

Language model optimization for mental health question answering application

Fardan Zamakhsyari^{1,2}, Agung Fatwanto²

¹Information Technology, Sekolah Tinggi Teknologi Cahaya Surya, Kediri, Indonesia

²Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia

Article Info

Article history:

Received Aug 18, 2024

Revised Jun 5, 2025

Accepted Jun 30, 2025

Keywords:

Bidirectional encoder representations from transformers

IndoBERT

MBERT

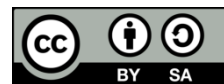
Natural language processing

Question answer

ABSTRACT

Question answering (QA) is a task in natural language processing (NLP) where the bidirectional encoder representations from transformers (BERT) language model has shown remarkable results. This research focuses on optimizing the IndoBERT and MBERT models for the QA task in the mental health domain, using a translated version of the Amod/mental_health_counseling_conversations dataset on Hugging Face. The optimization process involves fine-tuning IndoBERT and MBERT to enhance their performance, evaluated using BERTScore components: F1, recall, and precision. The results indicate that fine-tuning significantly boosts IndoBERT's performance, achieving an F1-BERTScore of 91.8%, a recall of 89.9%, and precision of 93.9%, marking a 28% improvement. For the model, M-BERT's fine-tuning results include an F1-BERTScore of 79.2%, recall of 73.4%, and precision of 86.2%, with only a 5% improvement. These findings underscore the importance of fine-tuning and using language-specific models like IndoBERT for specialized NLP tasks, demonstrating the potential to create more accurate and contextually relevant question-answering systems in the mental health domain.

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



Corresponding Author:

Agung Fatwanto

Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Kalijaga

Jl. Marsda Adisucipto, Yogyakarta 55281, Yogyakarta

Email: agung.fatwanto@uin-suka.ac.id

1. INTRODUCTION

Recent improvements to a family of neural models, known as bidirectional encoder representations from transformers (BERT), heavily rely on pre-training and have shown promise in a range of natural language processing (NLP) tasks, which include text categorization and dataset question-answering [1]. To learn word representations, BERT uses the transformer architecture, which consists of several encoder layers. Because it processes sequential input, like words in text, using an attention mechanism, the Transformer is regarded as remarkable within the discipline of natural language processing. A deeper comprehension of the text's context is made possible by the attention mechanism, which enables the model to accurately represent the relationship between words that are far apart [2].

The BERT framework consisted of two stages, pre-training and fine-tuning. Pre-training involves using several pre-training tasks to train the model on unlabeled data [3]. Using labelled data from earlier tasks, the first stage to be done in the BERT model is the pre-training parameters, then the process is followed by fine-tuning all of the parameters. Despite being started with the same pre-training parameters, every task has a unique fine-tuned model [4]. This study makes use of two BERT model versions that were created especially for Indonesian and multilingual languages: the BERT multilingual base model (cased) (google-bert/bert-base-multilingual-cased) and the IndoBERT base uncased (Rifky/Indobert-QA). Researchers

used two pre-existing BERT models monolingual BERT for Indonesian (called “IndoBERT”) and multilingual BERT (“mBERT”). MBERT is trained by combining 104 languages of Wikipedia documents, which also include Indonesian, and is effective for multilingual tasks that do not require much explanation [5].

IndoBERT is a BERT using transformer-based model [4], trained with the Huggingface framework with the default BERT base (uncased) configuration [6]. In contrast, another variant of BERT is multilingual-BERT (mBERT) that uses masked language modelling (MLM) to train it on the largest Wikipedia dataset across 104 languages, including Indonesian. Oversampling is used for small languages, and undersampling is used for large languages when dealing with unbalanced data. This makes it possible to employ the mBERT model on language datasets with limited resources, such as Indonesian [7]. The BERT configuration, 12 attention heads, 12 hidden layers of 768 each, and feed-forward hidden layers of 3,072 are shared by these two models [8].

We will use the datasets `Amod/mental_health_counseling_conversations` taken from the Huggingface website, to evaluate the two different models. Questions and responses from two online counselling and therapy platforms are included in this dataset. Qualified psychologists have answered the questions, which span a wide range of mental health concerns. The aim of this dataset is to enhance the quality of language models. The dataset is still in English, so the researcher manually translated it into Indonesian for more relevant testing. The evaluation of the dataset will use BERT-Score. This evaluation metric is very useful for evaluating the performance of question-answer [9].

Topics related to question-answer systems with Indonesians have been researched before, such as research conducted by Dzaky *et al.* [10], in his research creating a chatbot for mental health using a deep neural network and BERT models specifically for Indonesians. After testing, the results of this study showed an accuracy value of 71.73% [10]. In addition, in research conducted by Huzaeni *et al.* [11] where his research aimed to create a chatbot system to diagnose mental health symptoms using the BERT model, based on the results of tests that have been carried out, the accuracy value is 78.67% [11]. This research assesses the results of previous studies and focuses on improving performance for mental health question-answer systems by adopting Indonesian datasets.

This experimental research aims to demonstrate the efficiency in utilizing the M-BERT and IndoBERT models on an AI chatbot focusing on digital mental health. By conducting this research, it is expected to find out the relevance of BERT models that are suitable, efficient, and scalable for use in chatbot models. Other contribution of this research is:

- a. Optimized IndoBERT and MBERT model for QA task of the mental health domain in Indonesian language,
- b. The translation of the `Amod/mental_health_counseling_conversations` dataset to Indonesian language,
- c. Provide an overview of the mental health question-answer system that can be applied according to the intended specifications.

This type of research is potentially repeatable to be conducted for other kinds of foundational models' dataset. The result of that kind of study is possible to be applied to mental health questions answering system of different languages. It can be potentially implemented, especially in countries with large populations where sometimes suffer from insufficient mental health professionals.

2. METHOD

This study employs an explanatory research design with a quantitative approach. The explanatory technique was chosen since the study's goal is to describe and assess the cause-and-effect relationship between fine-tuning and performance improvement in the IndoBERT and MBERT language models [12]. The quantitative approach was utilized because the data acquired were numerical, indicating the level of model performance before and after optimization. The data was collected using experimental technique. Experiments were carried out by conducting fine-tuning to the IndoBERT and MBERT language models and then comparing their performance before and after this optimization process [13].

2.1. Instrument

2.1.1. Software and hardware

This research utilized Google Colab as the main development environment, supported by a Tesla T4 GPU. The T4 GPU, based on NVIDIA's Turing architecture, offers efficient performance for deep learning tasks with 16 GB GDDR6 memory and up to 8.1 TFLOPS of computer power. This setup enables faster model training and inference, especially for resource-intensive tasks in natural language processing. For offline preparation, we used a laptop with an Intel Core i5 processor, 8 GB RAM, and an NVIDIA GeForce MX230 GPU. Although local hardware was used for initial testing and code development, all fine-tuning and evaluation were conducted in the Colab environment to leverage cloud-based acceleration. This combination allowed seamless integration between code prototyping and scalable model training.

2.2. Material

2.2.1. Foundational model

In conducting this research, we used two language models developed based on the BERT model, namely IndoBERT and MBERT. IndoBERT is a transformer-based model in the BERT style. It was trained only as a masked language model using the Huggingface framework, adhering to the standard BERT-Base configuration (uncased) [14]. The model consists of 12 feed-forward hidden layers with 3,072d, 12 attention heads, and 12 hidden layers with 768d each. The model set the training to use 512 tokens per batch and extends the Huggingface framework to analyze distinct sets of text for distinct document blocks. A vocabulary of 31,923 Indonesian WordPieces was used to train IndoBERT [5]. The total amount of words used to train the IndoBERT model was over 220 million, gathered from three predominant sources: i) the 74 million words found on the Indonesian Wikipedia; ii) news items sourced from Kompas, Liputan6, Tempo (a total of 55 million words); and iii) the 90 million words found in the Indonesian Web corpus. Upon preprocessing the corpus into 512 tokenized document blocks, we were able to acquire 13,985 development examples and 1,067,581 training examples (without redundancy). Four Nvidia V100 GPUs (16 GB each), a batch size of 128 and a learning rate of $1e-4$, along with an Adam optimizer and a linear scheduler, were employed for the training. Over the course of two months, the model was trained for 2.4 million steps, or 180 epochs, and the development set's ultimate accuracy rate was 3.97, or comparable to BERT-base in English [15].

An Indonesian-language Wikipedia corpus consisting of 104 monolingual languages was used to pre-train the single-language multilingual BERT (M-BERT) model. Researchers can utilize the m-BERT model to tackle tasks in different languages because it has finished the pre-training process in an extensive variety of languages, expanding its usefulness [16]. The benefit of multilingual BERTs is that they facilitate multilingual learning. They are able to be coached on a specific task in one language and implement the same task quite well in another language, even if they are only trained for a monolingual task. The model uses 12 Transformer layers with 12 heads, 768 embedding dimensions, and 3072 feed-forward hidden layer dimensions, with a dropout rate of 0.1 and GELU activation. Multilingual BERT was optimized using Adam Optimizer, 88 batch sizes, and mixed-precision training with 10 epochs. Each model was trained on NVIDIA RTX Titan hardware with 24 GB of memory for approximately 20 hours [17].

2.2.2. Dataset

This research was conducted using the Amod/mental_health_counseling_conversations dataset taken from the Hugging Face website. This dataset originally contains conversations in English between counsellors and clients discussing mental health issues. To suit the context of this study, the dataset was translated into Indonesian. The translation process was done directly by the research team and not mechanically to ensure the accuracy and translation quality. The Amod/mental_health_counseling_conversations dataset consists of conversations covering a range of mental health-related topics, such as trauma, anxiety, relationship issues, and depression. Each conversation consists of multiple exchanges between the counsellor and the client, which allows researchers to analyze the dynamics of the conversation and effective counselling strategies.

The use of these datasets in research provides several advantages. Firstly, it provides a rich and realistic representation of mental health counselling situations, allowing researchers to explore the issue in a relevant context. Second, the translation into Indonesian allows this study to contribute to the growth of resources and technologies in Indonesian, which is still relatively lacking compared to other languages. Third, this dataset can be used to train and evaluate language models and conversational systems designed to assist in the context of mental health counselling.

The dataset translation process was done manually. In the process, we worked with a psychotherapy assistant and a lecturer in the Islamic psychology study program. In addition, the translator is a graduate of the Master of Islamic Psychology Education at Sunan Kalijaga State Islamic University Yogyakarta. The translator of the dataset has good Indonesian language skills and understands English as the main language of the translated dataset.

2.3. Procedure

In this research, there are several procedures that need to be done, starting from pre-processing data, and model training to model evaluation. as for the details that will be explained in Figure 1. Figure 1 shows the process flow in this research, where the first process is data collection which will be the main material in the question-answer system [18]. At this stage, the researcher uses a dataset from Amod/mental_health_counseling_conversations which will be translated into Indonesian. Furthermore, after the data is successfully obtained, data pre-processing will be carried out, where later the data will be made into tokens, and columns will be added to the dataset [19]. After pre-processing is complete, we will continue with model training for question-answer needs, where optimization will be carried out by fine-tuning the model to improve the quality of the model. Finally, the model will be tested using BERTScore to assess the optimization results [20].

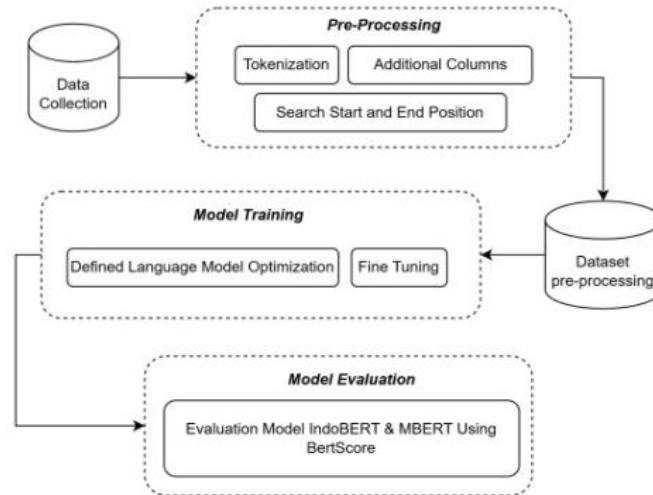


Figure 1. Research flow

2.3.1. Data pre-processing

Data needs to pass the important stage of pre-processing as preparation before it is used to train the question-answering model. This stage aims to transform unprocessed data into a suitable format that can be processed properly by deep learning models. Some of the data pre-processing steps that will be carried out are [21].

- Tokenization of questions and answers: The question and answer text is separated into individual tokens using a specialized tokenizer, such as the BERT tokenize [22]. This tokenization is important to break the text into units that can be understood by the model, such as words or sub-words. This stage ensures that the input data conforms to the format required by the BERT model [23].
- Search for the start and end position of the answer: The beginning and conclusion of the answer in the text factors are identified and marked [24]. This step helps the model learn the correlation among the question, the text context, and the actual position of the answer.
- Addition of start_positions and end_positions columns to the dataset: The start and end position information obtained from the previous step is added to the dataset as new columns, such as start_positions and end_positions [25]. These columns will be used as targets or labels in the model training process.

2.3.2. Model training

The model training stage is an important part of developing a deep learning-based system. This process involves training the transformer model using a pre-processed dataset. In this stage, hyperparameter determination and fine-tuning process will be executed on the model that has been determined [26].

- Defining language model optimization: To ensure a fair comparison, both IndoBERT and MBERT were fine-tuned using identical hyperparameters: learning rate of $2e-5$, batch size of 8, 3 training epochs, and a dropout rate of 0.1. Both models share the same architecture (12 hidden layers, 12 attention heads, and hidden size of 768). Optimization strategies include parameter initialization, tokenization, and regularization [27].
- Fine-tuning: Using labelled data from earlier jobs, the BERT model then initialized by using pre-trained parameters. The process then continues to adjustment of all parameters. Even while all tasks start with the same pre-training settings, each one has a unique refined model [28].

In performing question-answering tasks on Indonesian data, two transformer models can be used, namely IndoBERT and multilingual-BERT (MBERT). Both models are based on the BERT architecture that has proven reliable in several natural language processing tasks [29].

2.3.3. Model evaluation

In this research, the model performance evaluation uses BERTScore. BERTScore is an evaluation metric for language generation based on pre-trained contextual embedding BERT. BERTScore calculates the similarity between a couple of sentences as the total of the cumulative similarity between both sentences' embedding tokens. BERTScore assigns a true answer sentence (\hat{x}) and a prediction sentence (\tilde{x}). We use contextual embedding to represent the tokens and calculate the result utilizing cosine similarity, optionally importance reviewed with inverse document frequency (IDF) score [20]. BERTScore will match each token

in \bar{x} with a token in \bar{x} to calculate recall BERTScore, and each token in \bar{x} with a token in \bar{x} to calculate precision BERTScore. After that, the precision and recall values will be combined to calculate the F1 BERTScore size. The following is the formula used to calculate recall BERTScore, precision BERTScore, and F1 BERTScore [9]. For the evaluation of this question-answer system, we will test using BERTScore and compare the results before and after optimizing the model, in order to get clear and credible results.

3. RESULTS AND DISCUSSION

3.1. Data pre-processing

After the dataset is prepared, the data goes to the pre-processing phase. The dataset was readjusted before the question-answering model is trained. The goal of this step is to alter unprocessed input on a form that deep learning models can handle correctly. Additionally, the data was splitted in this process, with 80% of the data going toward training and 20% going toward validation.

3.1.1. Tokenization of questions and answers

At this stage, each question and answer were tokenized, to divide the text input in the data into smaller parts for easy processing. In the process, because there were two different models, the tokens used were also different between the IndoBERT and MBERT models. The IndoBERT model used the BERT Tokenizer from “Rifky/Indobert-QA” while MBERT used “bert-base-multilingual-cased”. In general, there was no difference in terms of the process, but the “Rifky/Indobert-QA” tokenization used an Indonesian dictionary, while “bert-base-multilingual-cased” used a dictionary of 100 languages in general, so for Indonesian specifically, “Rifky/Indobert-QA” is superior.

After determining the BERT tokenizer, iterate on each row and retrieve questions and answers from the dataset. Questions and answers were tokenized using the predefined tokenizer. The start and end positions of the answer in the input tokens were determined by finding the token [SEP] that separates the question and answer. If the [SEP] token is not found, the start and end positions were set to 0. The start and end positions were stored in the `start_positions` and `end_positions` lists. For determination, the [CLS] token was always at position 0, so after the [CLS] token was `start_positions`.

3.1.2. Addition of `start_positions` and `end_positions` columns to the dataset

After knowing the `start_positions` and `end_positions`, the `start_positions` and `end_positions` columns were added to the dataset. This was to share the labels or targets that the model will learn during the training process. In the question answering (QA) task, the model was required to determine the start and end positions of the answer from the context. By adding the `start_positions` and `end_positions` fields which put the starting and ending positions of the actual answer, the model could be trained to stick to patterns associated with the correct answer. During training, the model was receiving input in the form of questions and text context, and targets in the form of `start_positions` and `end_positions` of the actual answers. The model was learned to map the input to the correct targets so that during inference (judgment or usage), it could foresee the start and end positions of the correct answer for the given question and text context. Thus, the addition of the `start_positions` and `end_positions` fields to the training and validation datasets was an important step in preparing information for the Question Answering task so that the model could be trained to learn patterns that were associated with correct answers.

3.2. Model training

After determining the tokens in the dataset and adding the `start_positions` and `end_positions` columns, the next step was training the model. In the model training stage, the hyperparameters were determined and the fine-tuning process were performed on each model. This stage was the core process of this research.

3.2.1. Defining language model optimization

Determining the foundational model, strategy and optimal hyperparameter values, such as number of epochs, batch size, and learning rate, was very important to get good implementation from the model. In this research, we set the same value between the IndoBERT and MBERT models so that the comparison gets balanced results. The hyperparameters used in both models are Hidden Dropout Probability of 0.1, using Adam as the optimizer, learning rate of 2.00E-05 as well as using 3 epochs and batch size of 8. Foundational models of IndoBERT and MBERT have the same parameters, namely 12 hidden layers, 12 attention heads, and Hidden Size 768. For better model optimization, strategies such as parameter initialization, tokenization, model optimization, and regularization are needed.

Table 1 summarizes the optimization settings. The parameters follow the standard BERT configuration, and identical hyperparameters were applied to both models for a fair comparison. This

hyperparameter provision adapts to Devlin's research entitled “*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*”, where his research mentioned related hyperparameters that are optimal enough to be used in the fine-tuning process. The strategy used in this research includes parameter initialization, tokenization, model optimization, and regularization. For parameter initialization and tokenization, we use the foundation of IndoBERT and MBERT, while the regularization technique uses dropout probability, which was intended to reduce data overfitting during the fine-tuning process.

Table 1. Determining language model optimization

Variable	Item	IndoBERT	MBERT
Parameters	Hidden layers	12	12
	Attention heads	12	12
	Hidden size	768	768
Hyperparameters	Max length	512	512
	Learning rate	2.00E-05	2.00E-05
	Optimizer	Adam	Adam
	Epoch	3	3
	Batch size	8	8
Strategy	Parameter Initialization	Rifky/Indobert-QA	bert-base-multilingual-cased
	Tokenization	Rifky/Indobert-QA	bert-base-multilingual-cased
	Model optimization	Using Hyperparameter	Using Hyperparameter
	Regularization	Dropout Probability (0.1)	Dropout Probability (0.1)

3.2.2. Fine-tuning

After setting initial hyperparameters, fine-tuning was performed on both IndoBERT and MBERT. The process starts with tokenization, which differs from the earlier preprocessing step. It involves initializing the data frame, tokenizer, and text length, counting dataset size, and sampling from the “Question” and “Answer” columns. Special tokens like [CLS] and [SEP] were added, texts were truncated or padded to a maximum length, and inputs are converted to PyTorch tensors. Next, the model was built by adapting it to the question-answer task with the specified dataset and parameters. Training runs for 3 epochs on both training and validation data to optimize the model weights for better predictions. After training, the model and tokenizer were saved using `model.save_pretrained` and `tokenizer.save_pretrained` for later use without retraining.

3.3. Model evaluation

At this stage, the saved results of the fine-tuning process (the model) were evaluated using BERTScore. BERTScore is a language generation evaluation metric based on the pre-trained contextual embedding BERT. BERTScore calculates the similarity of two sentences as the total of the cumulative similarity between the embedding tokens.

As illustrated in Table 2, the IndoBERT model demonstrated a notable performance in terms of model evaluation using BERTScore. The model attained an F1-BERTScore of 91.8%, accompanied by a recall BERTScore of 89.9% and a precision BERTScore of 93.9%. In comparison, the MBERT model exhibited a comparatively lower performance, with an F1-BERTScore of 79.2%, a recall BERTScore of 73.4%, and a precision BERTScore of 86.2%. In addition, a comparison was made with the GPT-2 model, and the results obtained were F1-BERTScore 66.2%, recall BERTScore 68.0%, and precision BERTScore 65.5%. Based on the values obtained, it can be concluded that the IndoBERT model is superior to the MBERT and GPT-2 models in question and answer tasks using Indonesian-language mental health datasets.

Table 2. Determining language model optimization

Model	Item	F1 BERTScore	Precision BERTScore	Recall BERTScore
IndoBERT	Before Fine-Tuning	65.3%	73.8%	60.8%
	After Fine-Tuning	91.8%	93.9%	89.9%
MBERT	Before Fine-Tuning	74.2%	78.8%	72.2%
	After Fine-Tuning	79.2%	86.2%	73.4%
GPT-2	Before Fine-Tuning	58.4%	56.4%	60.0%
	After Fine-Tuning	66.2%	68.0%	65.5%

3.4. Discussion

The results of this study emphasize the importance of using language-specific models in NLP tasks. IndoBERT, which was pre-trained on Indonesian text, significantly outperformed the multilingual MBERT

after fine-tuning. IndoBERT achieved an F1-BERTScore of 91.8%, precision of 93.9%, and recall of 89.9%. In comparison, MBERT only reached 79.2% F1, with 86.2% precision and 73.4% recall. This 28% improvement in IndoBERT's F1 score compared to its performance before fine-tuning, versus only a 5% gain in MBERT, shows that fine-tuning is highly effective especially when the model architecture and training data are closely aligned with the target language. The lower performance of MBERT can be attributed to its multilingual training objective, which reduces its specialization in any single language, including Indonesian.

These findings support the idea that for specialized domains like mental health in a specific language, monolingual models are more suitable. Additionally, the optimized IndoBERT model has potential for real-world deployment in mental health chatbot applications. However, ethical and legal considerations remain essential, particularly regarding data privacy, content moderation, and appropriate response handling. Ensuring responsible AI design is critical before such systems are publicly implemented.

4. CONCLUSION

The intention of this work was to equalize the performance of BERT models in question-answering tasks, specifically IndoBERT and M-BERT models, utilizing Indonesian language datasets focused on mental health domain. The performance of IndoBERT surpassed M-BERT, with an F1-BERTScore of 91.8%, recall BERTScore of 89.9%, and precision BERTScore of 93.9%. Meanwhile, the MBERT model has a lower performance, with a value of F1-BERTScore of 79.2%, recall BERTScore of 73.4%, and precision BERTScore of 86.2%. These findings highlight the need to use language-specific models, such as IndoBERT for Indonesian, to improve the performance and relevance of responses in question-answer systems. In addition, this study demonstrates the effectiveness of fine-tuning methods in improving model performance which in this case the IndoBERT improve by 28% while MBERT improve by about 5%. The higher improvement of IndoBERT shows that models trained for a specific type of language (in this case the Indonesian language) can improve significantly when optimize for NLP tasks in that specific language. Other contribution of this research were; optimized IndoBERT and MBERT model for QA task of the mental health domain in Indonesian language, the translation of the Amod/mental_health_counseling_conversations dataset to Indonesian language, and provide an overview of the mental health question-answer system that can be applied according to the intended specifications. To summarize, this study not only demonstrate IndoBERT's superiority in Indonesian question-answering tasks compared to MBERT but also emphasizes the importance of investing in the development and optimization of language-specific models to improve the accessibility and quality of digital services, particularly in sensitive areas such as mental health. This study also shows a higher accuracy value compared to several previous studies with an F1-BERTScore of 91.8%, recall BERTScore of 89.9%, and precision BERTScore of 93.9%. These findings pave the way for future studies into the adaptation of NLP models to different cultural contexts and application domains.




REFERENCES

- [1] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, "COBERT: COVID-19 question answering system using BERT," *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 11003–11013, 2023, doi: 10.1007/s13369-021-05810-5.
- [2] A. Kesarwani, S. Das, D. R. Kisku, and M. Dalui, "Multi-scale vision transformer toward improved non-invasive anaemia detection using palm video," *Multimedia Tools and Applications*, vol. 83, no. 38, pp. 85825–85848, 2024, doi: 10.1007/s11042-024-20118-w.
- [3] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737–2757, 2023, doi: 10.1007/s10660-022-09560-w.
- [4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2019.
- [5] 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*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.
- [6] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: a pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 10660–10668, 2021, doi: 10.18653/v1/2021.emnlp-main.833.
- [7] Q. Li and Y. Zhang, "Improved text matching model based on BERT," *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 3, pp. 40–43, 2023, doi: 10.54097/fcis.v2i3.5209.
- [8] T. M. Luu, H. T. Le, and T. M. Hoang, "A hybrid model using the pretrained BERT and deep neural networks with rich feature for extractive text summarization," *Journal of Computer Science and Cybernetics*, vol. 37, no. 2, pp. 123–143, 2021, doi: 10.15625/1813-9663/37/2/15980.
- [9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTSCORE: evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020*, pp. 1–43, 2020.
- [10] A. A. Dzaky *et al.*, "Optimization Chatbot services based on DNN-BERT for mental health of university students," *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 13–21, 2024, doi: 10.30871/jaic.v8i1.7403.
- [11] H. Huzaeni, Z. K. Simbolon, and M. A. Firdaus, "Mental health disorder chatbot using NLP and forward chaining methods, case study of Cut Meutia Hospital," *Journal of Information Technology and Computers (JITC)*, vol. 4, no. 2, pp. 231–237, 2024.




- [12] M. Saunders, P. Lewis, and A. Thornhill, *Research methods for business students*. Pearson, 2007.
- [13] J. W. Cresswell, *Educational research planning, conducting, and evaluating quantitative and qualitative research*, 4th ed. Boston: Phoenix Color Corp, 2012.
- [14] R. Sutoyo, H. L. H. S. Warnars, S. M. Isa, and W. Budiharto, "Emotionally aware Chatbot for responding to Indonesian product reviews," *ICIC Express Letters*, vol. 19, no. 3, pp. 861–876, 2023, doi: 10.24507/ijicic.19.03.861.
- [15] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: a large-scale Indonesian dataset for text summarization," *arXiv preprint arXiv:2011.00679*, no. 1, 2020.
- [16] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
- [17] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 120–130, 2020, doi: 10.18653/v1/2020.repl4nlp-1.16.
- [18] A. Kazemi, J. Mozafari, and M. A. Nematbakhsh, "PersianQuAD: the native question answering dataset for the Persian language," *IEEE Access*, vol. 10, pp. 26045–26057, 2022, doi: 10.1109/ACCESS.2022.3157289.
- [19] A. Kesarwani, S. Das, D. R. Kisku, and M. Dalui, "Dual mode information fusion with pre-trained CNN models and transformer for video-based non-invasive anaemia detection," *Biomedical Signal Processing and Control*, vol. 88, p. 105592, 2024, doi: 10.1016/j.bspc.2023.105592.
- [20] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "Evaluating question answering evaluation," *MRQA@EMNLP 2019 - Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 119–124, 2019, doi: 10.18653/v1/d19-5817.
- [21] F. Baharuddin and M. F. Naufal, "Fine-tuning IndoBERT for Indonesian exam question classification based on bloom's taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, 2023, doi: 10.20473/jisebi.9.2.253-263.
- [22] X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, and X. J. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020, doi: 10.1007/s11431-020-1647-3.
- [23] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "QASiNa: religious domain question answering using Sirah Nabawiyah," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2023*, 2023, pp. 1–6, doi: 10.1109/ICAICTA59291.2023.10390123.
- [24] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bi-directional attention flow for machine comprehension," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–13, 2017, doi: abs/1611.01603.
- [25] B. Van Aken, A. Löser, B. Winter, and F. A. Gers, "How does BERT answer questions? A layer-wise analysis of transformer representations," in *International Conference on Information and Knowledge Management, Proceedings*, pp. 1823–1832, 2019, doi: 10.1145/3357384.3358028.
- [26] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," *AI (Switzerland)*, vol. 4, no. 1, pp. 54–110, 2023, doi: 10.3390/ai4010004.
- [27] A. Kesarwani, S. Das, D. R. Kisku, and M. Dalui, "Non-invasive anaemia detection based on palm pallor video using tree-structured 3D CNN and vision transformer models," *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–29, doi: 10.1080/0952813X.2023.2301401.
- [28] R. Calizzano, M. Ostendorff, Q. Ruan, and G. Rehm, "Generating extended and multilingual summaries with pre-trained transformers," in *2022 Language Resources and Evaluation Conference, LREC 2022*, no. June, pp. 1640–1650, 2022.
- [29] A. Fatwanto, F. Zamakhsyari, R. Ndungi, and L. Fitriyani, "Systematic literature review of BERT-based models for natural language processing tasks," *Infotel*, pp. 713–728, 2024, doi: 10.20895/INFOTEL.V16I3.1206.

BIOGRAPHIES OF AUTHORS



Fardan Zamakhsyari    studying to get a bachelor of computer science (S.Kom.) degree in information systems, at Sunan Ampel State Islamic University Surabaya in 2017, continuing his Master of Computer Science (M.Kom.) education in informatics, Sunan Kalijaga State Islamic University Yogyakarta in 2024. Currently working as a lecturer at the Department of Informatics, Faculty of Engineering, Sekolah Tinggi Cahaya Surya, Kediri, Indonesia. His research interests include web development, data science, natural language processing, and artificial intelligence. He can be contacted via email: masfardan99@gmail.com.



Agung Fatwanto    obtained a bachelor of science (S.Si.) in computer science and master's degree (M.Kom.) in Computer Science from Gadjah Mada University. He later pursued his doctoral education at the Department of Computer Science, Australian National University. He is currently a lecturer at the Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia. His research interests include software engineering, natural language processing, artificial intelligence and data science. He can be contacted via email: agung.fatwanto@uin-suka.ac.id.