




Optimizing IndoBERT for Revised Bloom's Taxonomy Question Classification Using Neural Network Classifier

Lazuardy Syahrul Darfiansa ^{1)*} , Fitriyani ²⁾ , Sza Sza Amulya Larasati ³⁾ 

¹⁾²⁾Telkom University, Bandung, Indonesia

¹⁾lazuardysyahrul@telkomuniversity.ac.id, ²⁾fitriyani@telkomuniversity.ac.id

³⁾Universitas Brawijaya, Malang, Indonesia

³⁾szaszaal@student.ub.ac.id

Abstract

Background: A major challenge in education is the dominance of exam questions that primarily assess basic thinking skills, such as remembering and understanding. Revised Bloom's Taxonomy (BT), which classifies cognitive skills into six levels, offers a framework to promote higher-order thinking through better-designed assessments. Deep learning-based systems have shown promising results in automatically classifying questions by BT levels, supporting educators in creating more meaningful exams.

Objective: This research aims to develop a classification system that can effectively classify Indonesian exam questions based on BT using IndoBERT pretrained models. These models were combined with Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) classifiers (referred to as IndoBERT-CNN and IndoBERT-LSTM) to determine the model with the highest performance.

Methods: The dataset utilized was self-collected and underwent several stages of preparation, including expert labeling and splitting. Furthermore, preprocessing was conducted to ensure the dataset was consistent and free from irrelevant features related to case folding, tokenization, stopword removal, and stemming. Hyperparameter fine-tuning was subsequently carried out on IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM. Model performance was evaluated using Accuracy, F-Measure, Precision, and Recall.

Results: The fine-tuned IndoBERT model results showed that IndoBERT-LSTM outperformed IndoBERT-CNN. The optimal hyperparameter configuration, batch size of 64 and learning rate of 5e-5, showed the highest performance, achieving Accuracy of 88.75%, Precision of 85%, Recall of 88%, and F-Measure of 86%.

Conclusion: IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM reflected promising results, although the performance of the models was significantly affected by respective architectures and hyperparameter settings. IndoBERT-LSTM achieved the highest accuracy with larger batch sizes, while IndoBERT and IndoBERT-CNN performed best under different configurations. However, IndoBERT faced limitations due to its language-specific focus and the limited interpretability of predictions against expert-labeled data.

Keywords: Bloom's Taxonomy, CNN, Hyperparameter Fine-Tuning, IndoBERT, LSTM, Question Classification

Article history: Received 23 November 2024, first decision 15 Mei 2025, accepted 17 June 2025, available online 22 July 2025

I. INTRODUCTION

Critical thinking and analytical ability are essential skills that should be cultivated, specifically in the current education system. However, many students in Indonesia have been observed to remain confined to simplistic thinking patterns that rely on binary logic, such as "if A, then Z," without considering alternative perspectives. For instance, a common misconception in parenting, namely "not forcing children means spoiling them," reflects limited higher-order thinking. This emphasizes deficiencies in advanced cognitive skills, which are systematically measured using the six levels of Bloom's Taxonomy (BT).

According to BT levels, manually classifying learning objectives requires not only a thorough understanding of the taxonomy but also the interpretive ability to accurately assess each objective [1], [2], [3], [4]. As stated in a previous research, the process can become overwhelming when utilizing large datasets [1], [2], [3], [4], and it is susceptible to human bias, as different individuals may interpret the same objective differently based on respective personal experiences or perspectives [5], [6]. These inconsistencies can undermine the reliability of classifications and, as a

* Corresponding author

result, impede the quality of educational assessments. These challenges emphasize the growing need for machine learning to automate question classification, helping educators accurately and efficiently categorize exam questions based on BT [7], [8], [9]. IndoBERT is becoming a method that can be leveraged in this classification automation process because IndoBERT has mechanisms to understand the context and complexity of text in exam questions. The question text is categorised based on keywords, the meaning and relationship between concepts are considered [1], [10], [11], [12]. Learning effectiveness can be increased by providing exam questions according to cognitive levels and reducing the time taken by educators to categorize questions manually.

Earlier research introduced utilizes classical classification approaches such as Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB), as well as the IndoBERT-trained language model to improve classification accuracy based on BT [13]. This method is expected to help teachers categorize questions automatically so that the prepared questions can be measured more according to the student's cognitive level. Experiments were conducted using 670 questions covering mathematics and Indonesian subjects from primary to secondary school levels. The model was trained and evaluated on the two types of questions with the results showing that the IndoBERT method provided the highest classification accuracy, 82% for math questions and 63% for Indonesian language questions. Question classification was also initiated in the research [14] with a fine-tuned IndoBERT model designed explicitly for classifying Indonesian exam questions based on BT. Dissimilar to traditional models, this model leverages the pretrained IndoBERT optimized to capture cognitive categorization nuances, which often need to catch up in capturing complexities. The experimental dataset consists of 449 Indonesian customized multiple-choice questions. The experimental results show high model performance, achieving 97% accuracy. Other research have investigated and compared the ability of various DL and ML techniques to classify questions according to BT [15]. The four models analyzed in this research include Long-Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), Random Forest (RF), and K-nearest neighbor (KNN). The research methodology involved developing and training each model using a set of questions categorized by BT levels. After the training phase, each model was evaluated based on its accuracy in classifying the questions to the correct category. The results showed that the LSTM model provided the best performance with an accuracy of 83%, followed by the RNN with 74%. Machine learning techniques RF and KNN, achieved 68% and 35% respectively.

Another previous research proposed LSTM-based DL model to classify Course Learning Outcomes (CLOs) and assessment items into different cognitive levels of BT, using English language question case studies [14]. This model predicted Bloom levels for CLOs and assessment items separately. Compared to other DL models in the literature, it features a simpler architecture and achieves a classification accuracy of 87%. Other explorations have explored both non-contextual word embedding approaches, such as Word2Vec, GloVe, and FastText, and contextual embedding methods, including BERT, RoBERTa, and ELECTRA, using two distinct datasets [16]. The experimental results reflected that FastText outperformed others on the first dataset, while RoBERTa performed best on the second. Interestingly, the results from the first dataset diverged from typical expectations in text classification, where contextual embeddings often surpass non-contextual ones. This research adopted the use of Convolutional Neural Network (CNN) instead of RNN, based on the premise that in exam question classification, extracting relevant features is more important than capturing data sequence patterns. The proposed CNN model achieved accuracy of approximately 86%.

Regardless of the recognized importance of critical and analytical thinking in education, the majority of Indonesian students were observed to continually operate in basic reasoning frameworks, reflecting a low mastery of higher-order thinking skills. BT remains the primary framework for measuring and cultivating cognitive abilities, although manually classifying questions according to its levels requires an in-depth understanding and is often subject to personal interpretation and bias. To address these challenges, various ML and DL approaches have been introduced, including SVM, Naïve Bayes (NB), CNN, LSTM, and IndoBERT, a model specifically trained to process Indonesian language. Although these methods have shown promising results, most existing investigations have predominantly focused on evaluating individual models or conducting isolated comparisons. Few have integrated the rich contextual representations of IndoBERT with external classifiers to enhance performance. Moreover, there is a scarcity of research that develops automatic classification systems specifically to classify Indonesian exam questions, specifically those reflecting the full spectrum of BT levels. In response to the gaps, this research proposes a novel classification model designed to categorize Indonesian senior high school exam questions based on BT levels C1–C6. The approach combines IndoBERT with CNN and LSTM architectures, enabling the model to explicitly consider the cognitive demands of each question. It is also important to state that hyperparameter tuning will be implemented to further improve classification performance.

II. LITERATURE REVIEW

BT serves as a conceptual framework for identifying thinking competencies, ranging from the lowest to the highest cognitive levels. This framework was originally developed to consist of six levels in the cognitive domain. The first three levels were categorized as Lower Order Thinking Skills (LOTS), while the latter three represent Higher Order Thinking Skills (HOTS) [17], [18], [19]. In accordance with the hierarchical nature of the taxonomy, students are expected to first master LOTS before progressing to HOTS [20], [21]. The six levels in the original cognitive domain are defined as follows, Knowledge (C1), which refers to the ability to recall or restate previously learned material. Understanding (C2) includes interpreting and explaining content based on inherent comprehensive ability, such as the ability to follow instructions or solve problems. Application (C3) is the capacity to apply concepts in new contexts. Analysis (C4) refers to the ability to deconstruct information into components to understand the interrelationships existing between each component. Synthesis (C5) includes generating new structures or ideas by combining existing elements. Evaluation (C6) is centered on the ability to make judgments or assessments based on specific standards or criteria. In response to developments in cognitive psychology, BT was revised in the 1990s by Lorin W. Anderson and David R. Krathwohl, both of whom were students of Bloom [22], [23], [24], [25]. Accordingly, the revised taxonomy was formally published in 2001 and introduced several changes. A key change among these changes includes the fact that the names of the cognitive levels were revised to reflect active verbs, and the positions of levels C5 and C6 were swapped. The revised cognitive levels include Remember (C1), Understand (C2), Apply (C3), Analyze (C4), Evaluate (C5), and Create (C6).

To effectively achieve the objectives of this research, a comprehensive literature review was conducted. Previous research on question classification, such as [13], explored the use of ML models, specifically SVM and NB, to classify exam questions at the secondary school level. Another research [14] applied BERT to classify multiple-choice questions based on BT levels, although the classification was limited to levels C1 through C4. Additionally, word embedding approaches have been adopted to classify English-language questions in related explorations [16].

TABLE 1
LITERATURE REVIEW COMPARISON

Author	Objective	Language	Dataset Size	Method	Accuracy
Wei, et. al [26]	Text Classification	English	20000	LSTM-SN	[78.64%]
Hasmawati, et. al. [13]	Questions Classification	Indonesian	670	SVM NB	[82%] [63%]
Ran Li, et. al [27]	Text Classification	Chinese	125064	BERT + CNN BERT + LSTM	[84.92%] [84.98%]
Darfiansa, et. al. [15]	Questions Classification	Indonesian	1322	LSTM	[83%]
Umer, et. al [28]	Text Classification	English	300000	CNN	[83%]
Gani, et. al. [16]	Questions Classification	English	2522	CNN	[86%]
Shaikh, et. al [14]	Questions Classification	English	829, 600	LSTM	[87%]
Gang Dou, et. al [29]	Text Classification	English	50000	LSTM	[88.58%]
Zhai, et. al [30]	Text Classification	Chinese	65000	CNN + LSTM	[93.41%]
Chen, et. al [31]	Text Classification	Chinese	15200	BERT + CNN	[96.2%]
Baharudin and Naufal [1]	Questions Classification	Indonesian	449	IndoBERT	[97%]

The application of CNN in text classification tasks has been widely explored through various experiments and analyses. For instance, a previous research [28] explored this topic by combining CNN with fasttext using publicly available datasets, namely Amazon. This research showed that CNN combined with fast text was very promising. Text classification on Chinese online news has long texts and complex structures, reducing the accuracy of text classification [31]. To improve the accuracy of long text classification, the present investigation proposes a model that combines CNN with BERT to overcome the limitations of the BERT model. The CNN model captures local features such as keywords so that the resulting feature vector can be used to predict text categories. The results of this research show that the combination of BERT and CNN can improve the accuracy of long texts. Another research related to text

classification conducted by [30] proposed a combination of CNN and LSTM, using CNN to obtain the features of text sentences and LSTM to capture the context of the text. The result vectors were combined to form a new feature vector, and then the softmax layer was used for classification. The results of the research showed that the text classification performance of the combination of these two models were significantly improved. LSTM as a classification text is also used [26], and the LSTM-SN method was proposed to deal with complex data by condensing the characteristics of the data set. The experimental results showed that the model could improve the accuracy of text classification. The effectiveness of the LSTM model for performing text classification tasks on IMDB data produced an accuracy of 88.58% [29]. Classification of long texts using traditional methods has insufficient feature extraction capabilities, leading to weaker classification effects than plain text [27]. As previously stated, this research proposed a combination of two models, BERT- CNN and BERT-LSTM. Baseline performance of BERT model achieved an accuracy of 82.73%, BERT-CNN showed an accuracy of 84.92% and BERT-LSTM produced the highest, at 84.98%. These enhancements were attributed to improved feature extraction and faster convergence during training. However, the imbalanced distribution of training data was found to influence the overall accuracy, suggesting that a more balanced dataset would further enhance model performance by improving generalization across all classes and yielding more reliable classification results.

Compared to traditional text classifiers, the application of CNN in text classification tasks has shown very significant effectiveness. CNN alone achieved accuracy of up to 86% in news text classification using Amazon dataset [28], while LSTM reached accuracy of 88.58% when applied to IMDB review classification [29]. Invariably, this literature review emphasizes the superior performance of hybrid models, which has been reported to consistently outperform the use of CNN or LSTM individually [27], [30], [31]. These combined models have shown significant improvements in classification accuracy and training efficiency. The results of previous research are summarized in **Table 1**, which presents a comprehensive overview of the literature reviewed in this research.

III. METHODS

This present research proposed a method for optimizing a pretrained IndoBERT model [32] for exam question classification by integrating CNN [33] and LSTM [34] architectures. The approach comprised several key stages, including data preparation, data preprocessing, hyperparameter fine-tuning, model training, testing, and comprehensive performance evaluation. A visual representation of the complete research methodology is presented in Fig. 1.

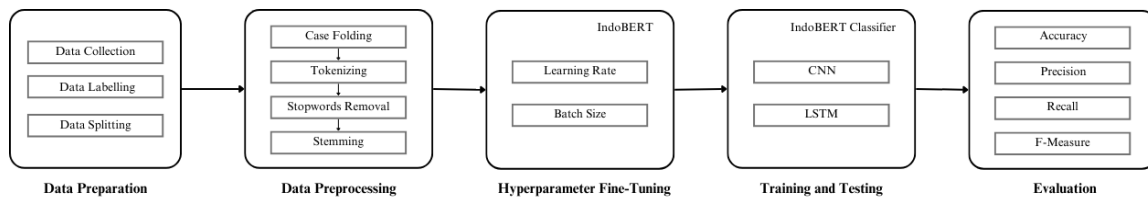


Fig. 1 Research Methodology

A. Data Collection and Labelling Process

This research was conducted using primary data obtained directly from educational documentation written in Indonesian. The data were sourced from teachers and Student Worksheets (Lembar Kerja Siswa, LKS) across various subjects at the secondary school level, including computer networking, software engineering, biology, physics, and mathematics. The dataset was organized in tabular format with two primary columns, namely question and cognitive level (label). Each question was manually labeled by an expert teacher into one of six cognitive levels including C1 (Remember), C2 (Understand), C3 (Apply), C4 (Analyse), C5 (Evaluate), and C6 (Create), based on the cognitive level keywords defined in BT and as outlined in Table 2. A total of 1,772 questions were compiled for the exploration, and the dataset utilized was split using an 80:20 ratio, with 80% used for training and 20% for testing. The distribution of each class across the dataset is shown in Fig. 2.

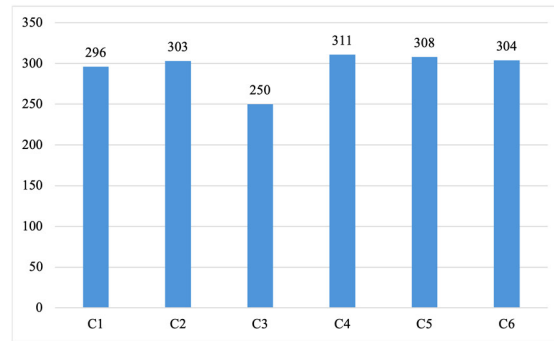


Fig. 2 Data Distribution

TABLE 2
EXPLANATION OF BLOOM'S TAXONOMY LABELS

Label	Explanation
C1	Students recall facts, terms, or concepts that have been learned.
C2	Students explain concepts or information in their own words.
C3	Students use their knowledge to solve a problem.
C4	Students decompose information into parts and see the relationship.
C5	Students assess information or situations based on certain criteria.
C6	Students create original products based on their knowledge.

B. Hyperparameter Fine-Tuning

IndoBERT was adopted as the baseline model for the question classification task in this research. As a pretrained model developed on extensive Indonesian text corpora, it was fine-tuned to extract contextual embeddings that captured the linguistic intricacies specific to Indonesian language. Hyperparameter fine-tuning was applied to improve the performance of the pre-trained IndoBERT model in identifying patterns related to BT levels. This process focused on optimizing batch size and learning rate that significantly influence model performance and stability in text classification tasks. Furthermore, experiments were conducted to identify the best configuration by testing various values for these parameters, as summarized in Table 3. Fine-tuning IndoBERT as pre-trained model on Indonesian language data may allow the model to adapt more effectively to specific classification tasks associated with BT levels. This optimization process enhances the model's ability to capture subtle patterns necessary for accurately classifying questions based on cognitive levels, thereby improving overall classification performance.

TABLE 3
HYPERPARAMETER FINE-TUNING CONFIGURATION

Hyperparameter	Value
Learning rate	5e-5, 3e-5, 1e-5
Batch size	32, 64

C. Model Training

The classification process based on BT levels was carried out using the fine-tuned IndoBERT model. Each fine-tuned model was combined with either CNN and LSTM (hereinafter referred to as IndoBERT-CNN and IndoBERT-LSTM). In the IndoBERT CNN model, the embedding results were passed to the convolutional layer to identify spatial patterns and proceed to dropout. In the IndoBERT LSTM combination, the embedding results are passed to the bidirectional LSTM. The architectures of IndoBERT-CNN and IndoBERT-LSTM are presented in Figure 3. Accordingly, the main parameters adopted in the development of IndoBERT-based classification models, including those for CNN and LSTM layers, were set using default values, as summarized in Table 4.

TABLE 4
PARAMETER INDOBERT, CNN, AND LSTM

Model	Parameter	Value
IndoBERT	Dropout Rate	0.3
	Number of filters	[2, 3, 4]
	Number of units in the fully connected layer	100
LSTM	Dropout rate	0.5
	Max Len	128
	Hidden Size	256
	Num Layers	2
	Dropout Rate	0.3

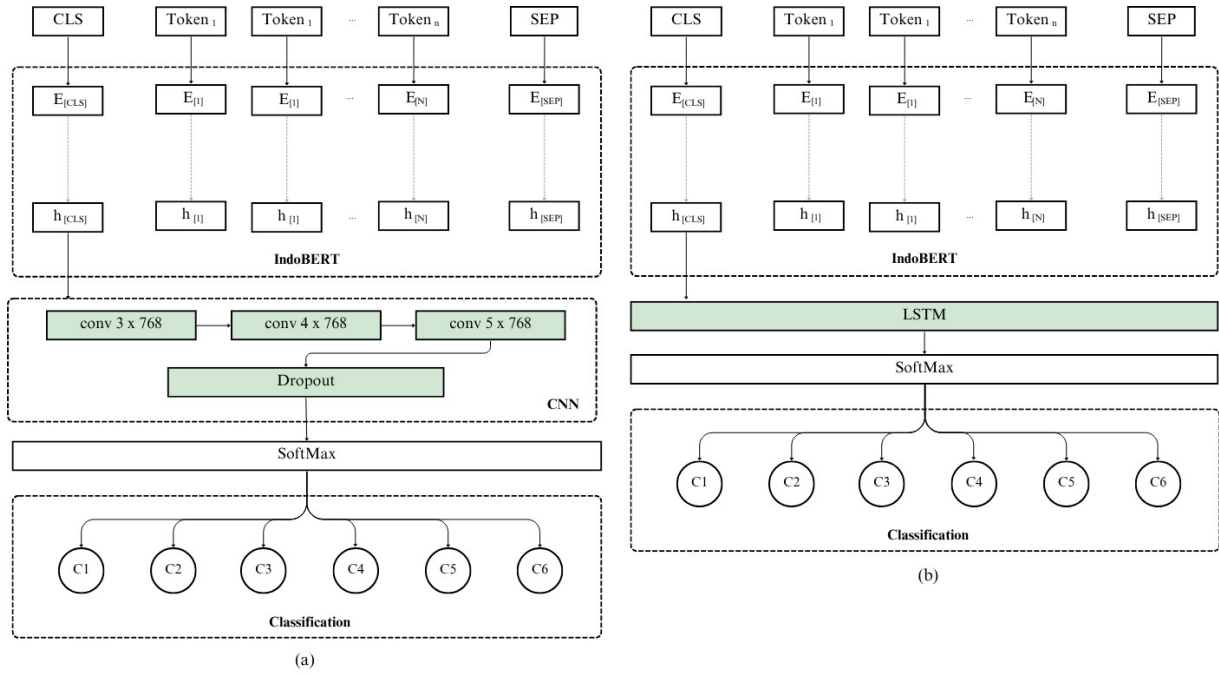


Fig. 3 (a) IndoBERT-CNN Architecture (b) IndoBERT-LSTM Architecture

D. Evaluation

The performance of the developed classification models was evaluated using four key metrics, namely accuracy, precision, recall, and F1-score. Accuracy was measured using Formula (1), which calculated the ratio of correctly predicted instances, categorized under true positives (TP) and true negatives (TN), to the total number of predictions, including false positives (FP) and false negatives (FN). Precision, calculated using Formula (2), measures the proportion of correctly predicted positive instances (TP) out of all instances predicted as positive (TP + FP). Recall, as shown in Formula (3), assesses the ability of the model to correctly identify actual positive instances, calculated by dividing TP by the sum of TP and FN. Lastly, F1-score was computed using Formula (4), which represented the harmonic mean of precision and recall.

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

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

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

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

IV. RESULTS

A. Question Label

In this research, the data were compiled from a collection of exam questions organized in a Microsoft Excel file format. The dataset was constructed through a manual labeling process, in which each question was assigned a cognitive level label based on BT categories. Examples of labeled questions representing each category are presented in Table 5. The word clouds representing classes C1 to C6, as shown in Figure 4 show the distribution of frequently occurring keywords found in exam questions or instructional prompts, all of which were categorized according to the six cognitive levels of BT. The term *technology* appears frequently in classes C1 to C3, probably due to the nature of the topics covered in the dataset.

TABLE 5
QUESTION AND LABEL

Question (Bahasa)	Question (English)	Label
Sistem operasi yang hanya terdapat pada komputer jenis Apple adalah	The operating system found only on Apple computers is	C1
Berikut yang bukan merupakan perangkat masukan/input device adalah	The following is not an input device is	C1
Apa saja teknologi yang digunakan dalam dunia industri kreatif untuk pembuatan animasi 3D?	What are the technologies used in the creative industry to create 3D animation?	C2
Apa perbedaan antara jaringan komputer kabel dan nirkabel	What is the difference between wired and wireless computer networks	C2
Bagaimana cara menghitung waktu tempuh perjalanan berdasarkan jarak dan kecepatan kendaraan?	How to calculate travel time based on distance and vehicle speed?	C3
Bagaimana cara menghitung persediaan barang berdasarkan data penjualan dan tingkat permintaan?	How to calculate inventory based on sales data and demand?	C3
Jelaskan bagaimana arus listrik dapat dianalisis melalui hukum Kirchhoff dalam rangkaian listrik.	Explain how electric current can be analyzed through Kirchhoff's law in an electric circuit.	C4
Bagaimana mekanisme kerja pompa jantung dapat dianalisis melalui prinsip fisika?	How can the working mechanism of a heart pump be analyzed through the principles of physics?	C4
Bandingkan cara kerja fotosintesis pada tumbuhan dengan respirasi seluler. Apa hubungan antara kedua proses tersebut?	Compare how photosynthesis works in plants with cellular respiration. What is the relationship between the two processes?	C5
Apa dampak dari penebangan hutan terhadap perubahan iklim? Apa langkah yang bisa diambil untuk mengatasi masalah ini?	What is the impact of deforestation on climate change? What steps can be taken to address this problem?	C5
Buatlah diagram untuk menjelaskan siklus nitrogen dan dampaknya terhadap ekosistem.	Create a diagram to explain the nitrogen cycle and its impact on the ecosystem.	C6
Buatlah analisis untuk menentukan pengaruh mode transportasi terhadap emisi karbon.	Create an analysis to determine the effect of transportation modes on carbon emissions.	C6

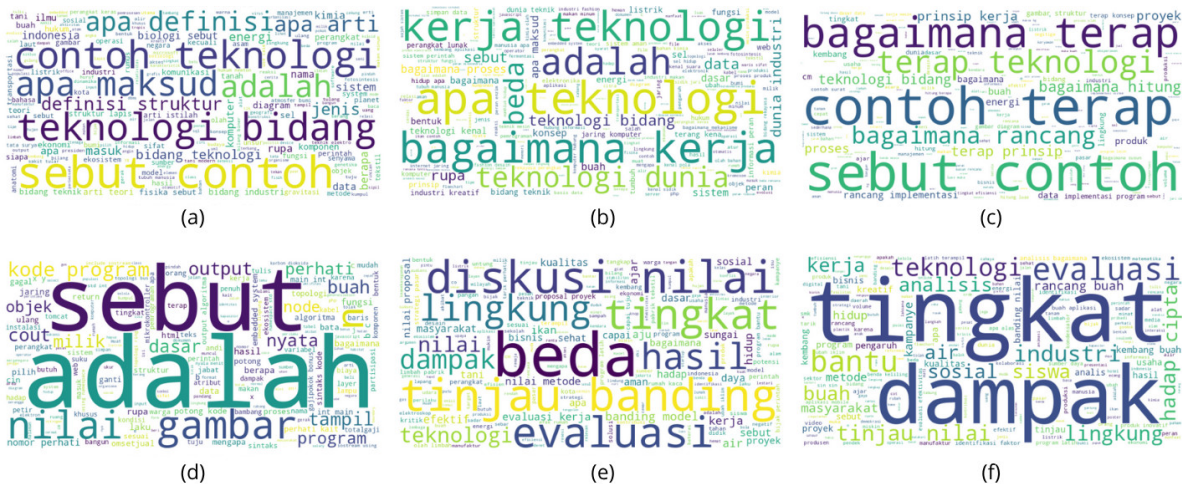


Fig. 4 Word Cloud of Question

As shown in Figure 4(a), dominant keywords namely *sebut* (mention), *adalah* (is), *apa maksud* (what), and *definisi* (definition) represented Class C1 (Remember), which invariably reflect activities associated with recalling or recognizing basic facts. Class C2 (Understand) features terms including *bagaimana* (how), *kerja* (work), and *adalah* (means), as presented in Figure 4(b), signifying comprehension of concepts and explanation of processes. Accordingly, Figure 4(c), representing Class C3 (Apply), emphasizes keywords such as *terap* (apply), *contoh* (example), and *rancang* (design), pointing to the application of knowledge in practical or new contexts. Class C4 (Analyze) includes words namely *sebut* (mention), *adalah* (is), *nilai* (value), and *gambar* (draw), as shown in Figure 4(d). These words are associated with examining structure and relationships in information. Moving to Figure 4(e), the word cloud for Class C5 (Evaluate) shows a dominance of terms, namely *diskusi* (discussion), *evaluasi* (evaluate), *tinjau* (review), and *banding* (compare), all of which pertain to the important assessment and comparison of information. Lastly, Figure 4(f) for Class C6 (Create) contains keywords such as *cipta* (create), *dampak* (impact), and

tingkat (stage), reflecting higher-order cognitive tasks pertaining to creation, solution design, and the development of structured argument. These patterns of word distribution underscore the linguistic correspondence between operational verbs and the corresponding cognitive levels in BT.

B. Training and Testing Result

This section presents the classification results obtained from experiments using three model configurations, namely IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM. Each model was examined using two different batch size values and three learning rate combinations, leading to a total of 18 test scenarios. The evaluation results of each model configuration are shown in Fig. 5.

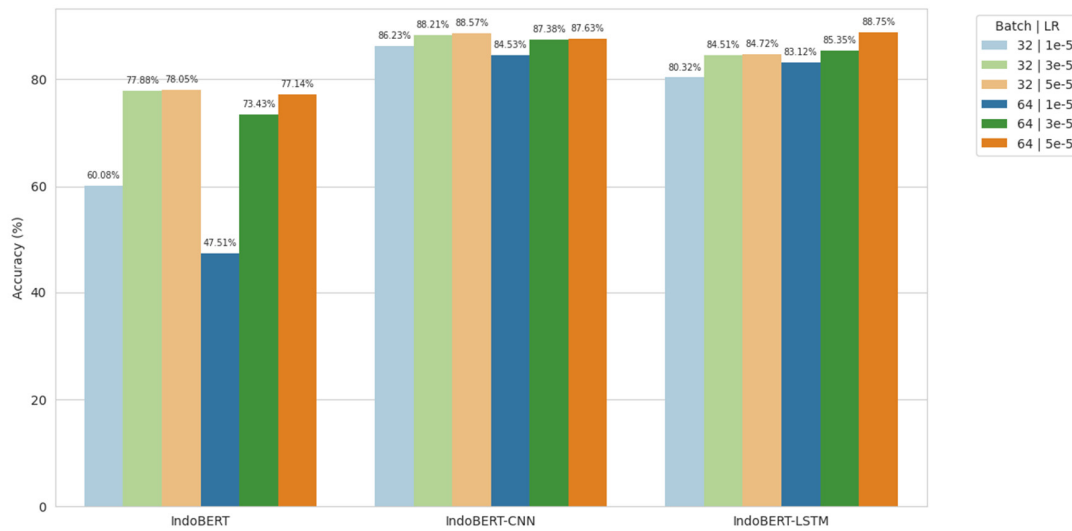


Fig. 5 Model performance

IndoBERT model achieved its lowest accuracy of 60.08% when trained with batch size of 32 and learning rate of 1e-5. This was attributed to the fact that learning rate of the model was significantly small, resulting in slow convergence. Increasing batch size to 64 while maintaining the same learning rate further decreased accuracy to 47.51%, possibly because the combination of a large batch size and low learning rate led to insufficient weight updates. Accordingly, performance was observed to significantly improve when learning rate was raised to 3e-5. At this rate, batch size of 32 produced accuracy of 77.88%, signifying a better balance between learning speed and weight update stability. However, increasing batch size to 64 with the same learning rate led to a slight decline in accuracy to 73.43%, though the result remained relatively effective. The highest performance was achieved with learning rate of 5e-5 and batch size of 32, reaching accuracy of 78.05%. This configuration allowed the model to learn more efficiently and adaptively. Meanwhile, using batch size of 64 at the same learning rate led to a slight drop in accuracy to 77.14%, probably due to the model's reduced responsiveness to rapid updates at larger batch sizes. In its entirety, training with batch size of 32 consistently outperformed training with batch size of 64, suggesting that smaller batch sizes facilitated more frequent parameter updates and improved the model's generalization ability. The optimal hyperparameter configuration for IndoBERT was therefore determined to be batch size of 32 and learning rate of 5e-5, which produced accuracy of 78.08%, precision of 75%, recall of 80%, and F1-score of 77%.

In the case of IndoBERT-CNN, the lowest accuracy recorded was 86.23% using batch size of 32 and learning rate of 1e-5. This low accuracy was attributed to the insufficient learning rate, causing slow optimization. When batch size was increased to 64 with the same learning rate, accuracy slightly decreased to 84.53%, possibly due to diminished learning dynamics from less frequent updates. Accordingly, with learning rate of 3e-5, accuracy improved to 88.21% at batch size 32, while increasing batch size to 64 caused a minor decrease to 87.38%, regardless of the fact that performance remained robust. The best performance was achieved with batch size of 32 and learning rate of 5e-5, resulting in accuracy of 88.57%, showing that this configuration supports faster and more effective learning. At the same learning rate but with batch size of 64, accuracy slightly decreased to 87.63%, signifying the model's reduced adaptability with larger batches. In essence, a smaller batch size consistently produced better results, probably due to more frequent weight updates that enabled finer gradient tracking and improved generalization. Based on these results,

the optimal IndoBERT-CNN configuration for classifying exam questions according to BT levels is batch size 32 and learning rate 5e-5, achieving accuracy of 88.57%, precision of 83%, recall of 89%, and F1-score of 86%.

IndoBERT-LSTM showed an opposite performance trend compared to the other models. The combination of batch size of 32 and learning rate of 1e-5 produced the lowest accuracy of 80.32%. However, increasing batch size to 64 while maintaining the same learning rate led to an improvement in accuracy to 83.12%. This suggests that a larger batch size may enhance model stability, even under a limited learning rate. Further improvement was observed when learning rate was increased to 3e-5, resulting in accuracies of 84.51% for batch size 32 and 85.35% for batch size 64. Using batch size of 32 and learning rate of 5e-5, the model achieved a slightly higher accuracy of 84.72%, reflecting stable learning at a higher rate. The best performance was achieved using batch size of 64 combined with learning rate of 5e-5, which produced the highest accuracy of 88.75%, along with precision of 85%, recall of 88%, and F1-score of 86%.

CNN architecture excelled at extracting local spatial patterns in text by focusing on key phrases or word groupings through convolutional operations. Meanwhile, LSTM architecture was particularly effective in identifying sequential dependencies and long-term relationships between words. Regardless of the observation that IndoBERT-CNN achieved a high accuracy of 88.57% using batch size of 32 and learning rate of 5e-5, IndoBERT-LSTM slightly outperformed it with accuracy of 88.75% using batch size of 64 at same learning rate. The ability of LSTM to capture richer contextual information and preserve word order probably contributed to its superior performance in classifying complex question structures. A summary of the highest accuracy, precision, recall, and F1-score achieved by each model is presented in Table 6.

TABLE 6
MODEL PERFORMANCE COMPARISON

Model	Precision	Recall	F-Measure	Accuracy
IndoBERT	75%	80%	77%	78.08%
IndoBERT-CNN	83%	89%	86%	88.57%
IndoBERT-LSTM	85%	88%	86%	88.75%

V. DISCUSSION

IndoBERT, *IndoBERT-CNN*, and *IndoBERT-LSTM* achieved accuracy of 52.45%. Although this model is capable of contextual understanding, it appeared to struggle when applied directly to classification tasks based on Revised BT. Dissimilar, LSTM model referenced in [20] performed slightly better, producing accuracy of 55.79%. As a sequence-based model, LSTM captures time-dependent relationships in data, which may benefit tasks related to word order. However, this model still lacks a comprehensive understanding of complex semantic relationships, suggesting the need for a context-aware model such as IndoBERT. A significant improvement was observed when using CNN alone, which reached approximately 86% accuracy. The ability of CNN to detect spatial features such as n-gram patterns and localized word structures proves advantageous in text classification, even without modeling sequential dependencies. This shows how local feature extraction can strongly influence performance.

TABLE 7
MODEL COMPARISON

Author	Method	Hyperparameter	Accuracy
Baharudin and Naufal [1]	IndoBERT	Learning Rate: 2e-5 Batch: 32 Dropout: 0.3	[52.45%]
Gani, dkk [16]	CNN	Learning Rate: 1e-5 Batch Size: 16 Dropout: 0.2	[86.12%]
Shaikh, dkk [14]	LSTM	Learning Rate: 1e-5 Batch Size: 16 Dropout: 0.2	[55.79%]
Darfiansa, dkk [15]	LSTM	Learning Rate: 1e-5 Batch Size: 128 Dropout: 0.3	[77.25%]
The proposed model in this research	IndoBERT-CNN	Learning Rate: 5e-5 Batch Size: 32 Dropout: 0.5	[88.57%]
	IndoBERT-LSTM	Learning Rate: 5e-5 Batch Size: 64 Dropout: 0.3	[88.75%]

Considering the strengths, hybrid models such as IndoBERT-CNN and IndoBERT-LSTM produced the most promising outcomes. IndoBERT-LSTM achieved the highest accuracy at 88.75%, closely followed by IndoBERT-CNN at 88.57%. These hybrid approaches leveraged the contextual language modeling of IndoBERT alongside the spatial sensitivity of CNN and sequence awareness of LSTM. By combining these architectural benefits, the models became better equipped to manage the cognitive complexity inherent in Bloom-level classification. In entirety, IndoBERT-LSTM model outperformed all standalone models, signifying that a hybrid architecture can deliver a more comprehensive representation of linguistic and sequential patterns. Accordingly, IndoBERT-LSTM effectively integrated context and dependency tracking, making it particularly adept at complex classification tasks. Regardless of the fact that IndoBERT-CNN typically benefits from fast convergence and gradient stability, particularly through its localized pattern recognition, the model may fall slightly short in capturing longer-range dependencies, thereby resulting in marginally lower accuracy.

Despite showing promising performance, this research has a significant limitation. The limitation includes the fact that the exploration did not evaluate the degree of correspondence between the predictions made by the model and the expert-assigned labels. This correspondence could be more rigorously assessed using statistical reliability measures such as Intraclass Correlation Coefficient (ICC) [35]. Although the labeling was conducted by subject matter experts to enhance the validity of the utilized dataset, the absence of a quantitative reliability check made it difficult to fully interpret how well the model replicates expert judgment. The use of ICC would provide deeper insight into the consistency between machine predictions and human classification, an aspect which is specifically relevant in tasks based on BT, where the distinction between cognitive levels can be subjective and comprehensive. From the model's perspective, its most prominent limitation is in the architectural complexity, which was built upon fine-tuning IndoBERT with relatively basic classifiers such as CNN or LSTM. Although effective for many text classification tasks, this architecture did not fully capture semantic depth or the intricate cognitive distinctions found in borderline-level questions, particularly those that straddle two adjacent cognitive categories, where misclassifications frequently occur. Lastly, IndoBERT is language-specific, designed exclusively for Indonesian language [1], which limits its transferability. The model cannot be directly applied to texts in other languages without retraining or substituting it with an equivalent pre-trained model suited to the target language.

VI. CONCLUSIONS

In conclusion, this research successfully provides valuable insights into the development of BT-based question classification system using IndoBERT. This was carried out by evaluating how architectural combinations and hyperparameter configurations influenced model performance on Indonesian-language data. In accordance with the observations made, classification of exam questions based on BT using IndoBERT baseline was observed to produce promising results, specifically with IndoBERT-CNN and IndoBERT-LSTM variants. This investigation systematically showed that each architecture responded differently to hyperparameter tuning. For instance, IndoBERT performed most stably with smaller batch sizes and moderated learning rates due to the nature of its deep attention layers. IndoBERT-CNN was observed to reach optimal performance at higher learning rates by leveraging its ability to capture local textual features, while IndoBERT-LSTM benefited from larger batch sizes, which helped stabilize gradient updates and allow better utilization of its sequential modeling strengths. Among the three observed models, IndoBERT-LSTM achieved the highest classification accuracy of 88.75%, using learning rate of $5e-5$ and batch size of 64. Moreover, the integration of CNN and LSTM as classifiers in IndoBERT framework significantly improved performance by reducing misclassification rates and enhancing accuracy of the model. Despite the achieved accuracy, the model remained constrained by its language specificity, as IndoBERT is trained exclusively on Indonesian data and cannot be directly applied to other languages without retraining or adaptation. Another limitation observed is the absence of reliability analysis between the model's predictions and expert-assigned labels, which restricted deeper interpretation of the model's ability to replicate human cognitive judgments. Future research may benefit from applying statistical methods such as ICC to measure the degree of agreement between machine predictions and expert classifications. Despite the outlined limitations, the primary strength of this research lies in its detailed exploration of the interaction between model architecture, hyperparameter design, and classification performance, particularly using Indonesian primary data, a contribution that has not been adequately emphasized in earlier research. The results from this investigation offer a meaningful advancement in the development of contextualized, NLP-based classification systems, with practical implications for the field of education.

Author Contributions: Lazuardy Syahrul Darfiansa: Conceptualization, Methodology, Writing - Original Draft, Supervision. Fitriyani: Investigation, Review & Editing. Sza Sza Amulya Larasati: Data Curation, Analysis, 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 research was collected independently and is not publicly accessible.

Informed Consent: There were no human subjects.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

ORCID:

Lazuardy Syahrul Darfiansa: <https://orcid.org/0000-0002-8983-8162>

Fitriyani: <https://orcid.org/0009-0002-7329-0769>

Sza Sza Amulya Larasati: <https://orcid.org/0009-0007-7654-0512>

REFERENCES

- [1] 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, Nov. 2023, doi: 10.20473/jisebi.9.2.253-263.
- [2] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTIS)*, IEEE, Nov. 2019, pp. 112–117. doi: 10.1109/IoTIS47347.2019.8980428.
- [3] S. K. Patil and M. M. Shreyas, "A Comparative Study of Question Bank Classification based on Revised Bloom's Taxonomy using SVM and K-NN," in *2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT)*, IEEE, Dec. 2017, pp. 1–7. doi: 10.1109/ICECIT.2017.8453305.
- [4] A. S. Callista, O. Nurul Pratiwi, and E. Sutoyo, "Questions Classification Based on Revised Bloom's Taxonomy Cognitive Level using Naive Bayes and Support Vector Machine," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, Sep. 2021, pp. 260–265. doi: 10.1109/IC2IE53219.2021.9649187.
- [5] M. İlhan and M. Gezer, "A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions," *Pegem Eğitim ve Öğretim Dergisi*, vol. 7, no. 4, pp. 637–662, Sep. 2017, doi: 10.14527/pegog.2017.023.
- [6] S. S. A. Larasati and F. A. Bachtiar, "Harnessing Residual Attention Networks for Stress Level Classification Using EEG Spectrograms," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICIC64337.2024.10956441.
- [7] Syahidah Sufi Haris and Nazlia Omar, "A rule-based approach in Bloom's Taxonomy question classification through natural language processing," *IEEE*, 2012.
- [8] S. HUILAN *et al.*, "Educational management in Critical Thinking Training Based on Bloom's Taxonomy and SOLO Taxonomy," in *2020 International Conference on Information Science and Education (ICISE-IE)*, IEEE, Dec. 2020, pp. 518–521. doi: 10.1109/ICISE51755.2020.00116.
- [9] K. Jayakodi, M. Bandara, and I. Perera, "An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy," in *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Dec. 2015, pp. 195–202. doi: 10.1109/TALE.2015.7386043.
- [10] S. F. Kusuma, D. Siahaan, and U. L. Yuhana, "Automatic Indonesia's questions classification based on bloom's taxonomy using Natural Language Processing a preliminary study," in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, IEEE, Nov. 2015, pp. 1–6. doi: 10.1109/ICITSI.2015.7437696.
- [11] 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, Mar. 2020, doi: 10.1371/journal.pone.0230442.
- [12] M. Ifham, K. Banujan, B. T. G. S. Kumara, and P. M. A. K. Wijeratne, "Automatic Classification of Questions based on Bloom's Taxonomy using Artificial Neural Network," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, Mar. 2022, pp. 311–315. doi: 10.1109/DASA54658.2022.9765190.
- [13] Hasmawati, A. Romadhony, and R. Abdurrohman, "Primary and High School Question Classification based on Bloom's Taxonomy," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2022, pp. 234–239. doi: 10.1109/ICoICT55009.2022.9914842.

- [14] S. Shaikh, S. M. Daudpotta, and A. S. Imran, "Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings," *IEEE Access*, vol. 9, pp. 117887–117909, 2021, doi: 10.1109/ACCESS.2021.3106443.
- [15] L. S. Darfiansa, F. Azzuri, F. A. Bachtar, and D. E. Ratnawati, "Comparative Analysis of Deep Learning and Machine Learning Techniques for Question Classification in Bloom's Taxonomy," in *2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC)*, IEEE, Oct. 2023, pp. 103–108. doi: 10.1109/ICONNIC59854.2023.10467502.
- [16] M. O. Gani, R. K. Ayyasamy, A. Sangodiah, and Y. T. Fui, "Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique," *Educ Inf Technol (Dordr)*, vol. 28, no. 12, pp. 15893–15914, Dec. 2023, doi: 10.1007/s10639-023-11842-1.
- [17] H. Sharma, R. Mathur, T. Chintala, S. Dhanalakshmi, and R. Senthil, "An effective deep learning pipeline for improved question classification into bloom's taxonomy's domains," *Educ Inf Technol (Dordr)*, vol. 28, no. 5, pp. 5105–5145, May 2023, doi: 10.1007/s10639-022-11356-2.
- [18] E. R. Setyaningsih and I. Listiowarni, "Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, IEEE, Apr. 2021, pp. 330–333. doi: 10.1109/EIConCIT50028.2021.9431862.
- [19] N. Barari, M. RezaeiZadeh, A. Khorasani, and F. Alami, "Designing and validating educational standards for E-teaching in virtual learning environments (VLEs), based on revised Bloom's taxonomy," *Interactive Learning Environments*, vol. 30, no. 9, pp. 1640–1652, Oct. 2022, doi: 10.1080/10494820.2020.1739078.
- [20] ANBUSELVAN SANGODIAH, ROHIZA AHMAD, and WAN FATIMAH WAN AHMAD, "TAXONOMY BASED FEATURES IN QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE," *J Theor Appl Inf Technol*, 2017.
- [21] A. Sangodiah, T. Jee San, Y. Tien Fui, L. Ean Heng, R. K. Ayyasamy, and N. A. Jalil, "Identifying Optimal Baseline Variant of Unsupervised Term Weighting in Question Classification Based on Bloom Taxonomy," *MENDEL*, vol. 28, no. 1, pp. 8–22, Jun. 2022, doi: 10.13164/mendel.2022.1.008.
- [22] L. S. Darfiansa and F. A. Bachtar, "Comparative Analysis of Term Weighting Methods for Question Classification in Bloom Taxonomy Using Machine Learning Approach," in *2023 IEEE International Conference on Computing (ICOCO)*, IEEE, Oct. 2023, pp. 259–264. doi: 10.1109/ICOCO59262.2023.10397821.
- [23] M. Mohammed and N. Omar, "Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF," *Int J Adv Sci Eng Inf Technol*, vol. 8, no. 4–2, pp. 1679–1685, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6835.
- [24] E. Subiyantoro, A. Ashari, and Suprpto, "Cognitive Classification Based on Revised Bloom's Taxonomy Using Learning Vector Quantization," in *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, IEEE, Nov. 2020, pp. 349–353. doi: 10.1109/CENIM51130.2020.9297879.
- [25] DHUHA ABDULHADI ABDULJABBAR and NAZLIA OMAR, "EXAM QUESTIONS CLASSIFICATION BASED ONBLOOM'S TAXONOMY COGNITIVE LEVEL USINGCLASSIFIERS COMBINATION," *J Theor Appl Inf Technol*, 2015.
- [26] W. Wei, X. Li, B. Zhang, L. Li, R. Damaševičius, and R. Scherer, "LSTM-SN: complex text classifying with LSTM fusion social network," *J Supercomput*, vol. 79, no. 9, pp. 9558–9583, Jun. 2023, doi: 10.1007/s11227-022-05034-w.
- [27] R. Li, W. Yu, Q. Huang, and Y. Liu, "Patent Text Classification based on Deep Learning and Vocabulary Network," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023, doi: 10.14569/IJACSA.2023.0140107.
- [28] M. Umer *et al.*, "Impact of convolutional neural network and FastText embedding on text classification," *Multimed Tools Appl*, vol. 82, no. 4, pp. 5569–5585, Feb. 2023, doi: 10.1007/s11042-022-13459-x.
- [29] G. DOU, K. ZHAO, M. GUO, and J. MOU, "MEMRISTOR-BASED LSTM NETWORK FOR TEXT CLASSIFICATION," *Fractals*, vol. 31, no. 06, Jan. 2023, doi: 10.1142/S0218348X23400406.
- [30] Z. Zhai, X. Zhang, F. Fang, and L. Yao, "Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model," *Multimed Tools Appl*, vol. 82, no. 14, pp. 20975–20988, Jun. 2023, doi: 10.1007/s11042-023-14450-w.
- [31] 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, doi: 10.1109/ACCESS.2022.3162614.
- [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," Nov. 2020.
- [33] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Dec. 2015.
- [34] C. B. Vennerød, A. Kjærø, and E. S. Bugge, "Long Short-term Memory RNN," May 2021.
- [35] V. S. Senthil Kumar and S. Shahraz, "Intraclass correlation for reliability assessment: the introduction of a validated program in SAS (ICC6)," *Health Serv Outcomes Res Methodol*, vol. 24, no. 1, pp. 1–13, Mar. 2024, doi: 10.1007/s10742-023-00299-x.

Publisher's Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.