.

An IndoBERT-Based Intelligent Chatbot for School Information Services

Soubran Ananda[1], Ari Muzakir[2]
[1,2] Information Systems, Faculty of Science and Technology, Universitas Bina Darma, Indonesia
sobranananda78@gmail.com

Submitted: August 4, 2026; accepted: September 23, 2026

Abstract

Fragmented school communication can delay access to official information and increase the number of repetitive administrative inquiries. This study developed and evaluated a practical school-information service that integrates an Indonesian-language IndoBERT intent classifier with a responsive web application at MA Islamiyah Bindu. The web application was implemented using PHP, Bootstrap 5, and MySQL, while the chatbot used the indobenchmark/indobert-base-p1 model for ten-class intent classification and controlled response retrieval. A balanced dataset of 600 school-related Indonesian utterances was constructed from official school information and recurring question patterns, with 500 utterances used for training and 100 held-out utterances for testing. IndoBERT was fine-tuned using a maximum sequence length of 64, batch size of 8, learning rate of 2×10−5, and 40 training epochs. The reported run achieved 100% training accuracy, while the held-out test set produced 94.0% accuracy, 94.3% macro-precision, 94.0% macro-recall, and 94.0% macro-F1. Class-wise analysis showed that most errors occurred between semantically adjacent school-information intents. End-to-end black-box testing also confirmed that all ten original functional scenarios, including confidence-based fallback handling, passed. The findings demonstrate the practical feasibility of integrating Indonesian intent classification with a controlled institutional knowledge base for school information services. However, the evaluation is limited by the absence of a separate validation set, repeated-seed analysis, same-dataset baseline comparisons, formal annotator agreement, and broader user-facing robustness testing.

Keywords*: educational chatbot, IndoBERT, intent classification, school information system, web application.*

1. Introduction

Digital transformation in education requires information services that are accurate, continuously accessible, and manageable by the institution. School websites commonly centralize profiles, announcements, academic schedules, admission information, activity documentation, and contact channels. When information remains distributed across bulletin boards and messaging groups, users must search multiple channels and school personnel repeatedly respond to similar questions. Recent research has therefore positioned conversational agents as a practical interface for reducing information-search effort and improving service accessibility [1], [2].

Educational chatbot research has expanded from rule-based question matching toward contextual language models. Systematic reviews published between 2023 and 2025 reported benefits such as rapid response, reduced routine workload, flexible access, and improved user engagement, while emphasizing the importance of controlled knowledge, transparent evaluation, and institutionally verified answers [1], [3]. For administrative school services, intent classification with controlled response retrieval is preferable to unrestricted text generation because official schedules, admission requirements, and institutional identities must remain consistent.

Intent classification identifies the purpose of a user utterance before the system selects the corresponding response. Recent Indonesian studies showed that language-specific transformer models can outperform conventional or multilingual alternatives in contextual text classification. IndoBERT achieved 94% accuracy in an Indonesian university-chatbot comparison [4], exceeded an SVM baseline in an institutional service chatbot [5], and produced competitive results in multi-label Indonesian classification tasks [6], [7]. These findings support the use of IndoBERT for short, domain-specific school questions with varied Indonesian word order and vocabulary.

MA Islamiyah Bindu previously relied on bulletin boards and WhatsApp groups as its primary communication media. This condition caused fragmented access to school profiles, facilities, schedules,

registration requirements, and contact information. The institution therefore required a centralized website and a chatbot capable of interpreting Indonesian questions and retrieving approved answers from a school knowledge base.

This study developed and evaluated an IndoBERT-based intelligent chatbot integrated into a responsive school-information website. The contribution of this study is positioned primarily as a practical information-system integration rather than a new NLP algorithm. The contributions are: (1) a centralized web application for accessing official school information; (2) a balanced intent dataset containing school-specific Indonesian utterances; (3) an integrated architecture combining IndoBERT-based intent classification with controlled response retrieval from a school knowledge base; and (4) an empirical evaluation using an independent held-out test set, class-wise analysis, a confusion matrix, and end-to-end functional testing.

2. Method

2.1. Research Design and Requirement Analysis

The study used a research-and-development approach with iterative requirement analysis, interface design, model development, system integration, and evaluation. Data requirements were identified through direct observation of the existing communication process, interviews with school personnel, review of official school information, and compilation of recurring questions from students and parents. The implementation site was MA Islamiyah Bindu, Ogan Komering Ulu, South Sumatra, Indonesia.

The functional scope included the homepage, school profile, vision and mission, facilities, activity information, new-student admission information, contact and map services, administrator-managed content, and a floating chatbot interface. Non-functional requirements included responsive display, understandable navigation, controlled content maintenance, secure administrator access, and reliable operation across desktop and mobile devices. The functional and non-functional requirements were defined based on the information needs identified during observation and interviews, with system quality considerations aligned with the ISO/IEC 25010 product quality model [8]. These requirements followed the contemporary ISO/IEC 25010 product-quality model (Table I).

Table I. System Requirements

Requirement	Implementation Evidence
Centralized information	Homepage, school profile, vision-mission, activities, PPDB, and contact pages
Conversational access	Floating chatbot widget connected to the IndoBERT inference service
Controlled official answers	Intent-to-response mapping retrieved from a school knowledge base
Responsive interface	Bootstrap 5 layout for desktop and mobile screens
Maintainability	Separated website content, model inference, and response-management components
Fallback handling	Questions below the confidence threshold returned a clarification response
Centralized information	Homepage, school profile, vision-mission, activities, PPDB, and contact pages

2.2. Dataset Construction and Labeling

The chatbot was formulated as a supervised ten-class intent classification problem. The dataset consisted of 600 Indonesian utterances representing information needs related to MA Islamiyah Bindu. The utterances were constructed from official school information and recurring question patterns identified during the requirement-analysis stage. The information categories were defined according to the services intended to be provided through the chatbot, including greetings, school profile, headmaster information, address and contact information, facilities, examination schedules, school hours, academic calendar, PPDB requirements, and curriculum information.

Each utterance was assigned to one of the ten predefined intents according to its intended information need. To maintain a balanced classification problem, each intent contained 60 utterances. Linguistic variation was introduced through alternative question forms, synonymous expressions, polite and informal wording, and different word orders while preserving the same underlying institutional meaning. Table II presents the distribution of the dataset across the ten intent categories. For each intent, 50 utterances were allocated to the training set and 10 utterances to the held-out test set, resulting in 500 training utterances and 100 test utterances. The balanced distribution ensured that each intent contributed equally to the model training and final evaluation.

Before dataset splitting, exact duplicate utterances were removed to avoid identical samples appearing in different partitions. The data were then divided using a stratified split into 500 training utterances and 100 held-out test utterances, with the same 50:10 distribution for every intent shown in Table II. The test set was isolated from the model fine-tuning process and was used only for final performance evaluation.

Table II. Intent Distribution And Example Utterances

Intent	Train	Test	Example Utterance
Greeting	50	10	Halo, apakah chatbot dapat membantu saya?
School profile	50	10	Apa visi dan misi MA Islamiyah Bindu?
Headmaster	50	10	Siapa kepala madrasah saat ini?
Address/contact	50	10	Bagaimana cara menghubungi pihak sekolah?
Facilities	50	10	Fasilitas apa saja yang tersedia?
Exam schedule	50	10	Kapan ujian semester dilaksanakan?
School hours	50	10	Jam masuk sekolah pukul berapa?
Academic calendar	50	10	Kapan kegiatan semester baru dimulai?
PPDB requirements	50	10	Dokumen apa yang dibutuhkan untuk PPDB?
Curriculum	50	10	Kurikulum apa yang digunakan sekolah?

The dataset construction process did not include a formal multi-annotator agreement procedure, and no quantitative inter-annotator agreement statistic was recorded. Therefore, the labels represent the intent assignments established during dataset construction rather than independently validated annotations. In addition, the available dataset documentation records the removal of exact duplicates but does not provide a quantitative semantic-similarity screening procedure for near-duplicate utterances. This limitation should be considered when interpreting the generalization performance because semantically similar utterances may still occur across the training and test partitions.

Text cleaning was limited to standardizing quotation marks, excessive spaces, and non-informative symbols. Stemming and stop-word removal were not applied because the contextual language model was intended to retain sentence structure and subword information. The resulting utterances were subsequently processed using the tokenizer associated with the *indobenchmark/indobert-base-p1* model.

2.3.Text Processing and IndoBERT Fine-Tuning

The implementation applied minimal text cleaning to remove inconsistent quotation marks and redundant spaces while retaining semantically meaningful words. Aggressive stemming and stop-word removal were not applied because contextual encoders are intended to preserve sentence structure and subword information [10]. The cleaned utterances were processed using the tokenizer associated with the *indobenchmark/indobert-base-p1* model [11], [12]. Padding and truncation were enabled, with the maximum input length set to 64 tokens.

Figure 1 illustrates the complete IndoBERT fine-tuning and deployment architecture used in this study. Each utterance was processed by the IndoBERT tokenizer and converted into input IDs and attention masks with a maximum length of 64 tokens. The representation combined token, position, and segment embeddings before passing through 12 bidirectional Transformer encoder layers with 12 attention heads and a hidden size of 76 [13]. The contextual representation of the [CLS] token was then passed through dropout and a linear classification layer containing ten output neurons. Softmax was applied to convert the resulting logits into intent probabilities, and the class with the highest probability was selected as the predicted intent [14]. At the application level, a confidence threshold of 0.70 was applied to determine whether the predicted intent was sufficiently confident for controlled response retrieval. If the highest probability was below this threshold, the system returned a fallback clarification response.

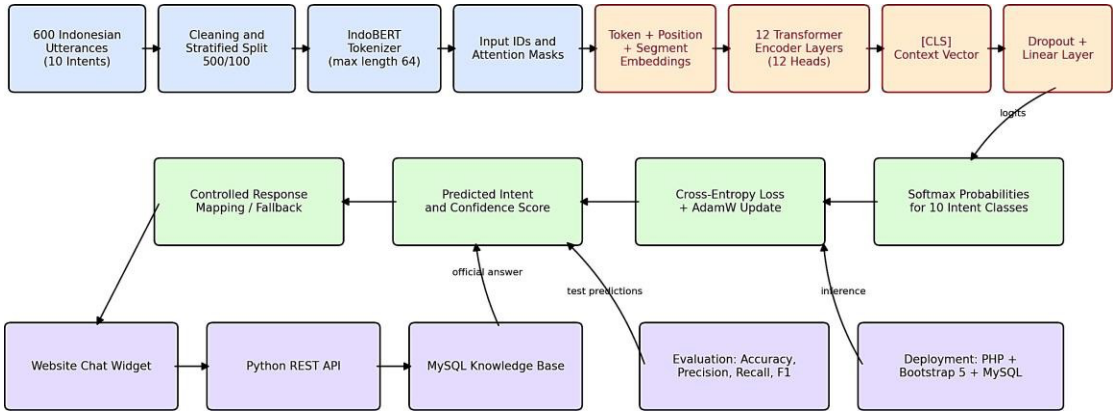


Fig. 1 IndoBERT fine-tuning and deployment architecture for the school-information chatbot

Fine-tuning used the AdamW optimizer with a learning rate of 2×10−5, weight decay of 0.01, a batch size of 8, and 40 training epochs. Cross-entropy loss was used to optimize the ten-class classification head. A fixed random seed of 42 was used for the reported experiment. Table III summarizes the model architecture, training configuration, confidence threshold, and evaluation measures used in the study.

Table III. Model And Training Configuration

Component	Configuration
Pre-trained model	indobenchmark/indobert-base-p1
Classification task	10-class intent classification
Encoder	12 Transformer layers; 12 attention heads; hidden size 768
Maximum sequence length	64 tokens
Training/test split	500/100 stratified utterances
Batch size	8
Optimizer	AdamW
Learning rate	2×10−5
Weight decay	00.01
Epochs	40
Random seed	42
Confidence threshold	0,048611111
Evaluation	Accuracy, macro-precision, macro-recall, macro-F1, confusion matrix

The reported experiment used the 500 training utterances for model fine-tuning, while the 100 held-out test utterances were reserved for final evaluation. No separate validation set was used in the reported experiment. Therefore, the 40 training epochs and the 0.70 confidence threshold were applied as fixed experimental settings rather than being selected through validation-set optimization. Likewise, the fixed seed of 42 represents a single experimental run and does not provide an estimate of performance variation across multiple random seeds. These limitations are considered when interpreting the reported test performance. Accuracy, macro-precision, macro-recall, and macro-F1 were calculated on the held-out test set. Macro-averaged metrics were reported to give equal importance to each intent class, while the confusion matrix was used to examine classification errors between individual intent categories.

2.4. System Architecture and Integration

The core website used PHP and Bootstrap 5, with MySQL storing website content and controlled chatbot responses. The IndoBERT classifier was deployed through a Python REST service [15]. The chat widget sent the user text to the service, which tokenized the input, generated class probabilities, applied the confidence threshold, and returned the predicted intent. The PHP application then retrieved the associated approved response from the knowledge base and displayed it in the conversation interface. Black-box testing was used to verify the end-to-end operation of the integrated application, including website navigation, content display, chatbot activation, in-domain queries, controlled response retrieval, and confidence-based fallback handling. The results of these functional tests are presented in Section 3.4.

3. Result and Discussion

3.1. Website and Chatbot Implementation

The resulting website provided a single access point for school identity, vision and mission, activities, PPDB information, facilities, contact details, and location. The responsive navigation and quick-access cards reduced the number of steps required to reach frequently requested information. The chatbot was embedded as a floating interface, allowing users to request information without leaving the active page. Figure 2 shows the main information page, the contact service, and representative chatbot responses.

3.1. Training Performance

The fine-tuning process showed a rapid improvement in training performance during the early epochs, followed by a progressively more stable pattern toward the end of training. Figure 3 presents the training loss and training accuracy across the 40 training epochs. The training loss decreased from 2.4970 in the first epoch to 0.0027 in the fortieth epoch, while training accuracy increased from approximately 18% to 100%. These results indicate that the model successfully fitted the training utterances and learned the linguistic patterns associated with the ten predefined intents.

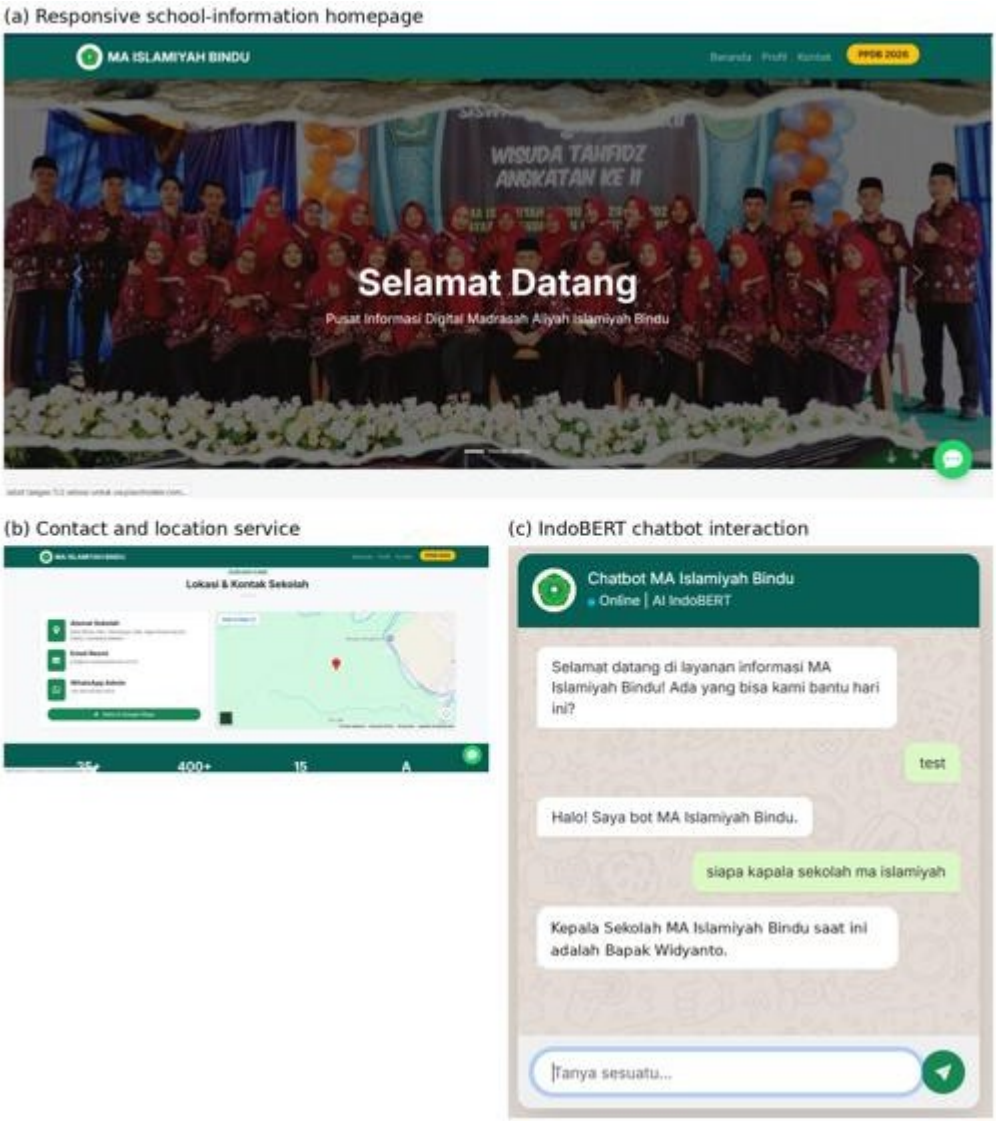


Fig. 2 Implementation of the responsive school-information website and chatbot

3.2. Training Performance

The fine-tuning process showed a rapid improvement in training performance during the early epochs, followed by a progressively more stable pattern toward the end of training. Figure 3 presents the training loss and training accuracy across the 40 training epochs. The training loss decreased from 2.4970 in the first epoch to 0.0027 in the fortieth epoch, while training accuracy increased from approximately 18% to 100%. These results indicate that the model successfully fitted the training utterances and learned the linguistic patterns associated with the ten predefined intents.

The observed optimization pattern is broadly consistent with previous studies that applied BERT-based and IndoBERT-based fine-tuning to Indonesian text classification tasks, where contextual pretrained representations were adapted to task-specific classification objectives [16], [17], [18]. The stable final segment of the curve showed that additional epochs produced only marginal improvement. However, these studies are not used as direct performance baselines for the present experiment because their datasets, task definitions, and experimental settings differ from those used in this study.

The training curves also show that the model reached a very high level of performance on the training data. However, the absence of a separate validation set means that the training curves alone cannot determine whether the model continued to improve in terms of generalization or whether overfitting occurred during the later epochs. In particular, the final training accuracy of 100% should not be interpreted

as evidence of optimal model performance. The generalization capability of the model is instead assessed using the held-out test set presented in Section 3.3.

The decrease in training loss and increase in training accuracy demonstrate successful optimization of the fine-tuning objective, but they do not provide sufficient evidence for determining the optimal number of training epochs. Because the reported experiment used 40 fixed training epochs without validation-based early stopping or model selection, the possibility of overfitting cannot be completely excluded. This limitation is particularly relevant given the relatively small number of training utterances. Therefore, the test-set results, rather than the training curves alone, are used as the primary evidence of classification performance.

Figure 3 demonstrates that the IndoBERT model was able to fit the task-specific training data effectively. Nevertheless, the absence of validation curves and repeated training runs means that the observed training trajectory should be interpreted as evidence of optimization for the reported run rather than as evidence of optimal hyperparameter selection or robust generalization across different random initializations.

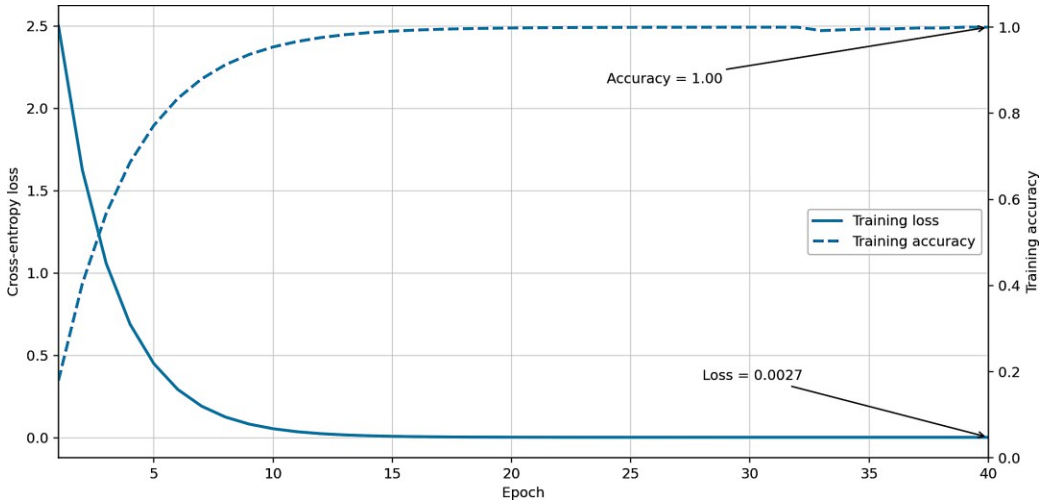


Fig. 3 Combined training loss and training accuracy

3.3. Independent Test-Set Evaluation

The classification performance was evaluated on the 100 held-out utterances not used during model fine-tuning. Table IV summarizes the overall classification performance on the test set. The model correctly classified 94 of the 100 utterances, resulting in an accuracy of 94.0%. The model also achieved a macro-precision of 94.3%, macro-recall of 94.0%, and macro-F1 of 94.0%. Because the test set contained an equal number of utterances for each intent, the macro-averaged measures provide an appropriate summary of performance across the ten intent categorie.

Table IV. Overall Classification Performance

Metric	Result
Accuracy	94.0%
Macro-precision	94.3%
Macro-recall	94.0%
Macro-F1	94.0%
Correct classifications	94 of 100
Incorrect classifications	6 of 100

The 94 correct classifications out of 100 test utterances correspond to an estimated accuracy of 94.0%. Using a 95% Wilson confidence interval for the binomial proportion, the estimated accuracy has an interval of approximately 87.5%–97.2%. This interval indicates that the observed 94.0% accuracy should be interpreted with consideration of the relatively small test-set size rather than as an exact estimate of performance on all possible school-information queries. Figure 4 presents the confusion matrix for the 100 held-out test utterances and provides the class-level distribution of correct and incorrect predictions. Greeting, Contact, Hours, and PPDB achieved 10 correct predictions out of 10 test samples. Profile, Headmaster, Facilities, Exam, Calendar, and Curriculum each had one misclassified sample, resulting in

nine correct predictions for those categories. The errors were concentrated among semantically related intents rather than being distributed randomly across all categories.

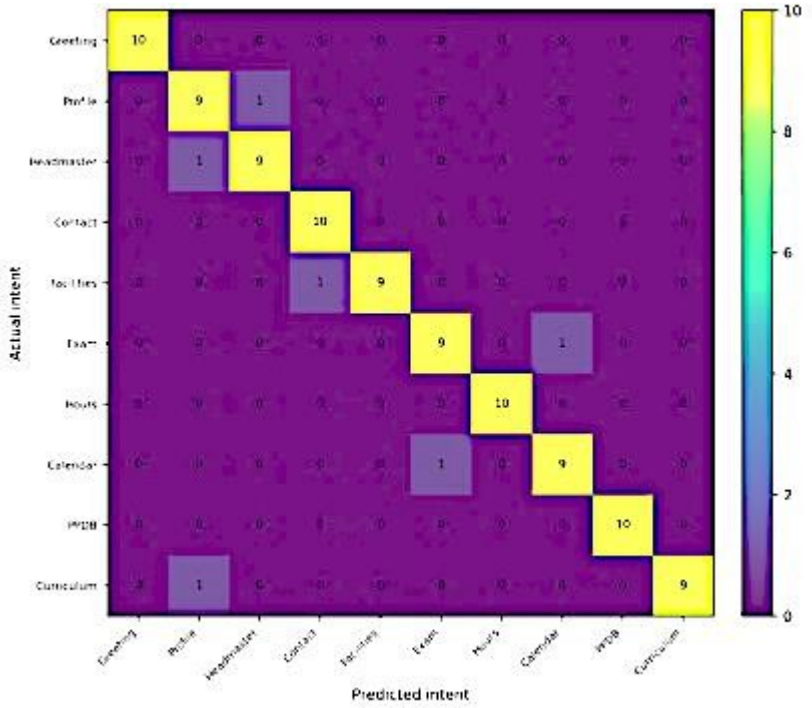


Fig. 4 Confusion matrix for the independent 100-utterance test set

The class-wise precision, recall, and F1-scores calculated from the confusion matrix are presented in Table V. Greeting, Hours, and PPDB achieved precision, recall, and F1-scores of 100%. Contact achieved a precision of 90.9%, recall of 100.0%, and F1-score of 95.2%. Facilities and Curriculum obtained F1-scores of 94.7%, while Headmaster, Exam, and Calendar each obtained an F1-score of 90.0%. Profile produced the lowest F1-score at 85.7%, mainly because one Profile utterance was predicted as Headmaster and one Headmaster utterance was predicted as Profile, which reduced the precision of the Profile class.

Table V. Functional Testing of The Integrated System

Intent	Precision	Recall	F1-score	Support
Greeting	100.0%	100.0%	100.0%	10
Profile	81.8%	90.0%	85.7%	10
Headmaster	90.0%	90.0%	90.0%	10
Contact	90.9%	100.0%	95.2%	10
Facilities	100.0%	90.0%	94.7%	10
Exam	90.0%	90.0%	90.0%	10
Hours	100.0%	100.0%	100.0%	10
Calendar	90.0%	90.0%	90.0%	10
PPDB	100.0%	100.0%	100.0%	10
Curriculum	100.0%	90.0%	94.7%	10

The confusion matrix in Figure 4 shows six classification errors. One Profile utterance was classified as Headmaster, while one Headmaster utterance was classified as Profile. Similarly, one Facilities utterance was classified as Contact, one Exam utterance was classified as Calendar, one Calendar utterance was classified as Exam, and one Curriculum utterance was classified as Profile. These errors indicate that the main classification difficulty occurred between intents that share institutional vocabulary or closely related information concepts. For example, questions about the school profile and headmaster may both contain references to the school and its institutional identity, while examination schedules and academic calendars both contain temporal information.

The obtained 94.0% accuracy is comparable to the 94% accuracy reported by Dwiyono et al. [4] for Indonesian university-chatbot intent classification. The result also exceeds the 87% IndoBERT accuracy reported by Santosa et al. [5] for an Indonesian overtime-service chatbot. However, these results should not be interpreted as evidence that IndoBERT is superior to the methods used in those studies, because the

datasets, intent definitions, number of classes, sample sizes, and experimental settings were different. In particular, the present study used a relatively small, balanced dataset with narrowly defined school-information intents. Therefore, the comparisons provide contextual evidence rather than controlled baseline comparisons.

The test-set results demonstrate that the proposed system can classify the predefined school-information intents with relatively high accuracy under the evaluated conditions. Nevertheless, the evaluation represents a single experimental run using a 100-utterance test set from one institutional domain. The absence of repeated random-seed experiments and same-dataset baseline models limits the strength of conclusions regarding the robustness and relative benefit of IndoBERT. These limitations are considered in the discussion in Section 3.5.

3.4. Functional Evaluation of the Integrated Application

Black-box testing was conducted to verify the end-to-end functionality of the integrated school-information website and chatbot. The evaluation focused on whether the main website modules, chatbot interface, intent prediction, controlled response retrieval, and confidence-based fallback mechanism operated according to their intended functions. Table VI summarizes the ten functional scenarios evaluated in the reported implementation.

Table VI. Functional Testing of The Integrated System

Scenario	Expected Result	Observed Result	Status
Open homepage	Identity, navigation, and quick links are displayed	Homepage and PPDB quick access displayed correctly	Passed
Open profile page	Official profile is readable	Profile, vision, and mission displayed	Passed
Open contact page	Contact and location are available	Contact details and map displayed	Passed
Open chatbot	Chat interface is activated	Floating chat window opened correctly	Passed
Greeting query	Greeting response is returned	Greeting intent and approved response returned	Passed
Headmaster query	Headmaster information is returned	Correct official answer displayed	Passed
PPDB query	Registration requirements are returned	Required documents displayed	Passed
School-hours query	Official school hours are returned	Stored school-hour answer displayed	Passed
Facilities query	Facility information is returned	Facility list displayed	Passed
Out-of-domain query	Fallback asks for clarification	Fallback response displayed below 0.70 confidence	Passed

As shown in Table VI, all ten planned functional scenarios passed. The website navigation, information pages, chatbot activation, representative intent queries, controlled response retrieval, and fallback mechanism operated as expected. The out-of-domain scenario also demonstrated that the application did not directly return an intent-specific institutional answer when the predicted confidence was below the predefined 0.70 threshold. Instead, the system returned a fallback clarification response.

The functional results confirm the technical integration of the website, chatbot inference service, and controlled knowledge base. In particular, the chatbot did not generate unrestricted responses; the predicted intent was used to retrieve an institutionally approved response from the knowledge base. This design supports consistency between conversational responses and the information maintained by the school. However, the functional evaluation should be interpreted as basic black-box functional verification rather than a comprehensive robustness or usability evaluation. The ten scenarios reported in Table VI did not systematically evaluate misspelled queries, ambiguous questions, mixed Indonesian-English inputs, multiple-intent queries, multiple out-of-domain query types, response latency, formal mobile usability, accessibility, or user satisfaction. Although the website was designed to be responsive for desktop and mobile devices, a formal mobile usability assessment was not conducted in the reported experiment.

Similarly, the functional testing did not involve a structured evaluation with students, parents, teachers, or other end users. Therefore, the current results demonstrate that the implemented functions operated according to their intended specifications, but they do not provide sufficient evidence regarding user acceptance, perceived usefulness, usability, or satisfaction. These aspects should be evaluated in future work through a broader test set and structured user evaluation.

3.5. Discussion

The implementation demonstrated that an Indonesian pre-trained language model can be integrated into a conventional school-information website without requiring the entire information system

to be rebuilt around an AI framework. The modular architecture separates the public website, NLP inference service, and controlled response management, allowing the information system to remain maintainable while the intent-classification component can be updated when new question variants become available. This observation is consistent with previous studies showing that pretrained Indonesian language models can be fine-tuned for task-specific text classification using relatively limited domain-specific data [5], [6], [7].

The classification results provide evidence that IndoBERT can effectively model the ten predefined school-information intents under the evaluated conditions. The model achieved 94.0% accuracy and 94.0% macro-F1 on the 100 held-out test utterances, while the class-wise analysis showed that most categories achieved F1-scores of 90% or higher. The strongest performance was observed for Greeting, Hours, and PPDB, each achieving 100% precision, recall, and F1-score. The comparatively lower performance for Profile was associated with confusion between Profile and Headmaster intents. These results suggest that the main challenge was not general failure to identify school-related questions, but distinguishing between semantically adjacent institutional categories.

The confusion patterns are also informative from a system-design perspective. Profile and Headmaster questions may share terms related to institutional identity, while Exam and Calendar questions can both contain temporal expressions. Similarly, Facilities and Contact questions may contain overlapping references to available school services or locations. These observations indicate that further improvement should focus not only on increasing the number of training samples but also on adding more discriminative utterance variants for closely related intents. In particular, examples should explicitly distinguish the information requested rather than relying only on shared institutional vocabulary.

The training results require a more cautious interpretation. Although the training loss decreased from 2.4970 to 0.0027 and training accuracy reached 100%, these results only demonstrate that the model fitted the training data effectively. Because no separate validation set was used, the training curves cannot determine whether the model continued to improve in generalization during the later epochs or whether overfitting occurred. The 100% training accuracy should therefore not be interpreted as evidence that the model is optimally trained. The held-out test-set performance provides a more appropriate indication of generalization within the evaluated dataset.

The reported 94.0% test accuracy is comparable to the 94% accuracy reported by Dwiyono et al. [4] for Indonesian university-chatbot intent classification and is higher than the 87% IndoBERT result reported by Santosa et al. [5] for an Indonesian overtime-service chatbot. However, these comparisons should be interpreted cautiously because the studies used different datasets, intent definitions, sample sizes, domains, and experimental settings. The present result therefore should not be interpreted as evidence of superiority over those methods. A controlled comparison using the same dataset and experimental protocol, such as TF-IDF-SVM, Logistic Regression, BiLSTM, mBERT, or another Indonesian language model, would provide stronger evidence regarding the specific benefit of IndoBERT.

The evaluation design also has several methodological limitations. First, the dataset contained only 600 utterances, with 500 used for training and 100 for testing. Although the test set was held out from fine-tuning, no separate validation set was used for model selection, hyperparameter tuning, epoch selection, or confidence-threshold optimization. Consequently, the 40 training epochs and 0.70 confidence threshold should be regarded as fixed experimental settings rather than empirically optimized values. Second, the reported result was obtained using a single random seed of 42. Repeated experiments with multiple random seeds would be required to estimate the variability of model performance and provide a more robust performance estimate.

The dataset construction process also limits the strength of the generalization claim. The utterances were derived from official school information and recurring question patterns, and exact duplicate utterances were removed before the train-test split. However, the available dataset documentation does not provide a formal quantitative semantic-similarity screening procedure for near-duplicate utterances. Likewise, no formal multi-annotator agreement statistic was recorded during dataset construction. Therefore, the test-set result demonstrates generalization to the held-out utterances in the constructed dataset, but it should not be interpreted as evidence of broad generalization to all possible Indonesian school-information queries or to other schools.

From an information-system perspective, the controlled intent-response design remains an important practical advantage of the proposed architecture. Instead of generating unrestricted responses, the chatbot predicts an intent and retrieves an approved response from the school knowledge base. This approach reduces the risk of providing unsupported institutional information and allows the school to retain control over official content. The trade-off is limited conversational coverage. Questions outside the predefined intents, spelling variations, ambiguous questions, mixed-language expressions, and multiple-

intent queries may not always be handled correctly. The system should therefore be regarded as a domain-specific information assistant rather than a general conversational agent.

The confidence-based fallback mechanism provides an additional layer of response control. When the highest predicted probability falls below 0.70, the system returns a clarification response instead of directly retrieving an intent-specific answer. The functional test demonstrated that this mechanism operated correctly for the reported out-of-domain scenario. However, because the 0.70 threshold was not optimized using a validation set and only a limited number of functional scenarios were evaluated, the current experiment does not establish that the threshold is optimal for detecting uncertain or out-of-domain queries.

The functional evaluation also confirms the technical feasibility of integrating the chatbot with the school website, but it should not be interpreted as a comprehensive usability or robustness study. The reported ten scenarios verified basic website functions, chatbot activation, representative intent queries, controlled response retrieval, and fallback handling. They did not systematically evaluate misspellings, ambiguous questions, mixed Indonesian-English input, multiple-intent queries, multiple types of out-of-domain questions, response latency, accessibility, formal mobile usability, or user satisfaction. Similarly, no structured evaluation involving students, parents, teachers, or other end users was conducted. Future evaluation should therefore include a broader robustness test set, latency measurement, formal mobile and usability assessment, and structured user evaluation. From an institutional perspective, the chatbot should complement rather than replace the school's official information system. The website remains the authoritative source of institutional content, while the chatbot provides a conversational access mechanism to the same controlled knowledge base. Consequently, sustainable deployment requires clear content ownership, update procedures, version control, and administrator validation. As noted in the current implementation, each response should ideally be associated with an owner, effective date, and revision history, particularly for information that may change over time such as examination schedules and admission requirements.

The contribution of this study is primarily practical rather than algorithmic. The study does not introduce a new NLP architecture or claim a new intent-classification algorithm. Its contribution lies in demonstrating a maintainable integration of an Indonesian pretrained language model, a school-specific intent dataset, controlled retrieval of institutional responses, and a conventional web information system. The empirical results support the feasibility of this approach for the evaluated school-information domain, while the limitations regarding validation, repeated runs, baseline comparison, annotation agreement, dataset size, and broader robustness testing indicate clear directions for further study.

4. Conclusion

This study developed an intelligent school information service by integrating an IndoBERT-based intent classifier with a responsive website built with PHP, Bootstrap 5, and MySQL. A balanced dataset of 600 Indonesian utterances covering ten school-information intents was divided into 500 training and 100 held-out test utterances. The model achieved 94.0% accuracy, 94.3% macro-precision, 94.0% macro-recall, and 94.0% macro-F1 on the test set, while all ten functional testing scenarios passed.

The findings demonstrate the practical feasibility of integrating Indonesian language understanding with controlled retrieval of institutional responses for school information services. The contribution is primarily practical and system-oriented rather than a new NLP algorithm. However, the evaluation was limited by the absence of a separate validation set, multiple random-seed experiments, same-dataset baseline comparisons, formal annotator agreement, and broader robustness and user evaluation. Future work should address these limitations through stronger experimental validation and more diverse user and query testing.

Reference

[1] L. Labadze, M. Grigolia, and L. Machaidze, “Role of AI chatbots in education: systematic literature review,” *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 56, 2023, doi: 10.1186/s41239-023-00426-1.

[2] S. Gökçearslan, C. Tosun, and Z. G. Erdemir, “Benefits , Challenges , and Methods of Artificial Intelligence (AI) Chatbots in Education : A Systematic Literature Review,” *Int. J. Technol. Educ.*, vol. 7, no. 1, pp. 19–39, 2024, doi: https://doi.org/10.46328/ijte.600.

[3] T. Debets, S. K. Banihashem, D. Joosten-Ten Brinke, T. E. J. Vos, G. M. de Buy Wenniger, and G. Camp, “Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts,” *Comput. Educ.*, vol. 234, p. 105323, 2025.

[4] A. Dwiyono, A. Abdiansah, and M. Fachrurrozi, “Analisis Perbandingan Klasifikasi Intent Chatbot

Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT," vol. 6, no. 1, pp. 595–606, 2024, doi: 10.47065/josh.v6i1.6051.

[5] R. Santosa, A. B. Nusantara, and S. Imron, "Comparative Analysis of SVM and IndoBERT for Intent Classification in Indonesian Overtime Chatbots," vol. 6, no. 3, pp. 258–270, 2025.

[6] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, pp. 5641–5652, 2023.

[7] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification.," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[8] H. Rojas, R. Renteria, V. M. Duran, Y. T. Gutiérrez, M. J. I. Cabrera, and A. Aminuddin, "Mapping the evolution and future directions of ISO/IEC 25010: A bibliometric and thematic analysis," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 27530–27541, 2025.

[9] H. Cheng *et al.*, "Generative ai for requirements engineering: A systematic literature review," *Softw. Pract. Exp.*, vol. 56, no. 2, pp. 141–170, 2026.

[10] A. Muzakir, U. Suriani, and U. Ependi, "A Hybrid Semantic Enrichment Approach for Multi-Label Toxic Speech Detection.," *Int. J. Saf. Secur. Eng.*, vol. 15, no. 3, 2025.

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

[12] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 10660–10668, 2021, doi: 10.18653/v1/2021.emnlp-main.833.

[13] F. Muftie and M. Haris, "Indobert based data augmentation for indonesian text classification," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, 2023, pp. 128–132.

[14] A. Chadha, A. Abdullah, and L. Angeline, "A Comparative Performance of Optimizers and Tuning of Neural Networks for Spoof Detection Framework," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 884–890, 2022, doi: 10.14569/IJACSA.2022.01304102.

[15] A. Maesya, Y. Arifin, A. Zahra, and W. Budiharto, "Transformer-Based Semantic Search Engine with Morphophonemic Stemming for Low-Resource Sundanese Language.," *Ingénierie des Systèmes d'Information*, vol. 30, no. 11, 2025.

[16] B. E. S. Tsaviola and A. Wibowo, "Sentiment Analysis in Indonesian on Twitter Using BERT Models with Lexicon-Based Preprocessing," in *2026 International Seminar on Intelligent Business and Edge-Computing Research (ISIBER)*, IEEE, 2026, pp. 1050–1055.

[17] G. Ramdhani and S. Yuliyanti, "Fine-Tuned IndoBERT-Based Sentiment Analysis for Old Indonesian Songs Using Contextual and Generating Augmentation," *Int. J. Appl. Sci. Smart Technol.*, vol. 7, no. 2, pp. 371–384, 2025.

[18] H. F. Karim and A. P. Wibowo, "Kinerja Metode Fine-Tuning IndoBERT untuk Klasifikasi Emosi Multi-Kelas pada Teks Informal Bahasa Indonesia," *Bull. Comput. Sci. Res.*, vol. 6, no. 1, pp. 63–74, 2025.