vanilla RNN	0.545	0.521	0.523	0.505
BERT+LSTM	0.518	0.543	0.480	0.532
BERT+GRU	0.516	0.516	0.480	0.505

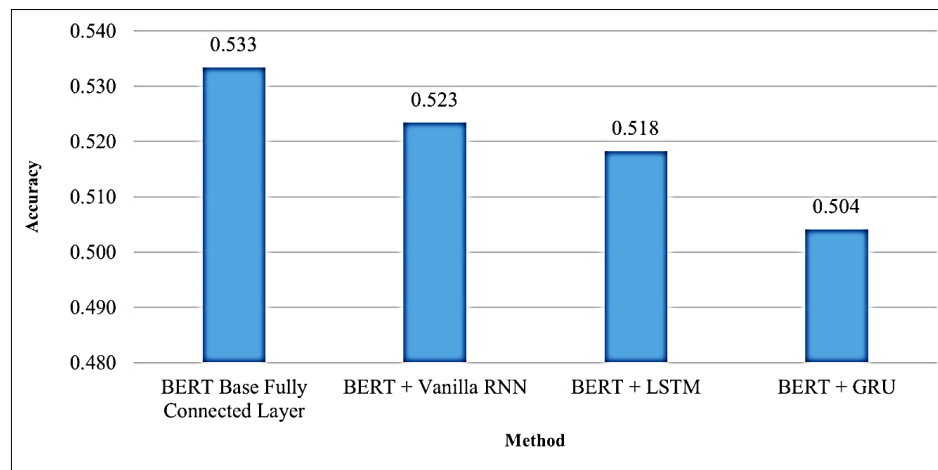


Figure 4. Average of experiment result

The E/I dimension refers to how a person directs energy and pays attention. This can be observed through their interaction style and emojis in tweets. The F/T dimension involves a preference for decision-making based on objective principles rather than personal feelings. When analyzing tweets, this dimension is reflected in the choice of words and tone, whether the author uses factual language or opts for a more empathetic and emotional tone. So, E/I and F/T do not require the addition of LSTM because they are sufficient to capture frequently appearing words and understand relationships directly without the need for a time sequence. Therefore, a fully connected layer is sufficient. In contrast, the N/S dimension focuses on a person's preference for acquiring information via the five senses or through patterns and possibilities. In tweets, this dimension pertains to how information is processed, whether the tweets convey reality using straightforward language or tend to be more conceptual and speculative. Lastly, the J/P dimension relates to an individual's lifestyle preference for either structure and definiteness or flexibility and adaptability. On the tweet, this dimension is reflected in their communication style, which may be either systematic and formal or spontaneous, exploratory, and often using abbreviations. It explains why BERT+LSTM is more effective for the N/S and J/P dimensions. It involves more complex sentence structures and a better understanding of the flow of language over time.

Compared with previous studies Datta *et al.* [33] was using BERT followed by random forest (RF) and extreme gradient boosting (XGB) as classifiers for MBTI personality detection on X data. Those studies reported the best accuracies of 0.441 and 0.424 for each classifier. In contrast, the proposed method in this study achieved a better accuracy of 0.562. The static embeddings produced by the BERT model in those studies may not be fully optimized for certain specific tasks, and tree-based classifiers do not inherently recognize sequential dependencies. In conclusion, the BERT-based fully connected layer consistently outperforms the others when it comes to directly capturing the meaning of text by identifying frequently occurring words. In contrast, the BERT+LSTM excels in understanding more complex sentence structures and managing temporal sequences. Therefore, the selection of the best model can be made by simply considering the highest accuracy for each label, eliminating the need for statistical testing. Although, this study investigates the impact of integrating BERT fine-tuning with RNNs by only applying fully connected layer. However, further comprehensive studies are necessary to ensure that the integration of

RNNs has a significantly impacts on evaluation results, particularly with regard to the suitability of the data type or structure used with the chosen model. This discovery provides conclusive evidence that this phenomenon is linked to changes in the classification part of fine-tuning, which can influence the model's performance and complexity.

#### 4. CONCLUSION

This study integrates fine-tuning BERT with RNNs for personality detection using X data. The RNNs used consist of vanilla RNN, LSTM, and GRU. In addition, this study also uses the BERT-based fully connected layer as a comparison. The results indicate that the BERT-based fully connected layer achieves the highest accuracy for class I/E and F/T, with scores of 0.562 and 0.538. This is attributed to its ability to effectively capture frequently appearing words and understand relationships directly without relying on a time sequence. On the other hand, the classes N/S and J/P require to understand more complex sentence structures and manage temporal sequences. Thus, the best-performing method for this class is BERT+LSTM, with accuracy scores of 0.543 and 0.532. This suggests that integration with RNNs can have a positively impact on certain classes. For future work, several improvements can be explored. First, optimizing hyperparameter settings is essential, as RNNs are highly sensitive to the hyperparameters used. Techniques such as Bayesian optimization can help identify optimal configurations. Additionally, it may be beneficial to alter the structure of the tweet data by segmenting each tweet for use with BERT. Maintaining the authenticity of each tweet's content is vital to avoid mixing information. Investigating methods like hierarchical BERT architectures could also be valuable in capturing the structural nuances of the text.

#### ACKNOWLEDGEMENTS

The authors would grateful to the anonymous reviewers for their insightful comments, constructive feedback, and helpful suggestions, which greatly enhanced the quality of this paper.

#### FUNDING INFORMATION

This study was funded by a grant from Directorate of Research, Technology and Community, Ministry of Education, Culture, Research, and Technology, Indonesia (PTM Grant 449A-70/UN7.D2/PP/VI/2023, 20 June 2023).

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Selvi Fitria Khoerunnisa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Bayu Surarso		✓		✓		✓		✓		✓		✓		✓
Retno Kusumaningrum	✓				✓					✓		✓	✓	✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

#### CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest associated with this paper.

#### DATA AVAILABILITY

Relevant data are available upon request, with appropriate permissions.

## REFERENCES





- [1] J. Carden, R. J. Jones, and J. Passmore, "Defining self-awareness in the context of adult development: a systematic literature review," *Journal of Management Education*, vol. 46, no. 1, pp. 140–177, Feb. 2022, doi: 10.1177/1052562921990065.
- [2] A. B. Bakker and M. V. Woerkm, "Strengths use in organizations: A positive approach of occupational health," *Canadian Psychology*, vol. 59, no. 1, pp. 38–46, 2018, doi: 10.1037/cap0000120.
- [3] A. E. M. van Vianen, "Person–environment fit: a review of its basic tenets," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 5, no. 1, pp. 75–101, 2018, doi: 10.1146/annurev-orgpsych-032117-104702.
- [4] R. M. Herr, A. E. M. V. Vianen, C. Bosle, and J. E. Fischer, "Personality type matters: Perceptions of job demands, job resources, and their associations with work engagement and mental health," *Current Psychology*, vol. 42, no. 4, pp. 2576–2590, 2023, doi: 10.1007/s12144-021-01517-w.
- [5] I. B. Myers, *Gifts differing understanding personality type*. Mountain View, California: Davies Black Publishing, 1995.
- [6] N. H. Jeremy and D. Suhartono, "Automatic personality prediction from Indonesian user on twitter using word embedding and neural networks," *Procedia Computer Science*, vol. 179, pp. 416–422, 2021, doi: 10.1016/j.procs.2021.01.024.
- [7] M. Frković, N. Čerkez, B. Vrdoljak, and S. Skansi, "Evaluation of structural hyperparameters for text classification with LSTM networks," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, pp. 145–150, doi: 10.23919/MIPRO48935.2020.9245216.
- [8] G. Ryan, P. Katarina, and D. Suhartono, "MBTI personality prediction using machine learning and smote for balancing data based on statement sentences," *Information*, vol. 14, no. 4, 2023, doi: 10.3390/info14040217.
- [9] N. Agarwal *et al.*, "Personality prediction and classification using Twitter data," *Social Networking and Computational Intelligence*, pp. 707–716, 2020, doi: 10.1007/978-981-15-2071-6\_59.
- [10] K. A. Nisha, U. Kulsum, S. Rahman, Md. F. Hossain, P. Chakraborty, and T. Choudhury, "A comparative analysis of machine learning approaches in personality prediction using MBTI," in *Computational Intelligence in Pattern Recognition*, Singapore: Springer, 2022, pp. 13–23, doi: 10.1007/978-981-16-2543-5\_2.
- [11] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews," *Procedia Computer Science*, vol. 179, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.
- [12] H. Naik, S. Dedhia, A. Dubbawar, M. Joshi, and V. Patil, "Myers Briggs type indicator (MBTI) - personality prediction using deep learning," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 2022, pp. 1–6, doi: 10.1109/ASIANCON55314.2022.9909077.
- [13] M. Maulidah and H. F. Pardede, "Prediction of Myers-Briggs type indicator personality using long short-term memory," *Jurnal Elektronika dan Telekomunikasi*, vol. 21, no. 2, pp. 104, 2021, doi: 10.14203/jet.v21.104-111.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4176.
- [15] H. Zhang, "MBTI personality prediction based on BERT classification," in *4th International Conference on Computer Science and Intelligent Communication (CSIC 2022)*, vol. 34, 2023, doi: 10.54097/hset.v34i.5497.
- [16] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Information Processing & Management*, vol. 58, no. 3, 2021, doi: 10.1016/j.ipm.2021.102532.
- [17] V. G. D. Santos and I. Paraboni, "Myers-Briggs personality classification from social media text using pre-trained language models," *JUCS - Journal of Universal Computer Science*, vol. 28, no. 4, pp. 378–395, 2022, doi: 10.3897/jucs.70941.
- [18] S. S. Keh and I.-T. Cheng, "Myers-Briggs personality classification and personality-specific language generation using pre-trained language models," *arXiv-Computer Science*, pp. 1–5, 2019, doi: 10.48550/arXiv.1907.06333.
- [19] A. N. Azhar, "Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis," *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*, 2020, doi: 10.1109/ICAICTA49861.2020.9428882.
- [20] L. Zhu and D. Luo, "A novel efficient and effective preprocessing algorithm for text classification," *Journal of Computer and Communications*, vol. 11, no. 03, pp. 1–14, 2023, doi: 10.4236/jcc.2023.113001.
- [21] T. H. Saputro and A. Hermawan, "The accuracy improvement of text mining classification on hospital review through the alteration in the preprocessing stage," *International Journal of Computer and Information Technology*, vol. 10, no. 4, pp. 2279–0764, 2021, doi: 10.24203/ijcit.v10i4.138.
- [22] B. A. H. Murshed, S. Mallappa, O. A. M. Ghaleb, and H. D. E. Al-ariqi, "Efficient Twitter data cleansing model for data analysis of the pandemic Tweets," in *Studies in Systems, Decision and Control*, vol. 348, 2021, pp. 93–114, doi: 10.1007/978-3-030-67716-9\_7.
- [23] A. L. Rio, M. Martin, A. Perera-Lluna, and R. Saidi, "Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-71450-8.
- [24] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The hybrid of BERT and deep learning models for Indonesian sentiment analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 591–602, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.
- [25] K. S. Rao, D. Valluru, S. Patnala, R. B. Devareddi, T. S. R. Krishna, and A. Sravani, "Phishing website detection using novel integration of BERT and XLNet with deep learning sequential models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, pp. 1273–1283, Nov. 2024, doi: 10.11591/ijeecs.v36.i2.pp1273-1283.
- [26] K. S. Nugroho, A. Y. Sukmadewa, D. W. H. Wuswilahaken, F. A. Bachtiar, and N. Yudistira, "BERT fine-tuning for sentiment analysis on Indonesian mobile apps reviews," in *SIET '21: Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, 2021, pp. 258–264, doi: 10.1145/3479645.3479679.
- [27] N. J. Johannesen, M. L. Kolhe, and M. Goodwin, "Comparing recurrent neural networks using principal component analysis for electrical load predictions," in *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2021, pp. 1–6, doi: 10.23919/SpliTech52315.2021.9566357.
- [28] M. A. Riza and N. Charibaldi, "Emotion detection in Twitter social media using long short-term memory (LSTM) and fast text," *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 3, no. 1, pp. 15–26, 2021, doi: 10.25139/ijair.v3i1.3827.
- [29] J. Shi, S. Wang, P. Qu, and J. Shao, "Time series prediction model using LSTM-transformer neural network for mine water inflow," *Scientific Reports*, vol. 14, 2024, doi: 10.1038/s41598-024-69418-z.
- [30] E. Ezhilarasi I and J. C. Clement, "GRU-SVM based threat detection in cognitive radio network," *Sensors*, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031326.







- [31] Y. Xiao, C. Zou, H. Chi, and R. Fang, "Boosted GRU model for short-term forecasting of wind power with feature-weighted principal component analysis," *Energy*, vol. 267, 2023, doi: 10.1016/j.energy.2022.126503.
- [32] 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*, Barcelona, Spain, Nov. 2020, pp. 757–770.
- [33] A. Datta, S. Chakraborty, and A. Mukherjee, "Personality detection and analysis using Twitter data," *arXiv-Computer Science*, pp. 1-9, 2023, doi: 10.48550/arXiv.2309.05497.

## BIOGRAPHIES OF AUTHORS







**Selvi Fitria Khoerunnisa**     received her B.S. degree (cumlaude) in informatics from Universitas Diponegoro, Semarang, Indonesia, in 2022. Now, she is postgraduate student at Master of Information System, School of Postgraduates Study, Universitas Diponegoro. She served as a Laboratory Assistant with the Department of Informatics, Universitas Diponegoro, from 2022 to now. She is currently a Research Assistant with the Laboratory of Intelligent Systems, Department of Informatics, Universitas Diponegoro. Her research interests include machine learning and natural language processing. She can be contacted at email: selvifkh@students.undip.ac.id.



**Bayu Surarso**     received his B.S. degree in mathematics from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 1987. His M.S. and Ph.D. degrees from Hiroshima University, Japan in 1995 and 1998, respectively. He is currently a lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro. Her research interests include algebra, combinatorics, and mathematical logic. He is also involved in expert system application used to identify issues and alternative solutions for secondary school students. He can be contacted at email: bayus@lecturer.undip.ac.id.



**Retno Kusumaningrum**     received her B.S. degree in mathematics from Universitas Diponegoro, Semarang, Indonesia, in 2003, and her M.S. and Ph.D. degrees from Universitas Indonesia, Depok, Indonesia in 2010 and 2014, respectively. She is currently a lecturer at the Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro. She also currently serves as the head of the Intelligent Systems Laboratory in Department of Informatics. Her research interests include natural language processing, machine learning, computer vision, pattern recognition, and topic modeling. She is a member of the IEEE Computational Intelligence Society, IEEE Computer Society, and ACM. Her awards and honors include the Sandwich-Like scholarship award from the Directorate General of Higher Education of Indonesia for visiting the School of System Engineering, University of Reading, Reading, U.K. as a student visitor in 2012, the Best Paper of the Second International Conference on Informatics and Computational Sciences in 2018, first place for Outstanding Lecturer-Universitas Diponegoro for the Science and Technology Category in 2019, and second place for the Best Paper Award of the Third International Symposium on Advanced Intelligent Informatics in 2020. She can be contacted at email: retno@live.undip.ac.id.