International Journal of Advances in Data and Information Systems
Vol. 7, No. 1, April 2026, pp. 121~133
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i1.1492

Multi-Output Classification of Cognitive Levels and Topics in Indonesian Questions using Deep Learning and Transformers

Orvalamarva[1], Oktariani Nurul Pratiwi[2], Hanif Fakhrurroja[3]
[1,2,3]Information System Study Program, School of Industrial Engineering, Telkom University, Bandung, Indonesia

Article Info

Article history:
Received Des 15, 2025
Revised Jan 06, 2026
Accepted Jan 29, 2026

Keywords:
Educational assessment
IndoBERT
Multi-output text classification
Revised Bloom's Taxonomy

ABSTRACT

Managing large-scale digital question banks struggles with manual metadata labeling, especially when identifying material topics and cognitive levels based on the Revised Bloom's Taxonomy. Current automated approaches usually treat these two attributes as separate tasks, which adds to the system's complexity and computational load. This study introduces a multi-output classification method using a shared encoder architecture with two task-specific heads to predict topics and cognitive levels simultaneously. We performed experiments on 685 Indonesian junior high science questions, covering 15 topic labels and four cognitive levels (C1–C4), with an imbalanced distribution in which lower cognitive levels accounted for more than 75% of the dataset. To handle this imbalance, we applied Focal Loss to taxonomy classification, and class weighting was used in the comparison model. A comparative study involved CNN, BiLSTM, DistilBERT, and IndoBERT. Our results demonstrate that IndoBERT delivered the best performance, with F1-macro scores of 0.78 for topics and 0.71 for cognitive levels and showed better performance in minority classes compared to standard cross-entropy-based models. These findings suggest that an integrated multi-output approach can boost the efficiency and accuracy of question labeling and offers potential for integration into Computer-Based Test systems and e-assessment platforms in real time.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Orvalamarva,
School of Industrial Engineering,
Telkom University,
Jl. Telekomunikasi Terusan Buahbatu No.1, West Java 40257, Bandung, Indonesia
Email: orvalamarvao@student.telkomuniversity.ac.id

1. INTRODUCTION

The move from static paper-based tests to large-scale digital question banks has fundamentally changed educational evaluation. Modern Learning Management Systems (LMS) now house millions of questions, serving as the backbone for adaptive testing [1]. Metadata quality plays a central role in this setup. It ensures curriculum alignment, balances difficulty, and guarantees reliable evaluation. If labeling is inaccurate, these databases risk turning into unusable "data swamps" [2].

However, managing metadata remains a bottleneck because it relies heavily on manual work. Annotating thousands of items is resource intensive and demands a level of consistency that is often difficult to maintain. Subjectivity often creeps in, particularly when separating close cognitive levels like Remembering (C1) and Understanding (C2) [3]. While NLP promises automation, most current methods handle topics and cognitive levels as distinct tasks. This split creates inefficiency. It forces systems to run separate models for each attribute, driving up computational costs in production.

Earlier studies favored CNNs and LSTMs for classification. While effective, these models typically handle single outputs [4], [5], [6]. The arrival of Transformers brought BERT-based models into the field with strong results [7], [8], [9]. Even so, existing multi-task research usually links related attributes, such as knowledge dimensions [10], [11]. Work on multi-output models for independent attributes in imbalanced Indonesian datasets is still rare [12]. This gap highlights a need for architectures that capture the unique needs of both tasks. Topic classification relies on keyword recognition, while cognitive leveling demands a grasp of semantic context. Thus, we need a model that is both accurate and efficient enough for real-world digital assessment systems.

We propose a multi-output classification framework to tackle these issues, predicting both topic and cognitive level simultaneously within a single integrated model. By using a shared encoder connected to two task-specific heads, this method reduces the computational redundancy found in single-task approaches. Our study compares four architectures: CNN, BiLSTM, DistilBERT, and IndoBERT to balance the need for linguistic complexity against computational efficiency in resource-constrained settings. Beyond its technical value to Natural Language Processing, this work implies significant cross-disciplinary benefits for Information Systems and educational technology. Specifically, the proposed multi-output approach supports the design of adaptive learning evaluation systems that integrate easily into digital platforms. Ultimately, this research bridges artificial intelligence, information systems, and educational evaluation, providing a solid groundwork for more consistent, sustainable automated assessment.

Most past studies treat topic classification and cognitive level as separate tasks or depend on multi-task learning for correlated attributes. In contrast, we position both attributes as independent outputs within one architecture. Our key innovation is using shared feature representations for two tasks with distinct linguistic needs: recognizing domain-specific keywords for topics while understanding complex semantic structures for cognitive levels, tested here on imbalanced Indonesian data. Furthermore, we combine comparisons of classical and Transformer models with specific strategies to handle class imbalance. The contribution here extends beyond just better prediction performance; it also simplifies system architecture and improves efficiency, setting it apart from earlier methods. Conceptually, this advances data-driven Information Systems by merging NLP and machine learning to handle large-scale educational text data.

2. RESEARCH METHOD

The research workflow adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, recognized as one of the most widely implemented data mining methodologies due to its iterative and flexible nature [13], [14]. CRISP-DM comprises six primary stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In this study, the framework is modified to accommodate the specific architecture of multi-output classification. The technical stages of the research are illustrated in Figure 1.

2.1. Data Understanding

The dataset utilized in this study comprises a collection of Natural Science questions at the Junior High School level. Data collection was sourced from various digital education platforms and online question banks, including Pinterran, Zenius, Defantri.com, Websiteedukasi.com, and Tryout.id.

The initial raw dataset consisted of 823 items. A rigorous data curation process was implemented to ensure training quality, comprising the following steps:

1. Duplicate Elimination: Removing identical questions across different sources.
2. Topic Filtering: Excluding topics with insufficient samples to ensure the model effectively learns latent topic patterns.
3. Taxonomy Adjustment: Due to data scarcity at higher cognitive levels, questions labeled C5 (*Evaluating*, 2 items) and C6 (*Creating*, 0 items) were excluded from the dataset because they provided insufficient samples for model generalization.

Following this curation, the final dataset comprised 685 items. The distribution reveals an important class imbalance. The breakdown of data by topic category is presented in Table 1.

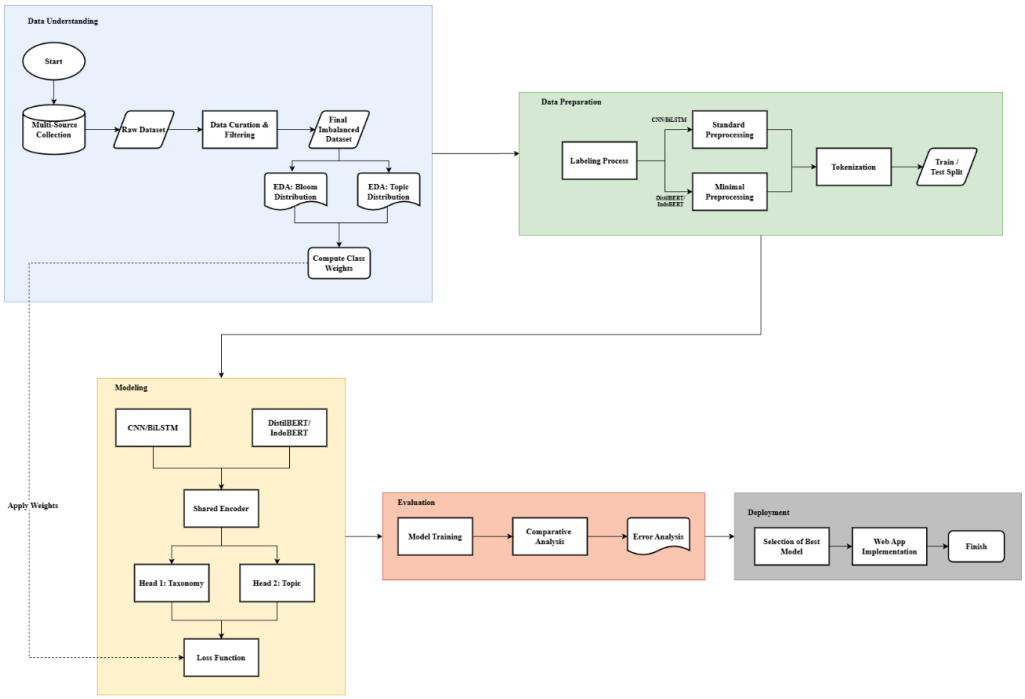


Figure 1. Research flow based on CRISP-DM framework

Table 1. Topic distribution in the dataset

Topic material	Total items
Forces	45
Food production	39
Electricity	50
Additives & addictive substances	98
Vibration	33
Science object	88
Waves	44
Classification of living things	36
Matter & its changes	41
Temperature & expansion	42
Biotechnology	23
Matter & its impact	34
Human respiratory system	26
Motion of objects	47
Work & simple machines	39

The distribution of data by cognitive complexity is presented in Table 2. The dataset is predominantly composed of lower-level questions (C1 and C2), presenting a distinct challenge for the classification process.

Table 2. Revised Bloom's Taxonomy distribution in the dataset

Cognitive level	Total items	Percentage
C1	323	47%
C2	202	29%
C3	89	13%
C4	71	11%

2.2. Data Preparation

The data preparation stage comprises two primary processes, namely labeling and text preprocessing, which are tailored to the specific model architecture.

1. Labeling Strategy

   Prior to the text preprocessing stage, data annotation was performed using an AI-assisted approach, with manual verification. The Gemini model was used as a pre-annotator to generate initial topic and cognitive-level labels, aiming to speed up the annotation process

and reduce manual workload. However, the generated labels were not used directly as final labels. All initial annotation results were manually verified and corrected by one researcher, who served as the final decision-maker in assigning labels. The verification process was carried out using written guidelines compiled based on the definition of Revised Bloom's Taxonomy [15], as well as a list of action verbs drawn from the scientific literature to distinguish cognitive levels across classes [16].

This approach was used to minimize ambiguity in distinguishing between closely related cognitive levels, such as C1 and C2. The determination of topic labels refers to the material categories available on digital learning platforms, specifically Pinterran and Zenius. Both platforms have a topic structure that is relatively consistent with the junior high school science curriculum, so the topic mapping process involved adjusting the terms and scope of material from each source. The annotation approach applied is designed to maintain consistency in the interpretation of cognitive levels and learning topics. In the context of evaluating the model architecture, the primary focus of this study, this strategy is considered adequate and relevant, while also providing room for exploring more collaborative annotation approaches in future studies.

2. Preprocessing

   This study applies two distinct preprocessing strategies (refer to Figure 1) contingent upon the model type:

   a. Standard Preprocessing

      This strategy aims to reduce feature sparsity in classic Deep Learning models. The pipeline includes text cleaning, case folding, stopword removal, and stemming utilizing the Sastrawi library.

   b. Minimal Preprocessing

      Preserving sentence structure and syntactic context is crucial for Transformer models. Consequently, stemming and stopword removal were omitted. The process is restricted to whitespace normalization and sub-word tokenization with a maximum sequence length of 128 tokens.

For the data partitioning phase, we split the dataset into training (80%) and validation (20%) sets using a fixed random_state of 42 to ensure consistency across experiments. We applied stratified sampling based on topic labels to keep the class distribution proportional. Due to the limited size of the dataset, we decided against creating a separate test set, prioritizing a larger sample size for training instead. The 20% validation subset served a dual purpose: monitoring performance to trigger Early Stopping (to prevent overfitting) and providing the basis for the final performance reports. Since these metrics reflect performance on data never used for weight updates, they offer a reliable benchmark for comparing the architectures in this study.

2.3. Modeling

At the modeling stage, a multi-output classification approach using four algorithms was implemented to predict Topics and Taxonomy concurrently. Across all tested architectures, we implemented the multi-output approach by optimizing two loss functions in parallel: one for topic prediction and one for the taxonomy level. To determine the total loss, we simply summed the two output losses (Total_loss = Topic_loss + Taxonomy_loss) using equal weights. We adopted this strategy to ensure both tasks contribute evenly to the learning process, preventing the model from prioritizing one output over the other.

1. Convolutional Neural Network

   The process began with text preprocessing, including standard tokenization and padding, followed by the transformation of target labels into One-Hot Encoding. The training and validation data were split using a stratified sampling technique to maintain class distribution. The model was constructed using a multi-output, with feature representations extracted by the Embedding and Conv1D layers. Adopting the basic CNN architecture for sentence classification, one-dimensional convolution operations were performed by shifting a filter of size h along the embedding matrix according to the equation [17]:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad (1)$$

Each word window $x_{i:i+h-1}$ is multiplied by the filter weight W, added to the bias b, and then passed to the non-linear activation function f. The result is the feature c_i which represents the local pattern (*n-gram*) at a specific position. All features generated from the filter shift form a feature map, which is then summarized using Global Max Pooling to extract the most dominant features. The final stage was performed by passing the representation to two separate output branches (Topic and Taxonomy), each with a Softmax activation. Training was conducted using the Adam optimizer with the standard Categorical Crossentropy loss function.

2. Bi-Directional LSTM
   This model is initiated with an Embedding Layer followed by a BiLSTM layer. BiLSTM is an extension of the Recurrent Neural Network (RNN) designed to capture long-term semantic dependencies while overcoming the *vanishing gradient* problem. The LSTM unit utilizes a gating mechanism (input gate i_t, forget gate f_t, and output gate o_t) to regulate the flow of information [18].

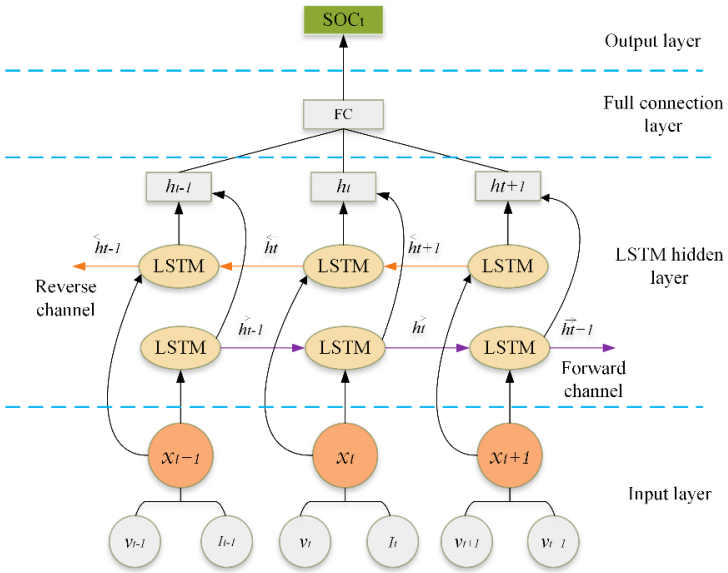


Figure 2. Structure of the BiLSTM network (adapted from [18]).

The Bi-Directional LSTM architecture is shown in Figure 2, where the text sequence is processed in two directions (forward and backward) so that the semantic context can be captured more comprehensively. The outputs from both directions are then concatenated and passed to the next layer to produce a richer feature representation. To mitigate class imbalance issues, this model integrated a Class Weighting mechanism into the Custom Weighted Categorical Crossentropy loss function. Training stability was maintained using the Adam optimizer (lr=1e-4), governed by Early Stopping and ReduceLROnPlateau.

3. DistilBERT
   The third model adopted the DistilBERT architecture as a shared encoder. DistilBERT utilizes the Multi-Head Self-Attention mechanism to capture information from different representation subspaces at different positions. Mathematically, this operation is defined as:

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head^1,\ldots,head^h)\,W^O \\ \text{where } head_i &= Attention\,(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (2)$$

Here *Q, K,* and *V* represent the *Query, Key,* and *Value*. Specifically, the model projects these matrices into several representation subspaces via different linear projections. The

results from all heads are then concatenated and projected back using the W^O weight matrix. This approach allows the model to capture various relational patterns between tokens simultaneously from multiple representation perspectives [19]. In its implementation, the contextual vector representation of the *[CLS]* token was distributed to two independent *Linear Heads*. The training process was implemented via a custom training loop using PyTorch with Mixed Precision Training (GradScaler) for GPU memory efficiency. The checkpointing mechanism was configured to save the best model based on the average *Macro-F1 score*.

4. IndoBERT

To capture the linguistic nuances of the Indonesian language, the IndoBERT base model (indobenchmark/indobert-base-p2) was utilized [20]. The novelty of this configuration lies in the application of Focal Loss to the Taxonomy output. Focal Loss is a modification of standard *Cross Entropy* with the addition of the modulating factor $(1 - p_t)^\gamma$. This factor serves to reduce the loss contribution from easy examples, allowing the model to assign higher weights to hard samples, with the equation:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log p_t \quad (3)$$

Where γ is the focusing parameter that reduces the impact of loss from easily classified examples (e.g., majority classes such as C1), compelling the model to focus more intensively on minority classes (e.g., C4) [21]. Meanwhile, the Topic output retained the use of standard *Cross Entropy*.

2.4. Evaluation

Model performance was evaluated using the Macro F1-Score metric, selected for its capability to provide a harmonic mean that is unbiased against the majority class [22]. The F1-Score is calculated as follows [23]:

$$F1 = 2.\frac{Precision.Recall}{Precision+Recall} \quad (4)$$

The F1-Score balances Precision and Recall, rendering it more informative than accuracy alone. Nevertheless, Accuracy was also recorded as a supporting metric. All experiments were executed on an NVIDIA Tesla T4 GPU. The classical Deep Learning model implementations utilized the TensorFlow/Keras framework, while the Transformer models employed PyTorch. Comprehensive hyperparameter configurations for all models are presented in Table 3.

Table 3. Hyperparameter configuration

Parameter	CNN	BiLSTM	DistilBERT	IndoBERT
Epoch	10	40	6	6
Batch Size	32	32	16	16
Learning Rate	1e-3	1e-4	3e-5	3e-5
Optimizer	Adam	Adam	AdamW	AdamW
Max Sequence	50	50	128	128

Hyperparameter selection was conducted through empirical experimentation, employing distinct strategies for each architecture type. For classical Deep Learning models, distinct training durations were observed; the CNN model achieved convergence rapidly, requiring only 10 epochs to capture local feature patterns. In contrast, the BiLSTM necessitated a prolonged training phase (40 epochs) with Early Stopping to effectively model complex sequential dependencies [24]. For Transformer-based models, a conservative fine-tuning strategy was applied adhering to recommendations from recent NLP literature. The learning rate was set minimally with a limited number of epochs (6) to mitigate *catastrophic forgetting*, a phenomenon where the model loses general linguistic knowledge due to drastic weight updates [25]. Furthermore, the AdamW optimizer was chosen over the standard Adam optimizer for Transformers to effectively manage *weight decay* in large-scale models [26].

3. RESULTS AND DISCUSSION

This chapter presents the performance evaluation results for the four proposed model architectures: CNN, BiLSTM, DistilBERT, and IndoBERT. The evaluation was conducted based on two primary metrics, Accuracy and Macro F1-Score, to measure the model's capability to classify Topics and Revised Bloom's Taxonomy concurrently.

3.1. Model Performance Comparison

A comprehensive summary of the performance of all models on the test dataset is presented in Table 4. Based on the experimental results, the IndoBERT model demonstrated the strongest performance among the evaluated architectures compared to classical Deep Learning models (CNN and BiLSTM) and DistilBERT.

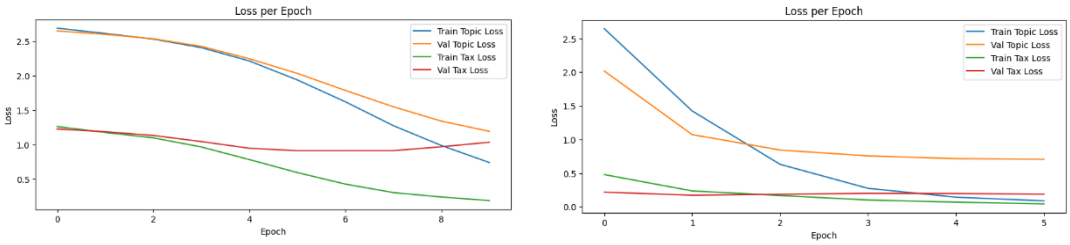
Table 4. Performance comparison

Model	Topic Accuracy	Topic F1-Macro	Taxonomy Accuracy	Taxonomy F1-Macro
CNN	0.6642	0.6221	0.6642	0.6371
BiLSTM	0.6131	0.6285	0.6131	0.6137
DistilBERT	0.4964	0.5157	0.6569	0.6232
IndoBERT	**0.8102**	**0.7807**	**0.7153**	**0.7120**

1. Impact of Language-Specific Pre-training
   The most obvious performance gap appears in topic classification: IndoBERT reached an F1 score of 0.78, while DistilBERT dropped to 0.51. Despite sharing the same Transformer backbone, DistilBERT struggled with the Indonesian science context. Since it was trained on English, the model tends to break Indonesian technical terms into disjointed sub-word units, which dilutes meaning and hurts accuracy [27]. We explicitly included DistilBERT to benchmark a general-purpose model, not because we expected optimal results. This comparison puts a number on the cost of linguistic mismatch—especially when dealing with domain-specific jargon. Ultimately, IndoBERT's edge isn't just about the architecture; it stems from the simple fact that its pre-training data actually aligns with the target language.
2. Local and Sequential Features Notably
   The CNN model (F1 Tax: 0.63) slightly outperforms the BiLSTM (F1 Tax: 0.61). This performance difference suggests that for datasets comprising relatively short questions, local features (*n-grams*) extracted by CNN such as detecting specific operational verbs (e.g., "explain," "mention") are more effective than BiLSTM's long-term sequence modeling [28]. Consequently, CNN serves as a highly competitive benchmark model due to its ability to rapidly capture salient "keywords."
3. Resilience to Class Imbalance
   Although CNN recorded competitive accuracy (~66%), these results must be interpreted with caution. Without advanced attention mechanisms and imbalance mitigation strategies, classical models tend to exhibit bias towards majority patterns. The significant increase in IndoBERT's performance (Topic Accuracy +15% and Taxonomy +5% compared to CNN) supports the effectiveness of the Focal Loss strategy in this experimental setting. This mechanism prevents the model from biasing predictions towards the majority class and compels it to learn patterns within the challenging minority classes, resulting in a much higher and more balanced F1-Macro score [29].

3.2 Training Dynamics and Convergence

In addition to the final metrics, training dynamics were evaluated using the Training Loss curve to assess model stability during training and detect signs of overfitting or underfitting as early as possible.



(a) (b)

Figure 3. Comparison of training stability,
(a) CNN Benchmark Model, (b) IndoBERT with Focal Loss

In Figure 3(a), the CNN model trained from scratch shows a moderate decrease in loss but plateaus early. This plateau indicates feature saturation: the model quickly learns surface-level patterns, such as *n-gram* keywords, but struggles to exceed a certain accuracy threshold due to its architecture's limitations in capturing complex semantic context. The consistent gap between the training and validation losses also suggests that the model struggles to generalize to unseen data.

In contrast, Figure 3(b) shows IndoBERT's curve declining sharply and aggressively in the early phase, which can be attributed to the benefits of transfer learning, where the model fine-tunes pre-existing linguistic representations. The model does not learn language from scratch, but rather fine-tunes its already extensive knowledge [7]. Furthermore, the integration of Focal Loss maintained gradient stability, as seen in the validation curve, which moved in tandem with the training curve without extreme fluctuations. This stability confirms that IndoBERT is capable of minimizing errors even on difficult samples without succumbing to severe overfitting.

3.3. Error Analysis

A primary challenge in this study is distinguishing among closely related cognitive levels. The following is an in-depth analysis of the best-performing model (IndoBERT) using the Confusion Matrix.

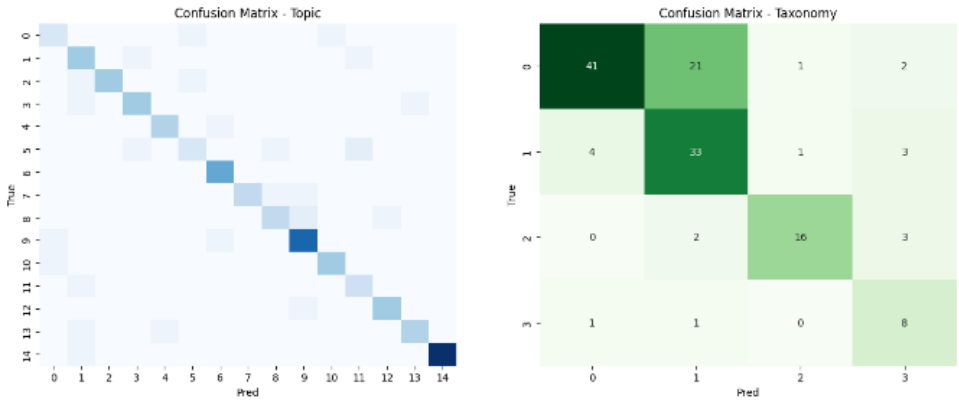


Figure 4. Confusion matrices of IndoBERT

Based on the Confusion Matrix above, several model behavioral patterns were identified:

1. Stability in Minority Class (C4)
   IndoBERT (Figure 4b) showed notably strong performance in Class 3 (C4/Analyzing). Despite being a minority class with the fewest samples, the model correctly predicted 8 out of 10 samples (Recall 80%). This high recall rate demonstrates that the Focal Loss strategy effectively forces the model to place greater weight on these "difficult" samples, preventing them from being overshadowed by the majority class.
2. Prediction Bias between C1 and C2
   In IndoBERT, there are 21 C1 questions that are incorrectly predicted as C2 (cell in row 0, column 1). This misclassification pattern indicates that the model tends to "overestimate" the difficulty of the questions, perceiving memorization questions (C1) as comprehension

questions (C2). This confusion is plausible, given that the boundary between "remembering facts" and "understanding concepts" often depends on subtle, implicit contexts [30].

3. Lexical Dependence on Topic
   In topic classification (left matrix), the model shows high accuracy on topic indices that have unique technical terms, such as "Addictive Substances" and "Electricity". Conversely, prediction errors are scattered across topics with more general vocabulary. This disparity suggests that the Transformer model still relies heavily on specific keywords (lexical cues) to distinguish subject matter [31].

3.4. Computational Efficiency and Performance

In terms of computational efficiency, an approximate comparison was conducted based on observed training duration under the same hardware environment (NVIDIA Tesla T4 GPU). The classical models (CNN and BiLSTM) demonstrated rapid convergence, completing the training phase within a few minutes. In contrast, Transformer-based models such as IndoBERT required noticeably longer training times under comparable epoch settings. While precise millisecond-level inference latency was not logged in this iteration, these observations indicate a clear trade-off between model complexity and computational cost.

BiLSTM emerged as a highly efficient contender in this experiment. Although it is a classical architecture that is significantly lighter than the Transformer family, it achieved a competitive Taxonomy F1-Score (61.37%), comparable to that of DistilBERT (62.32%). This observation suggests that in low-resource settings, well-optimized recurrent architectures remain a highly relevant and efficient alternative, especially when lightweight Transformer models (DistilBERT) are constrained by language limitations [32].

On the other end of the spectrum, IndoBERT emerges as a suitable choice when the primary priority is high-fidelity accuracy (F1-Score > 70%). The greater computational investment in IndoBERT pays off with its ability to manage language nuances and data imbalances with precision. Conversely, CNN, despite offering the fastest inference speed, proves less than ideal for this case because it is prone to majority class bias, rendering it a suboptimal candidate for assessment systems that require prediction fairness.

3.5. System Implementation

To validate the deployment feasibility of the proposed architecture, the highest-performing model (IndoBERT) was integrated into a web-based prototype system. The system was developed using the Streamlit framework, selected for its efficiency in building interactive data applications, and hosted on Hugging Face Spaces cloud infrastructure to ensure research accessibility and reproducibility.

The system workflow commences when the user enters the question text through the interface. The text input is processed by a pre-trained IndoBERT tokenizer to be converted into ID tokens and attention masks. Next, the model performs a *forward-pass inference* to generate two separate *logit* vectors. These vectors are then mapped to text labels (Topic Name and Cognitive Level) using the Softmax function, which are subsequently displayed to the user.

The user interface, depicted in Figure 5, is designed with a minimalist interface to simulate real-world use by educators. Initial testing indicates that the system can provide near real-time inference within a standard CPU environment during preliminary testing. This successful implementation demonstrates that the developed multi-output architecture has practical feasibility for deployment in digital education environments without requiring specialized hardware.

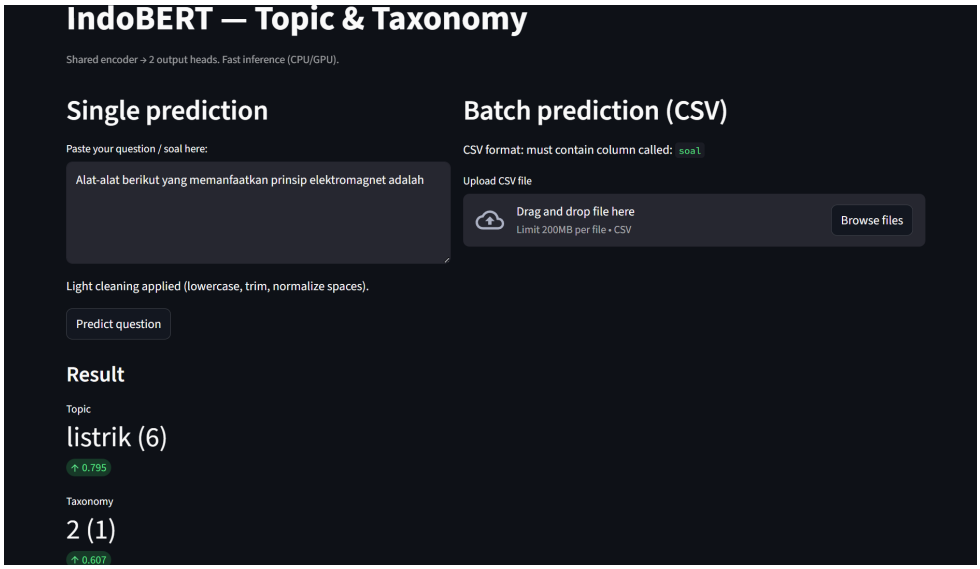


Figure 5. User interface of the automated question classification system

3.6. Discussions

This study set out to evaluate if a single, integrated multi-output model could effectively predict both topic and cognitive level in Indonesian educational questions. We specifically wanted to balance accuracy with efficiency. Our experiments offered several key takeaways regarding both the modeling process and its real-world application. Most notably, the data shows that pre-training language alignment is critical for topic classification. The gap between IndoBERT and DistilBERT isn't just about model differences; it spotlights the friction caused when pre-training data doesn't match the target language. We used DistilBERT here as a baseline and a representative of general-purpose Transformers to quantify the struggle models face with Indonesian technical terms. This highlights that linguistic representation quality plays a critical role for text-based topic classification.

Apart from language, class imbalance heavily skewed cognitive-level predictions. Since the dataset was dominated by lower-level questions (C1 and C2), standard loss functions naturally biased the models toward the majority. Implementing Focal Loss for the taxonomy output helped steer the learning process toward rarer, more difficult samples, improving the recognition of higher-order thinking. This proves that fixing class imbalance is essential for automating cognitive assessment. Architecturally, the multi-output approach wins on efficiency. By using a shared encoder with two task-specific heads, we avoid running feature extraction separately for each task, cutting down on computational redundancy. Interestingly, BiLSTM remained competitive, suggesting that lighter sequential models are still viable for resource-tight environments, even if they sacrifice some accuracy. However, a clear trade-off remains. Transformer-based models like IndoBERT provide stability against linguistic complexity and imbalance but demand more power. Classical models save resources but hit a performance ceiling. Consequently, the choice of architecture ultimately depends on the specific operational constraints of the educational system in use.

4. CONCLUSION

We explored a multi-output classification method to predict both topic and cognitive level in Indonesian educational questions simultaneously. Using a single architecture with a shared encoder and two task-specific heads, this approach reduces the computational redundancy often found in separate modeling schemes while maintaining competitive performance. The results highlight that language alignment during pre-training plays a critical role in topic classification, a pattern reflected in IndoBERT's stronger performance compared to general-purpose Transformer models. Applying Focal Loss improved the model's ability to identify minority cognitive levels under an imbalanced class distribution. Meanwhile, BiLSTM's stability suggests that lighter

architectures remain viable in resource-constrained environments. From a practical perspective, the proposed multi-output framework supports automatic and consistent question labeling in Computer-Based Test systems, making it suitable for scenarios where efficiency and low-latency operation are desirable. Nevertheless, this study is limited by dataset size and its focus on the science domain. Future research should therefore explore broader subject coverage, collaborative annotation strategies, and alternative data partitioning schemes to enhance generalization. The study balances technical depth with application context, making the findings relevant to a broad audience ranging from NLP researchers to information systems practitioners and educational platform developers.

REFERENCES

[1] I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, and P. Tudorache, "Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review," Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/educsci13121216.

[2] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational Data Mining: A Foundational Overview," *Encyclopedia*, vol. 4, no. 4, pp. 1644–1664, Dec. 2024, doi: 10.3390/encyclopedia4040108.

[3] A. Lubbe, E. Marais, and D. Kruger, "Cultivating independent thinkers: The triad of artificial intelligence, Bloom's taxonomy and critical thinking in assessment pedagogy," *Educ Inf Technol (Dordr)*, vol. 30, no. 12, pp. 17589–17622, Aug. 2025, doi: 10.1007/s10639-025-13476-x.

[4] M. Pota, M. Esposito, G. De Pietro, and H. Fujita, "Best practices of convolutional neural networks for question classification," *Applied Sciences (Switzerland)*, vol. 10, no. 14, Jul. 2020, doi: 10.3390/app10144710.

[5] S. Mondal, M. Barman, and A. Nag, "Question classification task based on deep learning models with self-attention mechanism," *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-19239-z.

[6] 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.

[7] C. Eang and S. Lee, "Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN_Bert_Based)," *Applied Sciences (Switzerland)*, vol. 14, no. 18, Sep. 2024, doi: 10.3390/app14188388.

[8] T. Miao and D. Xu, "KWM-B: Key-Information Weighting Methods at Multiple Scale for Automated Essay Scoring with BERT," *Electronics (Switzerland)*, vol. 14, no. 1, Jan. 2025, doi: 10.3390/electronics14010155.

[9] X. Chen, D. Zou, H. Xie, G. Cheng, Z. Li, and F. Lee Wang, "Automatic Classification of Online Learner Reviews Via Fine-Tuned BERTs," *International Review of Research in Open and Distributed Learning*, vol. 26, 2025.

[10] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang, "A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, May 2023, pp. 943–956.

[11] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Classifications of the summative assessment for revised Bloom's Taxonomy by using deep learning," *International Journal of Engineering Trends and Technology*, vol. 69, no. 3, pp. 211–218, Mar. 2021, doi: 10.14445/22315381/IJETT-V69I3P232.

[12] A. Indrawati, H. Subagyo, A. Sihombing, Wagiyah, and S. Afandi, "Analyzing The Impact Of Resampling Method for Imbalanced Data Text In Indonesian Scientific Articles Categorization," *BACA: Jurnal Dokumentasi dan Informasi*, vol. 42, no. 2, pp. 133–141, Dec. 2020, [Online]. Available: https://jurnalbaca.pdii.lipi.go.id/index.php/baca/article/view/563

[13] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[14] Rousyati, D. Pratmanto, A. Eko Widodo, K. Fatmawati, and R. Diva Saputra, "Analisa Sentimen Persepsi Masyarakat Terhadap Aplikasi Bea Cukai Mobile Menggunakan Algoritma Naive Bayes Dan K-Nearest Neighbors," *Jurnal Sains dan Manajemen*, vol. 12, no. 2, 2024.

[15] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing: Abridged Edition*. Allyn and Bacon, 2001. Accessed: Jan. 03, 2026. [Online]. Available: https://www.jsu.edu/online/faculty/blooms-revised-taxonomy-verb-list.pdf

[16] P. M. Newton, A. Da Silva, and L. G. Peters, "A Pragmatic Master List of Action Verbs for Bloom's Taxonomy," *Front Educ (Lausanne)*, vol. 5, Jul. 2020, doi: 10.3389/feduc.2020.00107.

[17] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Oct. 2014, pp. 1746–1751. [Online]. Available: http://nlp.stanford.edu/sentiment/

[18] B. Yang, Y. Wang, and Y. Zhan, "Lithium Battery State-of-Charge Estimation Based on a Bayesian Optimization Bidirectional Long Short-Term Memory Neural Network," *Energies (Basel)*, vol. 15, no. 13, Jul. 2022, doi: 10.3390/en15134670.

[19] A. Vaswani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 2017.

[20] M. D. Maulana and C. S. K. Aditya, "Perbandingan IndoBERT dan Bi-LSTM Dalam Mendeteksi Pelanggaran Undang-Undang ITE," *Sintech Journal*, vol. 8, no. 1, Apr. 2025.

[21] H. M. Ahmad and A. Rahimi, "Dynamic Focal Loss for Imbalanced Learning," *SN Comput Sci*, vol. 6, no. 6, Aug. 2025, doi: 10.1007/s42979-025-04198-5.

[22] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 1082–1090, 2023.

[23] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, p. 100026, Sep. 2023, doi: 10.1016/j.nlp.2023.100026.

[24] R. Katirci and H. Çelik, "Learning Rate Sensitivity In Transformer Models: A Case Study In Neural Machine Translation," in *Sivas International Conference On Scientific And Innovation Research-IV*, Sivas, May 2025, pp. 719–729.

[25] J. Lovón-Melgarejo, L. Soulier, K. Pinel-Sauvagnat, and L. Tamine, "Studying Catastrophic Forgetting in Neural Ranking Models," 2021, pp. 375–390. doi: 10.1007/978-3-030-72113-8_25.

[26] B. Hemanth Sai, S. Mukherjee, and S. R. Dubey, "Adaptive adam-based optimizers using second-order weight decoupling and gradient-aware weight decay for vision transformer," *Mach Vis Appl*, vol. 36, no. 3, May 2025, doi: 10.1007/s00138-025-01686-9.

[27] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics,* Barcelona: Online, Dec. 2020, pp. 757–770. [Online]. Available: https://huggingface.co/

[28] B. Jang, M. Kim, G. Harerimana, S. U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020, doi: 10.3390/app10175841.

[29] L. Cao, X. Liu, and H. Shen, "Adaptable Focal Loss for Imbalanced Text Classification," *Springer Nature Switzerland*, pp. 466–475, 2022, doi: 10.1007/978-3-030-96772-7.

[30] C. J. Stanny, "Reevaluating Bloom's Taxonomy: What measurable verbs can and cannot say about student learning," *Educ Sci (Basel)*, vol. 6, no. 4, Dec. 2016, doi: 10.3390/educsci6040037.

[31] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," Apr. 01, 2023, *MDPI*. doi: 10.3390/info14040242.
[32] M. Sharma, "A Comparative Analysis of Natural Language Processing Models: BERT and LSTM in Enhancing Business Communication," *Int J Innov Sci Res Technol*, pp. 1360–1365, Apr. 2025, doi: 10.38124/ijisrt/25apr943.