

Comparative Evaluation of AIML and IndoBERT on Chatbots School Information Services: Performance Analysis and User Experience

Adam Herlambang^{a,1*}, Farida Ardiani^{b,2}

^{a,b} Informatika, Universitas Teknologi Yogyakarta, Jl. Siliwangi, Jombor Lor, Sendangadi, Kec.
Mlati, Kab. Sleman, Daerah Istimewa Yogyakarta 55285
e-mail: ¹herlambangadam9@gmail.com, ²ardianifarida@gmail.com,
*Corresponding author

Submitted Date: December 23th, 2025
Revised Date: March 15th, 2026

Reviewed Date: March 1st, 2026
Accepted Date: March 20th, 2026

Abstract

School information services that rely on manual staff frequently face delays and inconsistent responses. Chatbots offer an automated solution; however, selecting the appropriate technological approach for educational contexts still requires sufficient empirical evidence. This study compares two fundamentally different approaches: AIML (Artificial Intelligence Markup Language), which operates through explicit rule-based text pattern matching, and IndoBERT, a Transformer-based model fine-tuned on a labeled dataset for intent classification and combined with a vector similarity retrieval mechanism to extract answers from the school knowledge base. This comparison is essential given that the fundamental differences in how these approaches work directly affect their ability to handle language variation and their feasibility for resource-constrained school environments. Evaluation was conducted through technical testing on a labeled dataset ($n \approx 225$) and a user study involving 60 participants comprising 20 students, 20 parents, and 20 teachers. Results show that the NLP approach outperformed AIML in intent classification (91.3% vs. 72.9%), macro F1 (0.89 vs. 0.72), coverage (97% vs. 84%), resolution rate (93% vs. 78%), and user satisfaction based on SUS scores (82.4 vs. 68.1). AIML, however, achieved lower response latency (127 ms vs. 213 ms). All differences were statistically significant ($p < 0.01$). The primary contribution of this study is a measurable comparative empirical evidence between AIML and IndoBERT+retrieval for secondary school information services, accompanied by a hybrid deployment recommendation that balances answer quality, responsiveness, and institutional resource constraints.

Keywords: Chatbot; Natural Language Processing; School Information Services; Comparative Evaluation; AIML

1. Introduction

The information services in schools cover various essential needs such as new student admissions, activity schedules, fee information, contact details, and official announcements. In practice, the management of these services still heavily depends on administrative staff, which often leads to delayed responses and inconsistent information, directly affecting service accessibility for students, parents, and teachers. This situation creates a need for an automated solution that can provide fast, accurate, and always-available responses.

Chatbots have been widely recognized as an automated solution for information services that can operate 24/7. Previous studies have demonstrated their effectiveness across various domains, including academic and educational services [1][2]. However, the implementation of chatbots in school environments faces its own challenges, such as variations in users' language styles (formal, informal, and abbreviations), limitations in information technology infrastructure, and the high demand for information accuracy. Systematic reviews in the education domain emphasize that although chatbots offer many benefits, there remains a gap between the technology developed and the actual implementation needs in educational institutions [1][3].

In general, chatbot development approaches can be categorized into three main groups. First, the rule-based approach using AIML (Artificial Intelligence Markup Language), which operates through explicit text pattern matching and is deterministic in nature [4][5]. Second, the machine learning approach based on deep learning, such as LSTM and Bi-LSTM, which is capable of recognizing language patterns from training data [6][7]. Third, the Transformer-based approach, such as BERT and IndoBERT, which leverages bidirectional contextual representations and has proven to excel in various NLP tasks, including intent classification [8][9][10]. Recent studies consistently show that Transformer-based models are better at handling lexical variations and non-standard sentence structures compared to traditional rule-based approaches [9][11].

However, comparative studies that directly compare AIML and IndoBERT in the context of information services in secondary schools are still very limited. Most studies tend to focus on developing a single approach without conducting a measurable comparative evaluation [2][6], or they are conducted in different domains, such as tourism services and healthcare [12][13]. Moreover, there are still few studies that evaluate both dimensions simultaneously, namely technical performance and user experience in an integrated manner within a single experiment in a school environment. In fact, the fundamental differences between AIML's rigid, pattern-based mechanism and IndoBERT's flexible, data-driven learning approach have significantly different implications for the system's ability to handle variations in user language, as well as for the cost and ease of implementation in schools.

This study aims to conduct a measurable experimental comparison between two approaches rule-based (AIML) and NLP-based (IndoBERT + retrieval) in the context of information services at SMP Negeri 5 Membalong. To assess user-related metrics, this study employs two evaluation instruments: the System Usability Scale (SUS), which consists of 10 items rated on a 1–5 scale to measure system usability, and a user satisfaction questionnaire rated on a 1–5 scale to evaluate answer relevance, response speed, and interaction comfort. These instruments were administered to 60 respondents, consisting of students, parents, and teachers, as representatives of school information service users. The main contributions of this study are: (a) quantitative and qualitative empirical evidence regarding differences in technical performance (intent accuracy, precision/recall/F1, response time, coverage) and user metrics (resolution rate, SUS, satisfaction) between the two approaches, (b) an analysis of the operational trade-offs of each approach within school environments with limited resources, and (c) practical recommendations for hybrid implementation scenarios that can be adopted by similar educational institutions.

2. Research Methodology

a. Data and Data Preparation

This study uses primary data in the form of a collection of official FAQs and informational documents from SMP Negeri 5 Membalong, such as admission guidelines, school regulations, fee lists, activity schedules, and other administrative documents. In addition, simulated conversation scripts are included to represent variations in user queries. The initial dataset consists of 600 questions classified into 10 intent classes: (1) admissions, (2) fees, (3) schedules, (4) extracurricular activities, (5) requirements, (6) facilities, (7) contact, (8) announcements, (9) curriculum, and (10) general.

The data distribution for each intent consists of 60 samples per class. To increase linguistic variation and enrich data representation, limited data augmentation was performed. This augmentation was not intended to balance the classes, as the initial dataset was already evenly distributed, but rather to enhance the model's ability to handle both formal and informal language variations. Examples of questions in the dataset include both formal and informal expressions, such as: "When does registration open?", "When can I enroll in the school?", and "How much is the school admission fee?". The use of such language variations is important because educational chatbots must be able to understand diverse natural language inputs from users, both formal and informal [14].

Before being used, all data underwent a preparation phase that included anonymizing sensitive information, text cleaning, and content normalization to ensure consistency across sources. The normalization process involved converting text to lowercase, removing non-text characters, and normalizing entities such as dates and numbers. The data was then annotated by two independent annotators using the same annotation guidelines, and any disagreements were resolved through an adjudication

process. The level of agreement between annotators was measured using Cohen's Kappa to ensure label quality.

To increase language variation without compromising label quality, data augmentation was performed using controlled paraphrasing techniques, such as synonym replacement and back-translation. These techniques have been shown to improve the performance of text classification models, particularly on small and imbalanced datasets [15]. In addition, IndoBERT-based augmentation approaches have shown the ability to generate text variations that preserve the original semantic meaning. The dataset was then split in a stratified manner into training (70%), validation (20%), and testing (10%) sets, using a fixed seed to ensure experiment reproducibility. This approach aligns with common practices in Transformer-based NLP research, which emphasize the importance of consistent data splitting for valid and reliable model evaluation [16].

The evaluation was conducted in a phased and measurable manner. During the technical testing phase, performance was measured using metrics such as intent accuracy, precision, recall, F1-score, coverage, and average response time based on system logs. Field testing was then conducted through a user study involving 60 participants (students, parents, and teachers) with standardized task scenarios, as well as measurements of resolution rate, System Usability Scale (SUS) scores, user satisfaction, and task completion time. To ensure the validity of the results, performance differences between the systems were analyzed using appropriate statistical tests, including the Shapiro-Wilk test for normality, the paired t-test or Wilcoxon signed-rank test for paired data, and the McNemar test for proportion comparisons.

b. Research Design

The proposed research aims to compare the effectiveness of rule-based (AIML) and NLP-based chatbots in information services at SMP Negeri 5 Membalong. The chatbots will be compared using a comparative approach with two conditions (AIML and NLP), assessed using technical metrics and a user study. The proposed research design is presented in a flowchart, as shown in Figure 1 below:

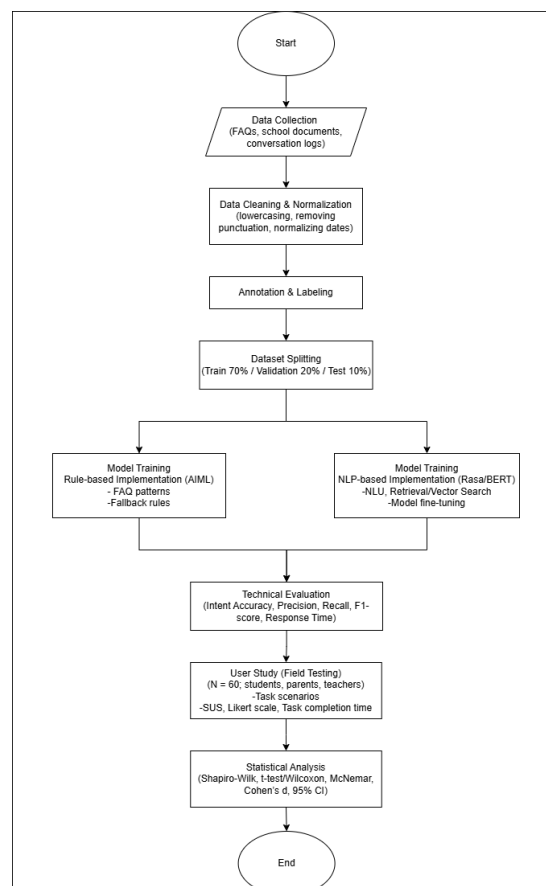


Figure 1. Research Flowchart

The flowchart in Figure 1 explains the research steps, starting with data collection and preparation, followed by parallel implementation of both chatbot systems using the same knowledge base. After implementation, a technical evaluation was conducted on a labeled test dataset, followed by a user study with 60 participants. The results of both evaluation stages were statistically analyzed to draw comparative conclusions.

c. Implementasi Chatbot

The chatbot implementation in this study began with the development of two different systems using the same data source and knowledge domain: a collection of FAQs and official information documents for SMP Negeri 5 Membalong. The first system was developed using an AIML-based rule-based approach, where question patterns are organized into several main intent groups, such as registration, fees, schedules, extracurricular activities, requirements, facilities, contact, announcements, curriculum, and general. In this implementation, AIML patterns are organized into approximately 100 basic patterns containing variations of core questions and variations of formal and informal sentences. Each pattern is mapped to a canonical answer to maintain response consistency. To handle questions that do not match any pattern, the AIML system uses a fallback mechanism by calculating the closest FAQ text similarity based on TF-IDF and cosine similarity. If the highest match score is below a threshold determined in the validation test, the system will return a fallback answer or direct the user to a clarification question.

The second system was built with a Natural Language Processing (NLP) approach using a modern NLU pipeline. The core model used is indobenchmark/indobert-base-p1, which is an IndoBERT base model pre-trained for Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) on the Indo4B corpus. This model was then fine-tuned for intent classification using the HuggingFace Transformers and PyTorch libraries. The training configuration used was a learning rate of $2e-5$, batch size of 16, epochs of 5, max sequence length of 128, AdamW optimizer, and early stopping with patience of 3. These settings were chosen to ensure stable and reproducible training, while still being suitable for Indonesian text classification tasks. BERT is designed for fine-tuning on various NLP tasks after pre-training, making this approach suitable for chatbot intent classification [17][18].

To increase language variation without compromising label quality, data augmentation was performed on the training data using controlled synonym replacement and back-translation techniques. These techniques were chosen because they have proven effective for classifying Indonesian text, especially in small or imbalanced datasets, and back-translation results are generally better than simple synonyms. Thus, augmentation not only increases the data but also helps the model recognize formal and informal language variations commonly encountered in user questions [19].

In the retrieval stage, each user question and each FAQ entry is converted into sentence embeddings using the SBERT / sentence-transformers approach, then compared using cosine similarity to obtain the most relevant answers. The SBERT approach is designed to produce sentence embeddings suitable for semantic similarity search, and research on academic FAQ chatbots has shown that SBERT + cosine similarity effectively handles both formal and informal language variations in Indonesian-language questions. The vector index can be stored and searched using FAISS for more efficient retrieval [20][21].

The implementation process for each chatbot response is recorded in a log containing the response time, predicted intent, confidence score, fallback type, and the answer retrieved from the knowledge base. The log data is then systematically analyzed to calculate predetermined technical evaluation metrics. Intent accuracy is calculated by comparing the system's predicted intents to the ground truth labels in the test dataset, while precision, recall, and macro F1-score are obtained by aggregating the classification results per intent class. Coverage is derived from the proportion of responses that do not trigger a fallback to the total number of queries, while the average response time is calculated from the difference between the request and answer timestamps in each log entry. This analysis produces quantitative data that serves as the basis for objective and reproducible comparisons of the performance of the two systems. With this implementation, both systems are ready for technical testing and user studies to compare their effectiveness.

3. Result and Discussion

a. Technical Evaluation

The technical evaluation was conducted on a test dataset of 60 data points obtained from a 10% stratified split of the total 600 data points, so that each intent was represented relatively equally. This test dataset was used to measure the ability of two chatbot models, namely rule-based (AIML) and NLP-based (IndoBERT + retrieval), to recognize intents and provide relevant answers. The test results showed that the NLP approach performed better in intent classification and answer relevance. Table 1 summarizes the key technical metrics:

Table 1. Technical Evaluation		
Metric	Rule based (AIML)	NLP based (IndoBERT + Retrieval)
Intent Accuracy	72.9%	91.3%
Precision (macro)	0.74	0.89
Recall (macro)	0.71	0.90
F1-score (macro)	0.72	0.89
Avg Response Time (ms)	127 ms	213 ms
Coverage (non-fallback)	84%	97%

In this study, coverage is defined as the percentage of questions in the test dataset that the system can answer without triggering a fallback. Therefore, a higher coverage value indicates that the system is able to answer more user questions directly. The test results show that the NLP system has higher coverage because the model can understand language variations and search for answers based on meaning, not just text pattern similarity.

Paired statistical analysis showed that the difference in intent accuracy between the two models was statistically significant (paired t-test, $t = 6.42$, $p < 0.001$) with a large effect size (Cohen's $d \approx 1.02$). The macro-F1 difference was also significant ($p < 0.001$). The average response time was lower in AIML ($p < 0.01$), but this difference was relatively small compared to the improvements in accuracy and coverage in NLP. These findings suggest that the superiority of NLP arises not only from a more modern model, but also because IndoBERT's semantic representation is able to capture synonyms, sentence structure variations, and OOV (out-of-vocabulary) that often appear in user questions. In contrast, AIML is faster because it works with direct pattern matching, but becomes less flexible when questions do not match predefined patterns.

The F1 improvement in the NLP model occurs because the model is better able to reduce false negatives on linguistically diverse questions, thus increasing recall, while simultaneously reducing false positives on semantically similar intents, thus increasing precision. In AIML, errors primarily arise when users write questions with paraphrases or structures that differ from the available patterns, for example, similar intents such as schedule and announcement are often confused. Meanwhile, errors in NLP occur more frequently when the knowledge base does not yet have the correct answer or when the question is ambiguous and requires context clarification. This explains why AIML's fallback rate is higher, at 16%, compared to NLP's 3%.

In discussion, these results indicate that NLP is superior when questions are diverse, informal, and do not always follow a standard pattern, while rule-based methods are still useful in highly structured, static domains that require very fast responses. In other words, AIML is suitable for clear and stable FAQs, while NLP is better suited for changing and more natural questions. Practical implications: schools can adopt a hybrid approach: AIML for the most frequent and stable questions, while NLP is used for questions that require deeper semantic understanding. This approach is more realistic for school environments that want to maintain service speed while improving answer quality.

b. User Study Evaluation

The user experience evaluation was conducted on 60 participants (20 students, 20 parents, and 20 teachers). Each participant ran 10 information search scenarios. Table 2 summarizes the test results:



Table 2. User Experience Evaluation

Metrik	Rule based (AIML)	NLP based
Resolution Rate	78%	93%
SUS (mean, 0–100)	68.1 (Acceptable)	82.4 (Excellent)
User Satisfaction (Likert 1–5)	3.6	4.4
Task Completion Time (s)	29.2	18.7
Fallback per 10 tasks	2.1	0.4

Non-parametric statistical tests on task completion time (Wilcoxon signed-rank test) and resolution rate (McNemar test) showed a significant difference in favor of NLP ($p < 0.01$). The difference in SUS was also significant (paired t-test, $p < 0.001$; Cohen's $d = 0.86$). These results demonstrate that the NLP model is not only technically superior but also provides tangible benefits to the user experience: more tasks can be completed without human intervention, faster completion times, and higher levels of satisfaction.

User interpretation was conducted through semi-structured interviews following the usability test, and the results revealed three main reasons for preference for the NLP model. First, answers felt more natural and relevant to informal language questions. Second, users did not need to adjust their questioning to fit the system's patterns. Third, users required fewer repeated clarifications. In contrast, participants rated AIML as fast but rigid and tended to repeat the same answer templates, resulting in a less flexible user experience. This pattern explains the higher SUS scores and user satisfaction for NLP.

Analytical evidence suggests that increased user satisfaction with NLP is directly related to three key factors: answer relevance, reduced fallbacks, and shorter task completion times. When a system is able to answer questions more accurately without forcing users to repeat or reformulate questions, perceived usability improves. This is crucial in the context of school information services, as primary users, such as students and parents, typically want quick, simple, and understandable answers. Thus, the user study results reinforce the technical finding that NLP is better suited to dynamic and diverse questions.

c. Error Analysis and Research Implications

Error analysis shows that AIML fails most frequently on questions that use paraphrases, synonyms, or sentence patterns that differ from the pre-compiled patterns. Other errors occur with semantically similar intents, such as the interchange between schedule and announcement, because pattern matching operates literally and doesn't understand the deep meaning relationships. Meanwhile, NLP tends to fail when the knowledge base doesn't yet have the correct answer, when the question is too ambiguous, or when the required context isn't available in the dataset. In other words, NLP errors are more related to the completeness of the knowledge base, while AIML errors are more related to the limited flexibility of the patterns.

These findings make it clear that this research isn't simply proving that NLP is "better," but rather explaining why the differences occur, under what conditions each method excels, and what the implications are for real-world implementations. AIML remains useful for FAQs that are highly stable, easy to maintain, and require low latency. NLP is more appropriate when users have varied, non-standard question habits and require semantic understanding. Therefore, for implementation in schools, a hybrid approach is the most rational solution: AIML handles very general and routine questions, while NLP handles more free-form and complex ones.

4. Conclusion

This study compares the effectiveness of rule-based (AIML) and NLP-based (IndoBERT + retrieval) chatbots in the information service of SMP Negeri 5 Membalong. The results show that the NLP approach consistently provides better performance than AIML, especially in the metrics of intent accuracy, F1-score, coverage, and user experience metrics such as resolution rate, System Usability Scale (SUS), and user satisfaction. This superiority is due to the NLP model's ability to understand language variations, synonyms, and non-standard sentence structures, making it more adaptive to diverse user question patterns. On the other hand, rule-based chatbots still have advantages in response speed and implementation simplicity, so they remain relevant for highly structured and stable scenarios. These findings indicate that neither approach completely replaces the other, but rather each has characteristics that are suitable for certain



conditions. Therefore, the main implication of this study is the importance of considering the needs of the implementation context, especially in school environments with limited resources and varied user needs. Based on the research results, it is recommended that chatbot implementation in educational settings use a hybrid approach, combining the speed and determinism of rule-based learning with the flexibility and language understanding capabilities of NLP. Furthermore, further development should focus on improving the quality of the knowledge base, enriching the dataset through real-world conversation logs, and exploring advanced methods such as Retrieval-Augmented Generation (RAG) and more sophisticated Indonesian language models. This way, the developed chatbot system will not only be technically accurate but also capable of providing an optimal and sustainable user experience in school information services.

References

- [1] M. Amin, K. Nazik, and A. Salwa, *Interacting with educational chatbots : A systematic review*. Springer US, 2023. doi: <https://doi.org/10.1007/s10639-022-11177-3>.
- [2] E. Mursidah, L. Ambarwati, and F. A. Karima, "Implementasi Chatbot Layanan Informasi Pendaftaran Mahasiswa Baru Program Pascasarjana Departemen eknik Informatika ITS," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 2, pp. 1–8, 2022, doi: <https://doi.org/10.21107/nero.v7i1.276>.
- [3] J. F. Urquiza-Yllescas, S. Mendoza, L. M. Sánchez-Adame, J. Rodríguez, and D. Decouchant, "Educational Chatbots: A Systematic Mapping Study," *J. Intell. \& Fuzzy Syst.*, vol. 49, no. 4, pp. 960–990, 2025, doi: <https://doi.org/10.1177/18758967251356870>.
- [4] G. Caldarini, S. Jaf, and K. McGarry, "A Literature Survey of Recent Advances in Chatbots," *Information*, vol. 13, no. 1. p. 41, 2022. doi: <https://doi.org/10.3390/info13010041>.
- [5] T. Debets, S. K. Banihashem, D. Joosten-Ten Brinke, T. E. J. Vos, G. Maillette 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, doi: <https://doi.org/10.1016/j.compedu.2025.105323>.
- [6] U. S. A. Dewi Arumsaria, Kharismab, "Sistem ChatbotLayanan Informasi Mahasiswa Menggunakan Algoritma Long Short-Term Memory," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 77–86, 2024. doi: <https://doi.org/10.30989/ijds.v2i2.1489>.
- [7] A. F. Zulhilmi, R. S. Perdana, U. Brawijaya, and P. Korespondensi, "Pengenalan Entitas Bernama Menggunakan Bi-LSTM pada Chatbot Bahasa Indonesia," vol. 11, no. 5, pp. 965–970, 2024, doi: <https://doi.org/10.25126/jtiik.1077968>.
- [8] F. Baharuddin and M. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 9, pp. 253–263, Nov. 2023, doi: <https://doi.org/10.20473/jisebi.9.2.253-263>.
- [9] A. Dwiyono, A. Abdiansah, and M. Fachrurrozi, "Analisis Perbandingan Klasifikasi Intent Chatbot Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT," *J. Inf. Syst. Res.*, vol. 6, no. 1, pp. 595–606, 2024, doi: <https://doi.org/10.47065/josh.v6i1.6051>.
- [10] N. R. P. Syallya, A. A. Pravitasari, and A. Helen, "NLP-Based Intent Classification Model for Academic Curriculum Chatbots in Universities Study Programs," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 1 SE-Artificial Intelligence, Feb. 2025, doi: <https://doi.org/10.29207/resti.v9i1.6276>.
- [11] N. Shahin and L. Ismail, "From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 271, 2024, doi: <https://doi.org/10.1007/s10462-024-10895-z>.
- [12] W. Hadinata and L. Stianingsih, "Implementasi Natural Language Processing pada Chatbot Untuk Helpdesk Informasi Wisata (Studi kasus : Tangerang Raya)," *J. Ilm. Inform.*, vol. 12, pp. 175–181, Sep. 2024, doi: <https://doi.org/10.33884/jif.v12i02.9279>.
- [13] R. Jullfikar, R. Megasari, and R. A. Sukamto, "Pengembangan Chatbot Menggunakan BERT untuk Pengelolaan Komunikasi Beasiswa Ikatan Alumni Universitas Pendidikan Indonesia," *Digit. Transform. Technol.*, vol. 5, no. 1 SE-Articles, pp. 391–400, Jul. 2025, doi: <https://doi.org/10.47709/digitech.v5i1.6367>.
- [14] A. S. Al-Ayyubi, A. Hadi, D. Novaliendry, and K. Budayawan, "Integrasi Chatbot Akademik



- Berbasis Natural Language Processing Pada Sistem Informasi Akademik di STEI Ar Risalah,” *J. Pendidik. Tambusai*, vol. 9, no. 2 SE-Articles of Research, pp. 27134–27141, Aug. 2025, doi: <https://doi.org/10.31004/jptam.v9i2.31214>.
- [15] Gilbert, S. Beatricia, A. Reinaldo, and M. Hasani, *Text Augmentation for Indonesian Intent Classification: Comparative Study*. 2024. doi: <https://doi.org/10.1109/ICITACEE62763.2024.10762810>.
- [16] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, “Revue d ’ Intelligence Artificielle Fine-Tuned IndoBERT Based Model and Data Augmentation for Indonesian Language Paraphrase Identification,” vol. 37, no. 3, pp. 733–743, 2023. doi: <https://doi.org/10.18280/ria.370322>
- [17] R. Santosa, A. B. Nusantara, and S. Imron, “Comparative Analysis of SVM and IndoBERT for Intent Classification in Indonesian Overtime Chatbots,” *J. Syst. Comput. Eng.*, vol. 6, no. 3, pp. 258–270, 2025. doi: <https://doi.org/10.61628/jsce.v6i3.2058>
- [18] Z. A. R. S. P. P. A. Annisa, “Kombinasi Intent Classification dan Named Entity Recognition pada Data Berbahasa Indonesia dengan Metode Dual Intent and Entity Transformer,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5 SE-Ilmu Komputer, pp. 1017–1024, Oct. 2024, doi: <https://doi.org/10.25126/jtiik.2024117985>.
- [19] I. A. L. H. S. Rahma, “Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6 SE-Ilmu Komputer, pp. 1329–1340, Dec. 2023, doi: <https://doi.org/10.25126/jtiik.2023107325>.
- [20] K. Peyton and S. Unnikrishnan, “A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs,” *Results Eng.*, vol. 17, p. 100856, 2023, doi: <https://doi.org/10.1016/j.rineng.2022.100856>.
- [21] R. M. Holis, P. E. P. Utomo, and B. F. Hutabarat, “Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services,” *Brill. Res. Artif. Intell.*, vol. 5, no. 2 SE-Articles, pp. 915–922, Oct. 2025, doi: <https://doi.org/10.47709/brilliance.v5i2.7027>.