Journal of
Information Systems Engineering
and Business Intelligence

Vol.9, No.2, October 2023
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy

Fikri Baharuddin [1)], Mohammad Farid Naufal [2)*]

[1,2)] *Teknik Informatika, University of Surabaya, Surabaya, Indonesia*

[1)]fikribaharuddin@staff.ubaya.ac.id, [2)]faridnaufal@staff.ubaya.ac.id

Abstract

Background: The learning assessment of elementary schools has recently incorporated Bloom's Taxonomy, a structure in education that categorizes different levels of cognitive learning and thinking skills, as a fundamental framework. This assessment now includes High Order Thinking Skill (HOTS) questions, with a specific focus on Indonesian topics. The implementation of this system has been observed to require teachers to manually categorize or classify questions, and this process typically requires more time and resources. To address the associated difficulty, automated categorization and classification are required to streamline the process. However, despite various research efforts in questions classification, there is still room for improvement in terms of performance, particularly in precision and accuracy. Numerous investigations have explored the use of Deep Learning Natural Language Processing models such as BERT for classification, and IndoBERT is one such pre-trained model for text analysis.
Objective: This research aims to build classification system that is capable of classifying Indonesian exam questions in multiple-choice form based on Bloom's Taxonomy using IndoBERT pre-trained model.
Methods: The methodology used includes hyperparameter fine-tuning, which was carried out to identify the optimal model performance. This performance was subsequently evaluated based on accuracy, F1 Score, Precision, Recall, and the time required for the training and validation of the model.
Results: The proposed Fine Tuned IndoBERT Model showed that the accuracy rate was 97%, 97% F1 Score, 97% Recall, and 98% Precision with an average training time per epoch of 1.55 seconds and an average validation time per epoch of 0.38 seconds.
Conclusion: Fine Tuned IndoBERT model was observed to have a relatively high classification performance, and based on this observation, the system was considered capable of classifying Indonesian exam questions at the elementary school level.

Keywords: IndoBERT, Fine Tuning, Indonesian Exam Question, Model Classifier, Natural Language Processing, Bloom's Taxonomy

Article history: Received 25 July 2023, first decision 28 August 2023, accepted 7 October 2023, available online 28 October 2023

I. INTRODUCTION

Critical thinking skills are essential in the field of education, serving as the foundation for all academic pursuits [1]. This skill have been widely assessed by educational institutions through teaching, comprehension, quality evaluation, and testing, with the aim of facilitating optimal learning among students. During this process, teachers are often saddled with the responsibility of designing course materials and specific course learning outcomes (CLOs) to emphasize the thinking abilities of students [2]. To achieve this objective, quality assessments were conducted by accreditation bodies and regulatory organizations, with exam questions playing a crucial role in the evaluation process. Furthermore, to facilitate the process, it was established that educational institutions, teachers, and accrediting bodies require a hierarchical system to distinguish the thinking behaviors of different individuals during learning [3]. In 1956, Benjamin Bloom introduced classification system known as "Bloom's Taxonomy" to categorize various thinking behaviors that are crucial in the learning process [4]. Conventionally, the categorization of exam questions into multiple levels of Bloom's taxonomy was performed manually by teachers and accreditation bodies based on their expertise on the subject matter. This manual method was time-consuming and susceptible to errors due to human biases, and based on these challenges, there was a growing need to automate the process, which fell within text classification. It is important to comprehend that the efficient increase in classification demanding substantial resources can be effectively addressed through the application of machine learning models. These models have been found to possess the capacity to grasp the required classification patterns, thereby automating classification process. Automation, in this context, offers the capability to significantly expedite the task of educators in grouping exam questions based on the thinking levels in Bloom's Taxonomy. Acknowledging the intricacy of examination questions

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.9.2.253-263

is of substantial significance in augmenting the cognitive aptitude of every student. Therefore, disregarding the difficulty level when organizing questions can lead to issues including questions being overly simple or too difficult and burdensome.

This research is centered on classification of Indonesian elementary school exam questions, leveraging the framework of Bloom's Taxonomy to assess the level of cognitive thinking. Within the domain of Machine Learning, Natural Language Processing (NLP) specializes in language and text data analysis. Meanwhile, Bidirectional Encoder Representations from Transformers (BERT) refers to a transformative model designed to pre-train deep bidirectional representations from unlabeled text, considering context from both left and right sides throughout all layers [5]. BERT is composed of a multi-layer bidirectional transformer encoder, renowned for its capability to effectively manage multiple tasks simultaneously [5]. IndoBERT, on the other hand, is a transformer-based model designed in accordance with BERT architecture, specifically trained as a masked language model using the Huggingface framework, and configured with the default configuration for BERT-Base (uncased) [6]. It is important to note that this model was specifically designed for the analysis of Indonesian text. In this research, IndoBERT pre-trained model was used for model training, which is an extension of the BERT architecture. However, while BERT is widely applied in tasks like sentiment analysis and article classification, it has been limitedly used for assessing the difficulty of exam questions, particularly in the context of Indonesian exam. The BERT model was leveraged in this investigation to map questions weights, thereby enabling their categorization based on the appropriate level within Bloom's Taxonomy. This system was also used to classify questions based on the verbs contained within. For example, questions containing the word "complete" were categorized into the C3 taxonomy (application), while some others using the same verb were classified under the C4 (analysis category). This differentiation underscored the rationale behind the use of the BERT architecture in this research.

In earlier research [7], the discussion concerning the classification of events revolved around the application of deep learning methods to categorize incidents shared in Bahasa Indonesia on Twitter. The investigation compared three different methods and underscored that the combination of CNN with NeuroNER produced the most favorable outcomes for multi-label classification. This exam also proposed that future research should explore alternative classification parameters and assess scenarios consisting of data from various sources. Subsequently, another previous exploration [8] concentrated on enhancing multi-label classification specifically for biomedical data, within the field of health-related questions-answer systems. This enhancement was implemented by the amalgamation of deep learning and machine learning methods to boost accuracy. From the obtained results, it was discovered that heterogeneous ensembles outperformed homogeneous ones, particularly when handling biomedical QA data characterized by lengthy text dimensions. Both of these prior studies used datasets compiled using Indonesian language and harnessed machine learning and deep learning methods for text classification. However, it is important to note that their investigations addressed different contexts compared to the present exploration. The earlier studies also did not leverage IndoBERT, which offers distinct advantages due to its extensive training in Indonesian text.

Previous research [9] has also been carried out to construct exam questions classifier grounded in Bloom's Taxonomy. The model designed in the investigation was trained using Word2Vec and TF-IDF methods, but its accuracy and certain other evaluation metrics remained somewhat low, indicating that there was room for improvement with the use of more recent methods. This exploration was carried out using an English-language dataset with a relatively limited sample of 140 data points. Accordingly, classification of exam questions within the context of Indonesian elementary schools was explored extensively in another previous investigation [10] using a different method. This classification was established using the REPTree algorithm and executed through the use of WEKA Tools.

As a novel exploration, this research aims to create classification model that is capable of assessing the difficulty level of Indonesian elementary school exam questions in line with Bloom's Taxonomy. It is important to understand that an adept classifier model was developed for categorizing the difficulty of elementary school exam questions in Indonesia. This model considered the specific aspects of Bloom's Taxonomy, and its classification performance was significantly improved through the use of fine-tuning configurations.

II. Literature Review

In order to effectively achieve the objective of this research, which includes creating a classifier model for categorizing Indonesian exam questions based on Bloom's Taxonomy, a comprehensive literature review was conducted. Accordingly, this review included exam of various studies and scientific articles on text classification such as journals, books, and relevant scientific papers that are in line with the thematic focus of the research.

In the context of questions classification, it was observed that previous research efforts have yielded similar models. In a particular study [9], three models were developed for classification of English questions grounded in Bloom's

Taxonomy. The first model used the TF-IDF method, while the second adopted the TF-POSIDF method, a modification of the basic TF-IDF method. The third model, abbreviated as W2V-TFPOSIDF, was created by combining Word2Vec and TF-POSIDF. These three models were trained on a dataset comprising 100 data points, and classification performance with precision ranging from 0.7 to 0.8 was achieved. Another research [10] also focused on developing a model for classifying the difficulty of exam questions using the REPTree algorithm. However, this study did not incorporate Bloom's Taxonomy as a reference for classification.

TABLE 1
LITERATURE REVIEW COMPARISON

Research Author	Dataset Size	Objectives	Method / Algorithm	Evaluation Aspect	Accuracy	Precision	F-Measurement
Mohammed and Omar [9]	141	Questions Classification	TF-IDF, TF-POSIDF, W2V-TFPOSIDF	Precision, Recall, F-measure	-	[0,77 – 0.86]	[0.724 – 0.837]
Baharuddin and Tjahyanto [10]	418	Questions Classification	REPTree	Accuracy	91.15%	-	-
Paul and Saha [11]	144,727	Text Classification	Fine-tuning BERT	F-measure	-	-	94%
Kaliyar, Goswami, and Narang [12]	20,800	Text Classification	BERT	Accuracy	98.90%	-	-
Rahmawati, Alamsyah and Romadhony [13]	2000	News Classification	IndoBERT	Accuracy	90%	-	-
Chen and Cong [14]	15200	Chinese News Classification	BERT + CNN	Accuracy	98.87%	-	-
Rahutomo and Pardamean [15]	1101	Indonesian Slang Language Classification	Fine-tuning IndoBERT	Accuracy	68%	-	-
Khan, Amjad, and Chang [16]	9312	Urdu Sentiment Analysis	BERT, Machine Learning, Deep Learning	F-measure	-	-	81.49%
Kulkarni, mandhane, and Joshi [20]	4779	Marathi Text Classification	BERT, LSTM, ULMFiT, CNN, Bi-LSTM	Accuracy	97.48%	-	-

Several previous investigations have leveraged the BERT architecture to address a variety of data analysis needs in textual form [17][18][19]. Some have used BERT to create specialized models, such as the CyberBERT model, designed for cyberbullying detection [11], and others have adapted BERT for classification of fake news [13][12]. BERT is widely used for text classification, and this method has been adopted in numerous research endeavors for the development of language-specific text classifier models, including Chinese-language news text classification [14], Indonesian slang for Trading [15], Urdu sentiment analysis [16], and text classification for Marathi [20] using datasets from [21]. Accordingly, this evidenced the versatility of the BERT architecture in classifying text data across multiple languages, including Indonesian. Further insights from previous results are presented in Table 1, which provides a comprehensive overview of the literature studies conducted in this regard.

III. METHODS

The methods leveraged in this study consist of several key stages, including data acquisition, dataset pre-processing, model training, and tuning, and the evaluation of classifier performance. The detailed research methodology used is visually presented in Fig. 1.

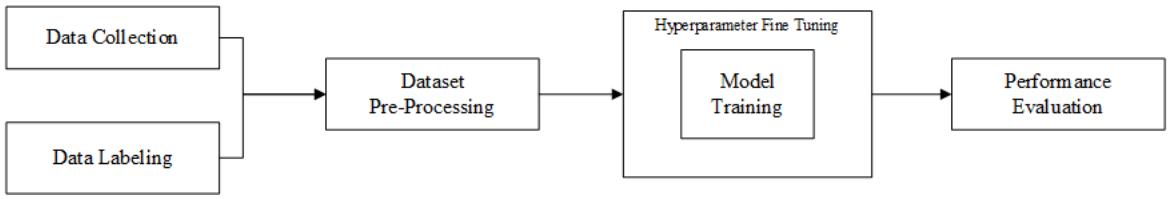


Fig. 1 Flow of applicated research method in this study

A. Data Collection

This case study focuses primarily on Indonesian elementary school level exam, and it necessitated the primary data to be centered around Indonesian exam questions suitable for elementary school students. The input comprised data extracted from multiple-choice questions. These questions were sourced from various written materials, including test preparation practice books, study guides, past examinations, and from multiple elementary schools in Sampang Regency, Madura, Indonesia. In total, 449 data points were compiled from the provided sources. For reference, the dataset is accessible via [22], and a sample of the collected data is presented in Table 2.

TABLE 2
SAMPLE OF COLLECTED DATA

Data (Bahasa Indonesia)	Data (English)	Taxonomy
Perbaikan ejaan untuk kata yang ditulis dalam huruf miring adalah sebagai berikut	*Spelling corrections for words written in italics are as follows*	C1 – Knowledge
Makna yang terkandung dalam kata kias "merah" pada syair di atas adalah	*The meaning contained in the figurative word "red" in the above poem is*	C2 – Understanding
Kata yang tepat untuk mengganti kata "penghabis" adalah	*The correct word to replace the word "ending" is*	C3 – Applicating
Nilai moral positif tokoh Reno pada cerita tersebut adalah	*The positive moral values from Reno in the story are*	C4 - Analyzing

B. Data Labeling

The collected data was systematically categorized and labeled according to the corresponding domains of Bloom's Taxonomy, as outlined in research [23] sourced from [24]. Bloom's Taxonomy is a widely recognized framework in educational systems. This framework comprises six distinct aspects namely memorizing or knowledge, understanding, applying, analyzing, evaluating, and creating. Accordingly, the labeling process was conducted manually and validated by two experienced Indonesian teachers in Sampang Regency. These teachers were responsible for tutoring Indonesian subjects at the elementary school level for over a decade. The labeling procedure included a discussion system, with one expert (Expert A) making the labeling decisions, while the other (Expert B) provided reinforcement and clarification. In cases of disagreement regarding the level of Bloom's Taxonomy, the final label used was based on the decision of Expert A, given their higher level of expertise and extensive experience in teaching Indonesian subjects.

As a result of the data collection and labeling process, the data was successfully categorized into four of the six Bloom's Taxonomy categories. These categories include C1 (Knowledge), C2 (Understanding), C3 (Application), and C4 (Analysis). However, is important to acknowledge that C5 (Evaluation) and C6 (Creation) were not used in this research dataset, as the collected data exclusively consisted of multiple-choice questions. The Evaluation and Creation aspects are typically used to assess the ability of students to respond to essay questions. The division of exam questions dataset based on Bloom's Taxonomy is shown in Fig. 2.

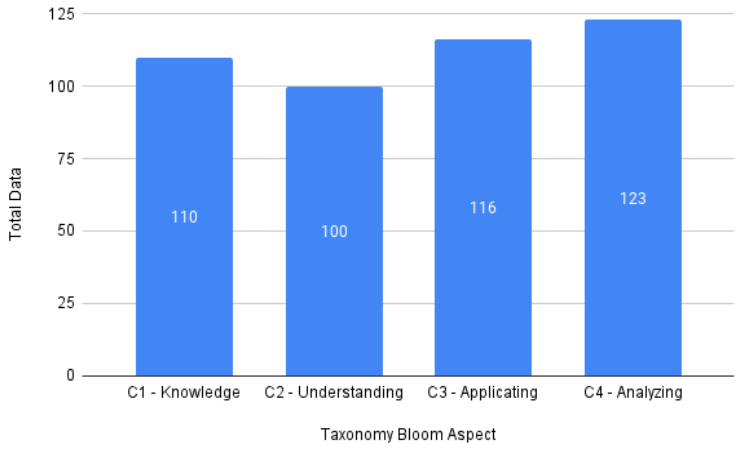


Fig. 2 Visualization of dataset

C. Dataset Pre-processing

The application of pre-processing methods to specific datasets makes it possible to significantly enhance text classification quality, with improvements of up to 80% being achievable [25][26]. The dataset pre-processing phase comprises various steps, including tokenization, stop words removal, punctuation removal, duplicate removal, and text case-folding. In this regard, tokenization, refers to the division of sentences into individual words or tokens. Stop word removal was used to eliminate common and frequently occurring words from the dataset. Punctuation removal removes punctuation marks such as periods, commas, questions mark, and exclamation points from the text. Duplicate removal ensures that data is free from redundancy, and text case-folding was performed to standardize all text to lowercase. These pre-processing steps collectively contributed to refining the dataset for more effective text classification.

D. Hyperparameter fine-tuning and Model Training

For model training, the IndoBERT pre-trained model was trained using 60% of the obtained data, while the remaining 40% was used for testing. It is important to establish that IndoBERT shares the same architectural foundation as the BERT-based model. Based on this understanding, it can be seen that achieving optimal classification performance requires fine-tuning, and this includes adjusting the hyperparameters used by the optimizer. In the BERT model, fine-tuning is essentially accomplished by introducing a special token to specify the intended task of the model. IndoBERT, like BERT, can conduct sequence classification to categorize text based on predefined classes or labels. The process of sequence classification in this context comprises the use of [SEP] token to demarcate the boundary between the input sequence and the label. In this research, the [SEP] token was used to separate questions text input, which subsequently served as the weighted input along with the label/class of questions text. Furthermore, the inclusion of the [SEP] token helped the model recognize context boundaries more effectively, which eventually led to higher classification accuracy. Fine-tuning process of the BERT architecture is shown in Fig. 3.

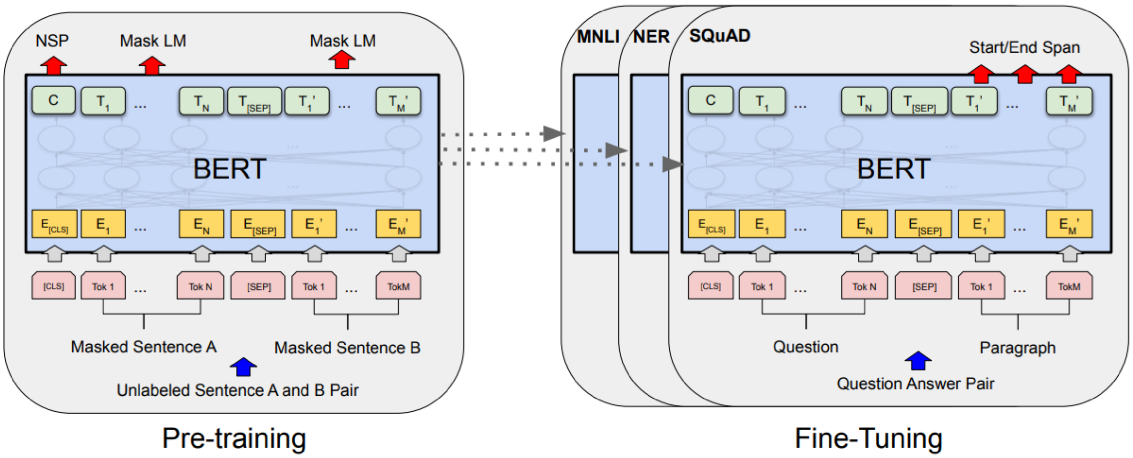


Fig. 3 BERT Pre-training and fine-tuning architecture[5]

Fine-tuning within the BERT architecture has been a prevalent practice in previous studies, as it efficiently plays the role of enhancing the performance of classifier models. For instance, prior exploration [15] adopted fine-tuning to build models aimed at understanding Indonesian slang used by stock traders. Similarly, some research have applied fine-tuning of the BERT architecture to a variety of other models, including sentiment classification[27], fake news classification [28], and customer review classification[29]. In line previous studies [5], to achieve the most favorable fine-tuning results, it was recommended to experiment with learning rate values such as 2E-5, 3E-5, and 5E-5 when using the Adam optimizer. In this study, a comparison of these three recommended learning rate parameter values was conducted with the aim of determining the optimal model performance. Furthermore, the model training in this investigation was carried out by using facilities provided by Google Colaboratory with detailed hardware specifications, as shown in TABLE 3.

TABLE 3
GOOGLE COLABORATORY RUNTIME ENVIRONMENT

Hardware	Specification
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
RAM	12.7 GB
GPU	Nvidia Tesla T4 16GB
DISK	78.2 GB

E. Performance Evaluation

Model performance evaluation was carried out by paying attention to the values of the accuracy, precision, and recall of the developed system. Measurement of the level of accuracy was performed using the calculation formula shown in (1), where accuracy is obtained by dividing the sum result between true positive (TP) and true negative (TN) with the sum results of true positive (TP), true negative (TF), false positive (FP), and false negative (FN).

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

Precision was calculated using (2), where the precision value was obtained from true positive (TP) divided by the sum results of true positive (TP) and false positive (FP).

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

Equation (3) was used to calculate the recall value. From the formula, it can be seen that this value was obtained by dividing the true positive (TP) by the sum result of the true positive (TP) and false negative (FN).

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

Lastly, F-measurement or F-score was calculated using the formula in (4). This value was obtained by dividing the result of the multiplication of the results of Precision and Recall, with the sum of their results.

$$F1\ Score = \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

The performance of the classifier model was comprehensively evaluated, with specific consideration on the calculated results of each assessment attribute mentioned earlier. The most favorable classifier model is one that shows the highest values across all relevant aspects. The evaluation process included the visualization of the level of accuracy and loss possessed by the model to identify potential issues of overfitting or underfitting.

IV. RESULTS

The pre-processed dataset was trained extensively, as this is necessary for the adequate construction of a classifier model. Furthermore, in this research, fine-tuning process was implemented to identify the classifier model with the highest performance. Fine-tuning, in this context, includes adjusting the hyperparameter configurations used during the model training. The specific hyperparameter settings used in this experiment are detailed in Table 4. It is also important to establish that the hyperparameter tuning in this investigation focuses solely on the learning rate aspect, with learning rate values of 2E-5, 3E-5, and 5E-5 being explored.

TABLE 4
HYPERPARAMETER CONFIGURATION FOR TUNING

Hyperparameter Configuration	Option
Optimizer	Adam
Learning Rate	2E-5, 3E-5, 5E-5
Batch Size	32

During the training phase, a callback function was implemented, which disrupted the training process prematurely if no improvements were observed in the validation accuracy over three consecutive epochs. In the absence of accuracy improvement, the callback automatically terminated the iterations and retrieved the best-performing data of the model.

This performance was subsequently evaluated across several aspects of the evaluation matrix, including accuracy, F1 Score, Recall, Precision, time taken per training epoch, and time taken per validation epoch.

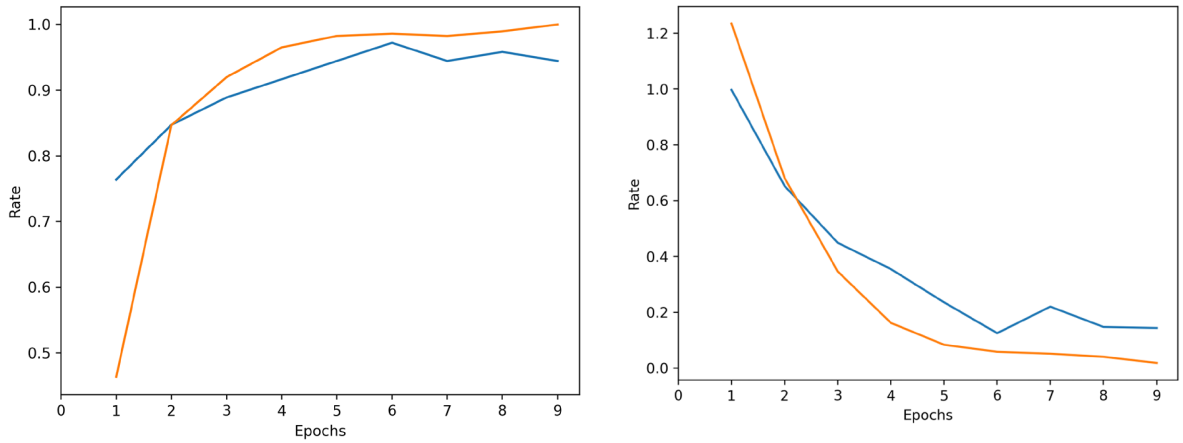
In accordance with the evaluation process, the experiment comprising a learning rate parameter of 2E-5 was applied to the optimizer during model training, the results showed classification performance with a validation accuracy rate of 97%, F1 Score of 97%, Recall of 97%, and Precision of 98%. On average, it took 1.55 seconds to process one training epoch and 0.38 seconds to process one validation epoch.

TABLE 5
VALIDATION PERFORMANCE COMPARISON AFTER TUNING

Learning Rate(LR)	Best Accuracy	F1 Score	Recall	Precision	Average Training Time/Epoch(second)	Average Validation Time/Epoch(second
2E – 5	0.97	0.97	0.97	0.98	1.55	0.38
3E – 5	0.93	0.93	0.93	0.94	1.48	0.43
5E – 5	0.94	0.95	0.95	0.95	1.84	0.43

The performance evaluation results presented in Table 5 show that fine-tuning with a learning rate parameter of 2E-5 produced the highest performance. In addition to assessing the results of the evaluation matrix, selecting the best model also includes considering the comparison of accuracy and loss observed during the model training process. This evaluation was carried out to determine whether the resulting model was suffering from overfitting or underfitting issues. Overfitting, in this context, occurs when a statistical machine learning model learns the training dataset so well that it performs poorly on unseen datasets, while underfitting happens when the model includes too few predictors [30].

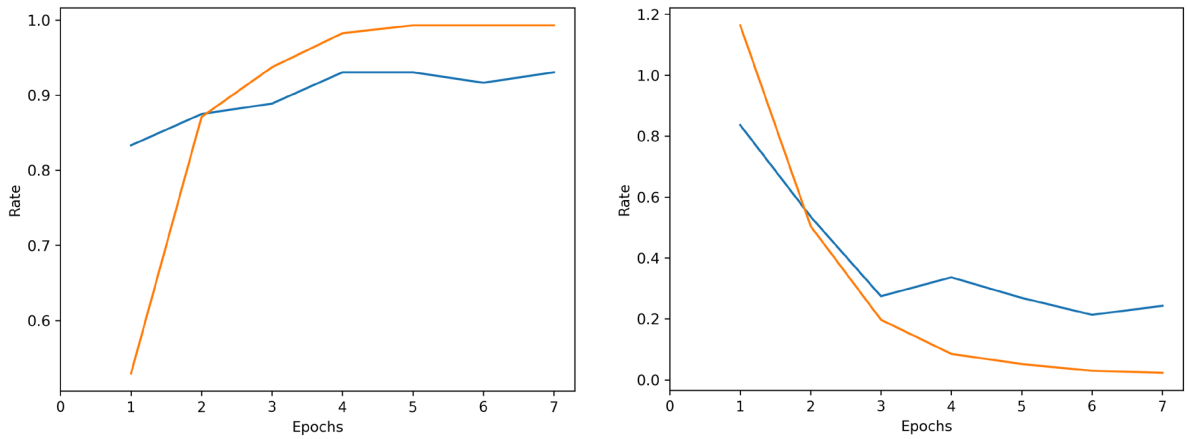
Fig. 4 provides a comparison of accuracy and loss during the training and validation phases when the model was trained with the Adam optimizer using a learning rate parameter value of 2E-5. Based on the accuracy comparison, no evidence of underfitting was observed, as there was no significant decline in accuracy. Moreover, the comparison of loss levels did not show overfitting, as these levels remained within tolerable limits, and the loss-to-accuracy ratio sustained relative stability.



(a) (b)

Fig. 4 Tuning model using LR 2E-5 (a) accuracy graph, (b) loss graph

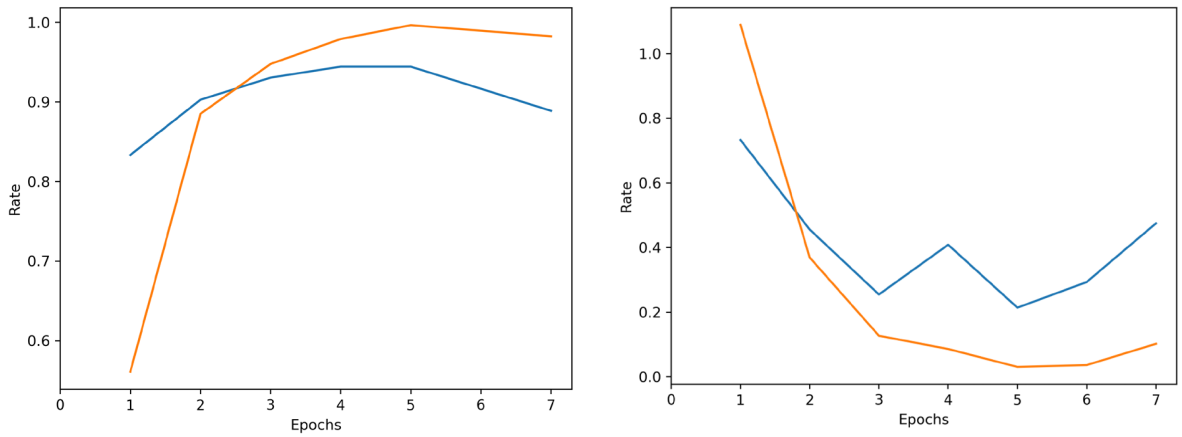
The comparison graph for accuracy and loss in the second fine-tuning experiment, which used a learning rate of 3E-5, is shown in Fig. 5. Based on the accuracy chart, it is evident that no underfitting issues were observed during the model training process. In accordance with this, the training accuracy consistently increased, which showed effective training. In the loss chart, the model also did not exhibit signs of overfitting, as the difference in loss ratios was not excessively significant and remained within acceptable limits. The loss range in this research appeared to be higher compared to fine-tuning models using the 2E-5 learning rate, but it was still manageable.



(a) (b)

Fig. 5 Tuning Model using LR 3E-5 (a) Accuracy Graph, (b) Loss Graph

The experiments conducted with learning rates set at 5E-5 showed a comparison of accuracy and loss, as can be observed in Fig. 6. During this experiment, it was evident that the training accuracy experienced a significant decrease and the validation accuracy was substantially reduced, showing overfitting. The loss comparison chart shows a significant increase in validation loss, which surpassed the acceptable tolerance limit. This provides further evidence that the model in this third experiment experienced overfitting.



(a) (b)

Fig. 6 Tuning Model using Learning Rate 5E-5 (a) Accuracy Graph, (b) Loss Graph

Considering the results from the evaluation metrics and the comparison of accuracy and loss graphs, it becomes evident that the model trained with a learning rate of 2E-5 stood out as the best-performing model. A separate validation of this top-performing model was executed by testing its performance using new Indonesian elementary school exam data. The outcomes of this separate classification test showed excellent classification results, as presented in Table 6.

V. Discussion

Categorizing exam questions according to cognitive domains, such as Bloom's Taxonomy, has become a widely adopted method in educational assessment. Consequently, there has been a growing need for a dependable system that can aid teachers in the automatic, rapid, and accurate grouping of exam questions based on their cognitive levels within Bloom's Taxonomy. This research focuses on classification of exam questions using the deep learning architecture of IndoBERT. In line with the experiments conducted in references [12], [13], [20], it becomes evident that deep learning

methods excel in text classification. This is supported by the obtained classification accuracy results, which reached an impressive 97% in this research. The results obtained from this comparison are presented in Table 7.

TABLE 6
CLASSIFICATION TEST USING SEPARATED DATASET RESULT

Questions Text (Bahasa Indonesia)	Questions Text (English)	Class Predicted	Prediction Accuracy*
Perbaikan penggunaan tanda baca pada kalimat di atas adalah	*The improvements to the use of punctuation in the sentence above are:*	C1 - Pengetahuan	99.325%
Makna kata pengganti dalam kalimat ini dapat dijelaskan sebagai	*The meaning of the word substitute in this sentence can be explained as*	C2 - Pemahaman	93.332%
Latar tempat pada cerita tersebut adalah	*The setting of the story is*	C4 - Analisis	95.595%
Susunan kalimat yang tepat agar menjadi paragraf deskripsi adalah	*The proper sentence structure to be a description paragraph is*	C3 - Aplikasi	86.133%
*Optimizer = Adam, Learning Rare = 2E-5, Batch Size = 32, Epoch = 9			

TABLE 7
CLASSIFIER MODEL PERFORMANCE COMPARISON

Research	Dataset	Dataset Availability	Methods / Algorithms	Total Dataset	Accuracy	F1 Score	Recall	Precision
Kaliyar et al.[12]	Fake News Dataset	Public	BERT + CNN	20800	98%	-	-	-
Rahmawati et al.[13]	Hoax News Dataset	Private	IndoBERT	2000	90%	90%	90%	90%
Kulkarni et al.[20]	Marathi News Dataset	Public	IndicBERT	4779	99%	-	-	-
Proposed model in this study	Indonesian Exam Questions with Bloom's Taxonomy Dataset	Public	Fine Tuning IndoBERT	449	97%	97%	97%	98%

This research underscored the impact of fine-tuning on classifier model performance. Several experiments were conducted on the learning rate parameter, and three different parameter values, namely 2E-5, 3E-5, and 5E-5, were tested in line with the recommendations from previous studies [5]. The results of these experiments showed that fine-tuning IndoBERT with a learning rate parameter of 2E-5 produced the best model performance. This was evident from the significantly higher accuracy, precision, F-score, and recall results in comparison to the experiments using different parameter values. Additionally, these results emphasized the substantial influence of the learning rate parameter on model performance. It was observed that excessively low parameter values can lead to prolonged training times and the risk of overfitting.

It is important to acknowledge that this research has some limitations. Firstly, the dataset used for model construction consisted of only 449 data points, which had to be divided into training and validation data at a 60:40 ratio. Secondly, the distribution of data among the labels or classes appears unbalanced, with certain classes having a dominant amount of data compared to others. In this regard, future research endeavors could benefit from a larger dataset with a more balanced distribution. Thirdly, the investigation exclusively used multiple-choice exam questions data, which did not comprise all cognitive aspects of Bloom's Taxonomy. Further studies can explore a broader range of data, potentially including a combination of multiple-choice questions and essays to include all cognitive aspects of the used framework.

VI. CONCLUSIONS

In conclusion, through a series of experiments, a classifier model that excels in categorizing Indonesian elementary school level exam questions based on Bloom's Taxonomy was successfully developed in this research. IndoBERT pre-trained models in the experiment were fine-tuning, and this process significantly elevated the model performance. Additionally, it is important to note that the choice of the learning rate parameter in the optimizer during model training significantly influenced the performance of the classifier. The proposed model was found to possess the potential to serve as a valuable tool for educators, which can aid in categorizing and classifying exam questions according to the aspects related to Bloom's Taxonomy at Indonesian elementary school level.

This exploration possesses certain limitations, first, the dataset used had uneven distribution among the taxonomy classes, and it was confined to multiple-choice exam questions, leaving some aspects of the leveraged approach unexplored. Secondly, the dataset exclusively comprised Indonesian language exam questions at the elementary school level. Thirdly, the investigation exclusively used multiple-choice exam questions data, which did not comprise all cognitive aspects of Bloom's Taxonomy. Future studies addressing these limitations could benefit from more comprehensive datasets that cover all aspects of the leveraged framework, along with ensuring a more balanced distribution of data for each class. Moreover, future investigations may explore the development of models for different case studies or adopt alternative methods. The model could also be evolved into an integrated application usable across various platforms for enhanced accessibility.

Author Contributions: *Fikri Baharuddin:* Conceptualization, Methodology, Data Collection, Writing - Original Draft. *Mohammad Farid Naufal*: Methodology, Review, Supervision, Review & Editing.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The dataset used in this study is publicly accessible. The dataset can be accessed at https://doi.org/10.5281/zenodo.8331563.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Fikri Baharuddin: https://orcid.org/0000-0002-2495-5878
Mohammad Farid Naufal: https://orcid.org/0000-0002-1194-2760

REFERENCES

[1] N. Omar *et al.*, "Automated analysis of exam questions according to Bloom's taxonomy," *Procedia-Social and Behavioral Sciences*, vol. 59, pp. 297–303, 2012.

[2] W.-C. Chang and M.-S. Chung, "Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items," in *2009 Joint Conferences on Pervasive Computing (JCPC)*, IEEE, 2009, pp. 727–734.

[3] S. U. Monrad *et al.*, "What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy," *Med Teach*, vol. 43, no. 5, pp. 575–582, 2021.

[4] B. S. Bloom and D. R. Krathwohl, *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. longman, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," *arXiv preprint arXiv:2011.00677*, 2020.

[7] S. R. Anggraeni, N. A. Ranggianto, I. Ghozali, C. Fatichah, and D. Purwitasari, "Deep Learning Approaches for Multi-Label Incidents Classification from Twitter Textual Information.," *Journal of Information Systems Engineering & Business Intelligence*, vol. 8, no. 1, 2022.

[8] A. F. Abdillah, P. Putra, C. Bagus, S. Juanita, and D. Purwitasari, "Ensemble-based Methods for Multi-label Classification on Biomedical Question-Answer Data.," *Journal of Information Systems Engineering & Business Intelligence*, vol. 8, no. 1, 2022.

[9] M. Mohammed and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, vol. 15, no. 3, p. e0230442, 2020.

[10] F. Baharuddin and A. Tjahyanto, "Peningkatan Performa Klasifikasi Machine Learning Melalui Perbandingan Metode Machine Learning dan Peningkatan Dataset," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 1, pp. 25–31, 2022.

[11] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification," *Multimed Syst*, vol. 28, no. 6, pp. 1897–1904, 2022.

[12] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed Tools Appl*, vol. 80, no. 8, pp. 11765–11788, 2021.

[13] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax News Detection Analysis using IndoBERT Deep Learning Methodology," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2022, pp. 368–373.

[14] X. Chen, P. Cong, and S. Lv, "A long-text classification method of Chinese news based on BERT and CNN," *IEEE Access*, vol. 10, pp. 34046–34057, 2022.

[15] R. Rahutomo and B. Pardamean, "Finetunning IndoBERT to Understand Indonesian Stock Trader Slang Language," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, IEEE, 2021, pp. 42–46.

[16] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, "Multi-class sentiment analysis of urdu text using multilingual BERT," *Sci Rep*, vol. 12, no. 1, p. 5436, 2022.

[17] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci China Technol Sci*, vol. 63, no. 10, pp. 1872–1897, 2020, doi: 10.1007/s11431-020-1647-3.

[18] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med Res Methodol*, vol. 22, no. 1, p. 181, 2022.

[19] A. Esmaeilzadeh and K. Taghva, "Text classification using neural network language model (nnlm) and bert: An empirical comparison," in *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*, Springer, 2022, pp. 175–189.

[20] A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, J. Jagdale, and R. Joshi, "Experimental evaluation of deep learning models for marathi text classification," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, Springer, 2022, pp. 605–613.

[21] D. Kakwani *et al.*, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.

[22] F. Baharuddin, "Elementary School Indonesian Exam Questions with Bloom's Taxonomy Dataset," Sep. 2023, doi: 10.5281/ZENODO.8331563.

[23] I. M. Arievitch, "The vision of Developmental Teaching and Learning and Bloom's Taxonomy of educational objectives," *Learn Cult Soc Interact*, vol. 25, p. 100274, Jun. 2020, doi: 10.1016/J.LCSI.2019.01.007.

[24] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2021.

[25] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, p. e0232525, 2020.

[26] A. Kathuria, A. Gupta, and R. K. Singla, "A Review of Tools and Techniques for Preprocessing of Textual Data," in *Computational Methods and Data Engineering*, V. Singh, V. K. Asari, S. Kumar, and R. B. Patel, Eds., Singapore: Springer Singapore, 2021, pp. 407–422.

[27] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, IEEE, 2019, pp. 1–5.

[28] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *J Healthc Eng*, vol. 2022, 2022.

[29] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews," *Electronic Commerce Research*, pp. 1–21, 2022.

[30] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Overfitting, Model Tuning, and Evaluation of Prediction Performance," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, O. A. Montesinos López, A. Montesinos López, and J. Crossa, Eds., Cham: Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.