Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

Reliable Intent Detection in Public Service Chatbots Using Hybrid IndoBERT and Bidirectional Long Short-Term Memory with Confidence-Based Decision Strategy

Barka Satya*[1], Mei Parwanto Kurniawan[2], Toto Indryatmoko[3], As'adurrofiq[4]

[1,2,3,4]Universitas Amikom Yogyakarta, Indonesia

Email: [1]barka.satya@amikom.ac.id

Received : Apr 7, 2026; Revised : Apr 13, 2026; Accepted : Apr 14, 2026; Published : Jun 15, 2026

Abstract

The rapid digitalization of public services has increased the demand for intelligent information systems capable of providing accurate and responsive assistance to citizens on a 24/7 basis. However, many existing public service chatbots still rely on rule-based mechanisms or single-model natural language processing (NLP) approaches, which often fail to handle linguistic variations, informal expressions, and ambiguous user queries. This study proposes a Hybrid Natural Language Understanding (NLU) architecture that integrates a fine-tuned IndoBERT model with a Bidirectional Long Short-Term Memory (BiLSTM) network to improve intent detection performance in public service chatbots. To enhance system reliability, a confidence-based decision-making mechanism is introduced, enabling the system to dynamically select the most reliable prediction or activate a fallback pattern-matching module when confidence thresholds are not met. The proposed approach was evaluated on a custom dataset comprising 53 public service intents, spanning formal and informal Indonesian language use. Experimental results demonstrate that the hybrid architecture achieves an intent classification accuracy of 86.8%, outperforming single-model approaches while maintaining an acceptable response time for practical deployment, particularly in public service scenarios where accuracy and reliability are prioritized over response speed. Furthermore, integrating a continuous learning mechanism enables the system to adapt to low-confidence queries over time, thereby improving robustness in real-world applications. These findings indicate that hybrid NLP architectures with confidence-aware decision mechanisms offer a practical and scalable solution for intelligent public service chatbots.

Keywords: *Chatbot, Confidence-Based Decision, Hybrid NLP, IndoBERT, Intent Detection, LSTM*

1. INTRODUCTION

Digital transformation has significantly reshaped how governments provide services and information to citizens. In recent years, many public institutions have begun adopting artificial intelligence technologies to improve the accessibility, efficiency, and responsiveness of public services. Among these technologies, conversational agents or chatbots have become an increasingly popular solution because they enable automated interaction between citizens and digital systems without requiring continuous human involvement. By providing information services that are available at any time, chatbots can reduce operational costs while improving service availability in digital government environments [1],[2].

Despite these advantages, implementing chatbots in public service environments presents several challenges. Citizens often communicate using informal language, abbreviations, typographical errors, and mixed linguistic expressions when interacting with conversational systems. These variations can significantly degrade the performance of natural language understanding models, particularly when chatbots rely on rule-based mechanisms or single-model classification. Consequently, the system may

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

fail to correctly identify user intent or generate responses that do not correspond to the actual request [3],[4].

Recent advances in natural language processing have led to the development of powerful deep learning models capable of capturing complex semantic relationships within textual data. Transformer-based architectures, especially those derived from the Bidirectional Encoder Representations from Transformers (BERT) model, have demonstrated substantial improvements in various NLP tasks, including text classification, question answering, and conversational intent detection. These models utilize self-attention mechanisms to capture contextual relationships between words, enabling more accurate semantic representation compared with traditional machine learning approaches [5],[6].

However, transformer-based architectures also present several practical limitations. Although they provide strong contextual representations, these models may produce unstable predictions when processing noisy or unfamiliar inputs. In contrast, recurrent neural network architectures such as Long Short-Term Memory (LSTM) networks remain effective for modeling sequential relationships in textual data and often provide efficient inference for short conversational queries. Because of these complementary strengths and weaknesses, combining transformer models with sequential neural networks has become an attractive strategy for improving conversational AI performance [7], [8].

Hybrid NLP architectures that integrate transformer-based representations with recurrent neural networks have recently received increasing attention in conversational AI research. Such hybrid approaches aim to leverage the contextual representation capability of transformers while maintaining the efficiency and sequential learning properties of recurrent networks. Several studies have shown that hybrid models can improve text classification accuracy and enhance model stability across diverse conversational inputs [9],[10],[11].

Another emerging concern in conversational AI research is the reliability of model predictions. Neural network models may produce predictions even when their confidence level is relatively low. In sensitive application domains such as public services, inaccurate responses may mislead users and reduce trust in digital government systems. Therefore, recent studies emphasize the importance of integrating confidence-aware mechanisms that allow conversational systems to evaluate prediction reliability before generating responses [12],[13].

In addition to prediction reliability, conversational systems must also adapt to evolving user interaction patterns. Language usage in real-world environments constantly changes, particularly in online communication contexts where informal expressions and slang frequently emerge. Continuous learning approaches allow conversational systems to collect and analyze difficult queries, which can later be incorporated into training datasets to improve future model performance [14],[15].

Furthermore, recent research highlights the growing role of conversational AI in public sector digital transformation. Intelligent conversational systems are increasingly used to support administrative processes, citizen information services, and government communication platforms. As governments continue to digitize their services, reliable and adaptable chatbot architectures become an essential component of modern public service systems [16],[17].

Motivated by these challenges, this study proposes a hybrid Natural Language Understanding architecture that integrates IndoBERT and Bidirectional Long Short-Term Memory (BiLSTM) models for intent detection in public service chatbots. In addition, a confidence-based decision mechanism is introduced to enhance prediction reliability and activate fallback strategies when model confidence is insufficient. The proposed system also incorporates a continuous learning mechanism that records low-confidence user queries to support future model improvements.

The main contributions of this research can be summarized as follows:

1. The development of a hybrid IndoBERT–BiLSTM architecture for intent detection in public service chatbots.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

2. The integration of a confidence-based decision mechanism to improve response reliability.
3. The implementation of a continuous learning framework that enables adaptive system improvement.

Through experimental evaluation using a dataset of public service queries, the proposed approach aims to demonstrate that hybrid architectures combined with confidence-aware decision strategies can improve intent classification performance while maintaining response efficiency suitable for real-world deployment.

2. METHOD

2.1. System Overview

This study proposes a hybrid Natural Language Understanding (NLU) architecture designed to improve intent detection performance in public service chatbots. The proposed system integrates two complementary deep learning models: IndoBERT and Bidirectional Long Short-Term Memory (BiLSTM). IndoBERT is utilized to capture contextual semantic representations from user queries, while BiLSTM is employed to model sequential patterns within textual data. By combining these two approaches, the system aims to achieve both contextual understanding and efficient sequential learning capabilities in conversational text processing [18],[19].

In the proposed architecture, both models run in parallel during inference. The user query is processed simultaneously by the IndoBERT model and the BiLSTM classifier, enabling the system to generate multiple intent predictions. This parallel design enables the system to leverage the strengths of both transformer-based contextual representations and recurrent neural network sequential modeling.

In addition to the hybrid architecture, the system incorporates a confidence-based decision mechanism that evaluates prediction reliability before determining the final intent classification. When the primary model's confidence score falls below a predefined threshold, the system activates a fallback mechanism based on pattern matching to ensure the chatbot can still provide meaningful responses. This design helps reduce the risk of incorrect responses in public service environments where information accuracy is essential [20],[21].

Furthermore, the architecture integrates a continuous learning component that records user queries associated with low-confidence predictions. These queries are stored for further analysis and can later be incorporated into the training dataset during subsequent model updates. Continuous learning mechanisms enable conversational AI systems to adapt to evolving language patterns and improve their performance over time [22],[23] . The overall architecture of the proposed hybrid NLU system is illustrated in Figure 1.

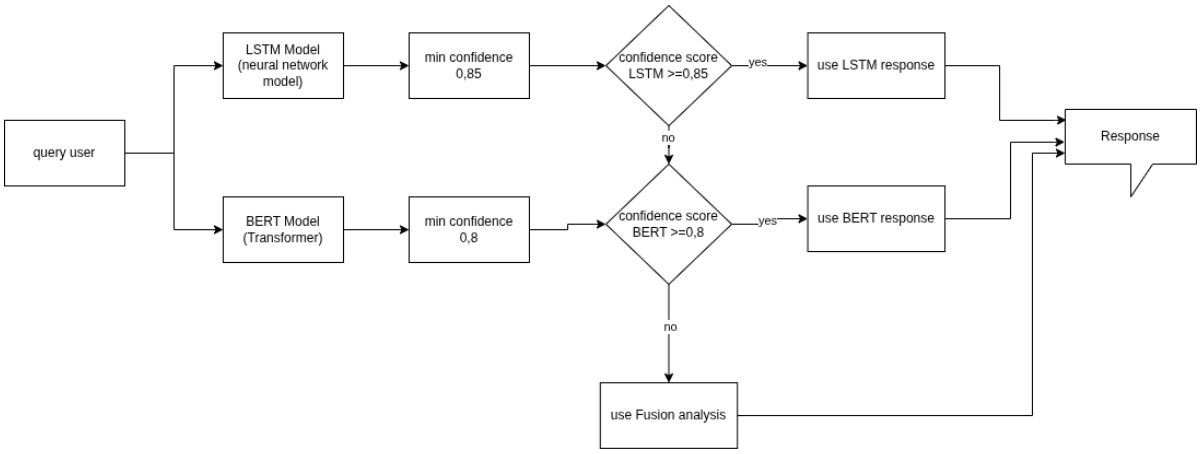

Figure 1. Proposed Hybrid NLU Architecture for Public Service Chatbots

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

2.2. Dataset Collection

The dataset used in this study was compiled from several public service information domains. User queries were categorized into 53 intent classes representing common citizen information needs, including administrative procedures, document requirements, and other inquiries related to public services.

To reflect realistic conversational scenarios, the dataset includes both formal and informal language expressions commonly used by citizens when interacting with digital systems. Variations such as abbreviations, typographical errors, and colloquial expressions were intentionally preserved to simulate natural chatbot interactions.

Previous studies indicate that incorporating linguistic diversity during dataset preparation can significantly improve the robustness of conversational systems deployed in real-world environments [24],[25]. Some intent labels appear more than once in the table because they represent variations of user queries collected from different conversational contexts within the same intent category. The distribution of the dataset across different intent categories is presented in Table 1.

Table 1. Distribution of Intent Classes in the Dataset

No	Intent	No	Intent	No	Intent
1	salam	19	pn_info	37	akta_lahir_info
2	terimakasih	20	dpmptsp_info	38	akta_mati_info
3	bantuan	21	imigrasi_info	39	akta_general
4	saran	22	disdukcapil_info	40	kk_info
5	keluhan	23	kemenag_info	41	surat_pindah_luar
6	out_of_topic	24	dinsos_info	42	surat_pindah_masuk
7	bappenda_info	25	bank_tegal_info	43	surat_pindah_dalam
8	kpp_info	26	bpjs_kerja_info	44	pindah_general
9	pdam_info	27	bpjs_sehat_info	45	loakk_info
10	bank_jateng_info	28	jaksa_info	46	sicantik_info
11	polres_info	29	mpp_info	47	ktp_info
12	samsat_info	30	nib_info	48	out_of_topic
13	pos_info	31	ak1_info	49	kabar
14	pupr_info	32	itr_info	50	bot_info
15	hukum_info	33	sls_info	51	miss_info
16	taspen_info	34	sls_info	52	kominfo_info
17	perintransnaker_info	35	nib_info	53	kabar
18	koperasi_info	36	lkpm_info		

2.3. Text Preprocessing

Before training the models, several preprocessing procedures were applied to clean and normalize the textual data. These preprocessing steps aim to reduce noise while preserving the semantic information contained in user queries.

The preprocessing stage includes case folding, punctuation normalization, tokenization, and slang normalization. Case folding converts all characters to lowercase format, while punctuation normalization removes unnecessary symbols that may interfere with model training. Tokenization is then applied to split sentences into individual tokens that serve as input units for the deep learning models.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

Normalization techniques are particularly important in conversational datasets because users often write messages using informal or abbreviated expressions. Proper preprocessing improves the consistency of input data and enables deep learning models to learn more reliable linguistic representations [26]. Examples of preprocessing rules applied in this study are summarized in Table 2.

Table 2. Text Preprocessing Rules and Examples

Preprocessing Stage	Rule / Logic Description	Input Example	Output Example
Case Folding & Stripping	Converts all characters to lowercase and removes leading/trailing spaces.	" HOW To Make a Family Card?! "	"how to make a family card"
Gibberish Filtering	Filters out strings with repetitive characters (>4x) or impossible consonant clusters (>5x).	"aaaaaa" or "asdfghj"	(Discarded/Empty)
Critical Correction	Fixes domain-specific typos that change the core meaning of the query.	"bapenda", "kt"	"bappenda", "ktp"
Regex Normalization	Uses pattern matching to standardize varied spellings of specific terms.	"bappppenda"	"bappenda"
Slang Normalization	Converts informal/abbreviated words into formal Indonesian (Applied in Aggressive mode).	"gimana", "pkl"	"bagaimana", "jam"
Whitespace Normalization	Compresses multiple whitespaces or tabs into a single space.	"check status"	"check status"
Contextual Cleaning (BERT)	Preserves semantic punctuation (?, !, ., ,) to maintain sentence structure.	"where is it?!"	"where is it?!"
Aggressive Cleaning (LSTM)	Removes all punctuation and non-alphanumeric characters to minimize feature noise.	"where is it?!"	"where is it"

2.4. Hybrid IndoBERT–BiLSTM Model

The main component of the proposed system is a hybrid model that integrates IndoBERT and BiLSTM architectures for intent classification.

In this architecture, IndoBERT first processes the user query to generate contextual embeddings that represent semantic relationships between words in the sentence. These embeddings allow the system to capture deeper contextual meaning compared with traditional bag-of-words or n-gram representations.

In parallel, the BiLSTM model processes tokenized query sequences to capture sequential dependencies within the sentence. BiLSTM networks analyze textual sequences in both forward and backward directions, enabling the model to understand contextual dependencies between words across the entire query.

By combining contextual representations from transformer-based models with sequential learning from recurrent neural networks, the hybrid architecture leverages the strengths of both approaches and improves intent classification performance [27]. The structure of the hybrid IndoBERT–BiLSTM model is illustrated in Figure 2.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

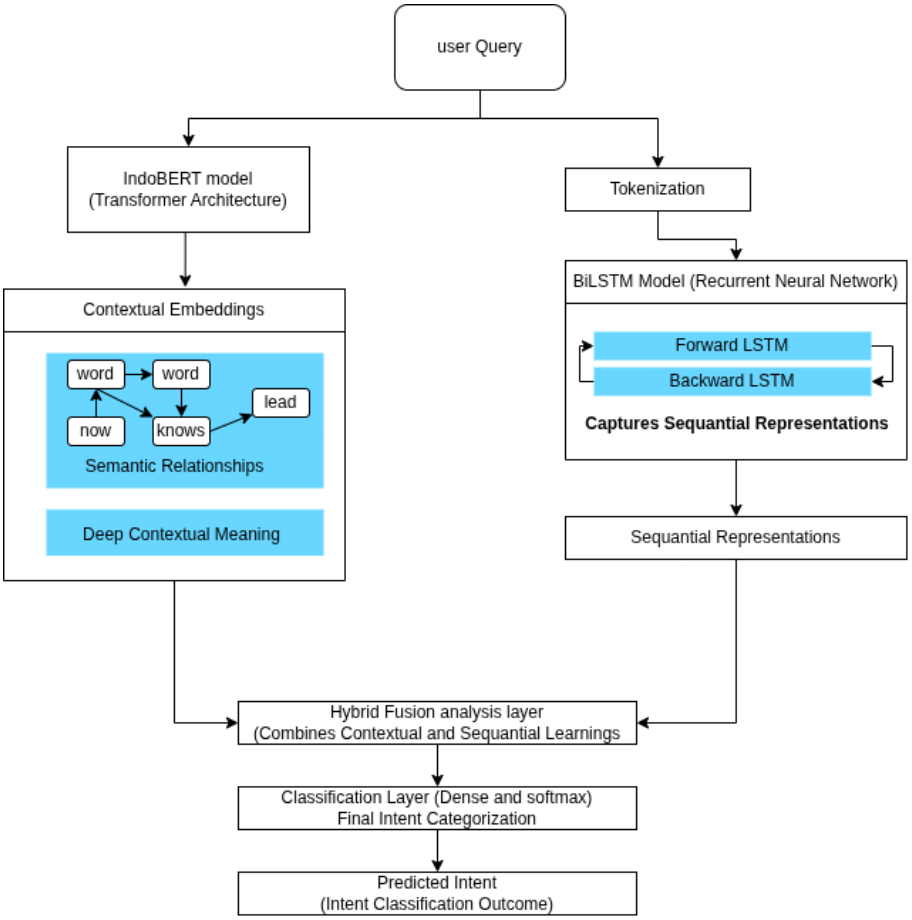

Figure 2. Hybrid IndoBERT–BiLSTM Model Architecture

2.5. Confidence-Based Decision Mechanism

To improve prediction reliability, a **confidence-based decision mechanism** is implemented during classification. Each predicted intent generated by the models is associated with a probability score representing the confidence of the prediction.

Let P_{max} denote the highest probability score among all predicted intent classes. If P_{max} is greater than or equal to a predefined threshold θ, the predicted intent is accepted as the final classification result. Otherwise, the system activates a fallback mechanism based on rule-based pattern matching. The final intent prediction is determined using a confidence-based decision rule, as defined in Equation (1).

$$Intent = \begin{cases} argmax(P_i), & \text{if } P_{max} \geq \theta \\ Fallback, & \text{otherwise} \end{cases} \quad (1)$$

Where:
P_i represents the predicted probability for intent class i,
P_{max} denotes the highest probability among all predicted classes, and
θ is the predefined confidence threshold.

This mechanism ensures that only high-confidence predictions are accepted, while low-confidence cases are handled using a fallback strategy to improve system reliability.

Confidence-aware decision mechanisms have been widely adopted in conversational AI systems to prevent unreliable predictions and improve system trustworthiness [28]. The workflow of the confidence-based decision process is illustrated in Figure 3.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

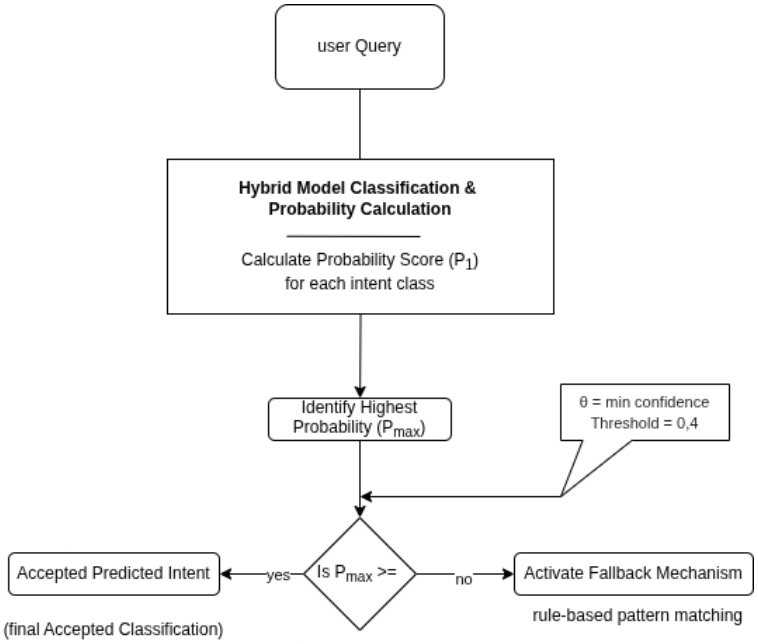

Figure 3. Confidence-Based Decision Process

The models were trained using the Adam optimizer with a learning rate of 2e-5 for IndoBERT and 1e-3 for the BiLSTM model. The batch size was set to 32, and training was conducted for 10 epochs. The dataset was split into training, validation, and test sets at 80:10:10 to ensure reliable model evaluation.

3. RESULT

This section reports the experimental results of the proposed hybrid NLU system, focusing on classification performance, error characteristics, fallback effectiveness, latency, and continuous learning outcomes.

3.1. Intent Classification Performance

The performance of the proposed system was evaluated by comparing three different classification approaches: a standalone LSTM model, a standalone IndoBERT model, and the proposed hybrid IndoBERT–BiLSTM architecture. All experiments were conducted using a dataset consisting of 53 intent classes representing public service queries.

The evaluation metrics used in this study include accuracy, precision, recall, and F1-score. These metrics were calculated from the classification results on the test dataset. Table 3 presents the overall performance of the three evaluated models.

Table 3. Classification Performance Comparison

Metode	Accuracy	Precision	Recall	F1-Score
LSTM only	78.9%	78.9%	78.9%	78.9%
Bert only	82.9%	82.9%	82.9%	82.9%
Hybrid System	86.8%	86.8%	86.8%	86.8%

The identical values across evaluation metrics indicate a balanced classification performance across intent classes, although further analysis such as confusion matrix evaluation is required to provide deeper insights into class-level performance.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

The results show that the hybrid architecture achieves the highest classification accuracy among the three evaluated models.

3.2. Performance of the LSTM Model

The first experiment evaluated the LSTM-only model on the test dataset. The model was trained to classify user queries based on sequential patterns in the text input. The evaluation results show that the LSTM model correctly classified 78.9% of the test queries, while 21.1% were misclassified. The model's average confidence score was 62.5%. The distribution of the LSTM model's classification results is illustrated in Figure 4.

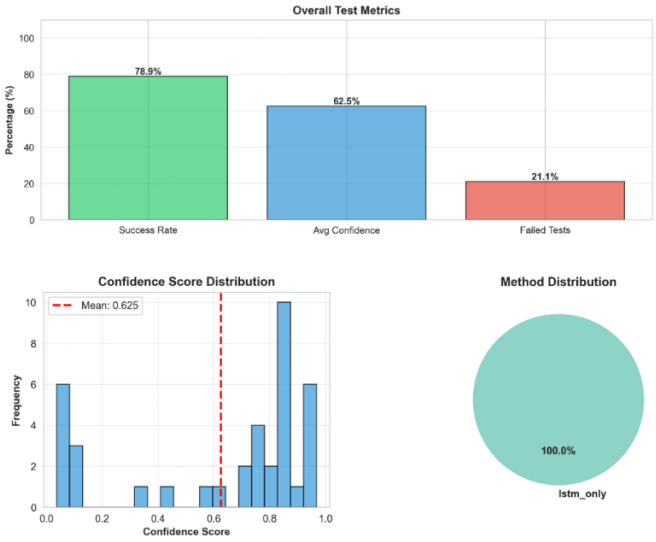

Figure 4. Overall Performance of the LSTM Model

3.3. Performance of the BERT Model

The second experiment evaluated the performance of the BERT-only model using the same dataset. The model achieved a success rate of 82.9%, an improvement over the LSTM-only model. The average confidence score generated by the BERT model was 86.7%, while the proportion of failed predictions was 17.1%. The overall performance metrics of the BERT model are presented in Figure 5.

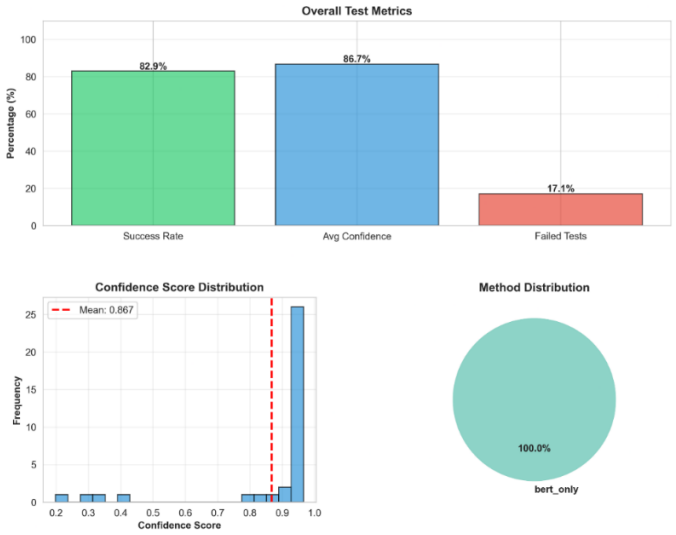

Figure 5. Overall Performance of the BERT Model

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

3.4. Performance of the Hybrid IndoBERT–BiLSTM Model

The final experiment evaluated the proposed Hybrid IndoBERT–BiLSTM architecture, which combines predictions from both models using a confidence-based decision mechanism. The hybrid model achieved the highest classification performance among all tested approaches. The system correctly classified 86.8% of the test queries, with a failure rate of 13.2%. The average confidence score generated by the hybrid model was 74.0%. The distribution of the classification results is presented in Figure 6.

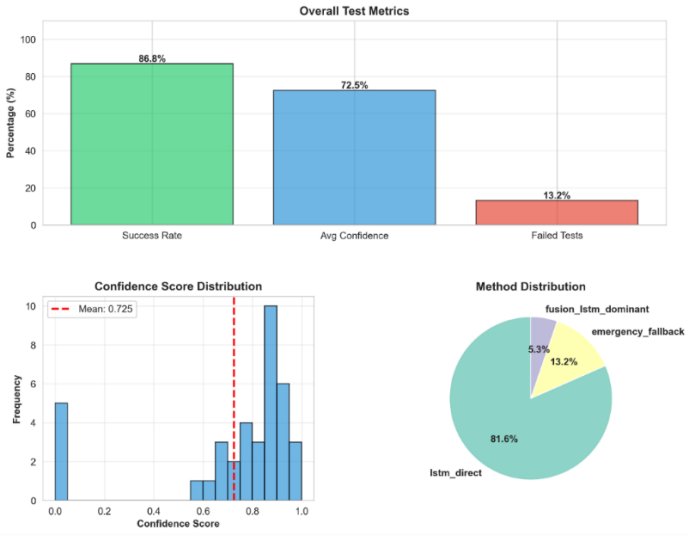

Figure 6. Overall Performance of the Hybrid IndoBERT–BiLSTM Model

3.5. Response Time (Latency)

In addition to classification accuracy, system responsiveness was also evaluated. The response time was measured from the moment a user query was received by the FastAPI server until the chatbot returned a response to the user through the Telegram platform. Table 4 summarizes the average latency observed during the evaluation process.

Table 4. System Response Time

Model Type	Test Status	Avg. Response Time (ms)	Response Time (s)
LSTM (Single)	Average (Success & Failed)	13,650 ms	13.65 s
BERT (Single)	Average (Success & Failed)	13,650 ms	13.65 s
Hybrid (LSTM + BERT)	Successful	15,850 ms	15.85 s
Hybrid (LSTM + BERT)	Failed	17,150 ms	17.15 s
Hybrid (Overall)	Total Average	16,500 ms	16.50 s

4. DISCUSSION

The experimental results demonstrate that integrating contextual language models with sequential neural networks can significantly improve intent detection performance in public service chatbot systems. The hybrid IndoBERT–BiLSTM architecture consistently outperformed the individual models across all evaluation metrics, indicating that the combination of contextual semantic understanding and sequential pattern learning provides a more robust representation of user queries.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

The LSTM-only model achieved an accuracy of 78.9%, indicating that sequential neural networks can capture structural patterns in textual data. However, the relatively lower confidence scores observed in this model suggest that purely sequential architectures may struggle to interpret semantic nuances in conversational language. This limitation has also been reported in previous studies, where recurrent neural networks often struggle to process complex contextual relationships in natural language inputs [29].

In contrast, the BERT-based model demonstrated improved performance, achieving 82.9% accuracy and a significantly higher average confidence score. Transformer architectures such as BERT are designed to model contextual relationships between words using self-attention mechanisms, allowing the model to capture semantic dependencies across entire sentences rather than relying solely on sequential order. As a result, BERT-based models have become widely adopted in modern NLP systems and have shown superior performance in tasks such as text classification, sentiment analysis, and conversational intent detection [30],[24].

Despite this improvement, the results also indicate that transformer-based models alone may not always provide optimal performance in conversational environments. Although BERT effectively captures contextual meaning, it may occasionally produce unstable predictions when dealing with noisy input, informal language, or rare linguistic patterns. Conversational systems must be able to process diverse linguistic expressions and maintain stable language understanding performance when interacting with users. Several studies have proposed methods to improve natural language understanding capabilities in conversational agents without degrading model performance across different input variations [31].

The hybrid IndoBERT–BiLSTM architecture proposed in this study addresses these limitations by combining the contextual understanding of transformer models with the sequential pattern-recognition ability of recurrent neural networks. The hybrid model achieved 86.8% accuracy, demonstrating that integrating these two learning mechanisms can improve classification reliability. Similar improvements have been observed in hybrid NLP architectures that combine transformer embeddings with recurrent layers to enhance performance in text classification tasks [27],[32].

Another important aspect of the proposed system is the implementation of a confidence-based decision mechanism. This mechanism evaluates the probability score of each prediction before determining the final classification result. When the predicted confidence score falls below the predefined threshold, the system activates a fallback strategy based on rule-based pattern matching. This approach helps reduce the risk of incorrect responses, particularly in public service applications where information accuracy is critical. Confidence-aware decision strategies have been increasingly explored in conversational AI research to improve system reliability and user trust [33],[34].

The integration of a fallback mechanism also contributes to system robustness when the model encounters ambiguous or previously unseen queries. Instead of returning an incorrect classification, the system attempts to provide a reasonable response using predefined patterns. This strategy ensures that the chatbot remains functional even when the machine learning model fails to produce a reliable prediction.

In addition to accuracy improvements, the proposed architecture maintains a reasonable level of system responsiveness. The average response time observed during the experiment was approximately 16.5 seconds, including preprocessing, parallel inference, and hybrid decision-making. Although the response time is higher than that of typical real-time chatbot systems, the observed latency remains acceptable for public service applications where response accuracy and reliability are prioritised over response speed. This trade-off reflects the system design objective of minimising incorrect responses in sensitive service domains. Similar latency characteristics have also been reported in other NLP-based

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

conversational systems that integrate deep learning models with practical deployment environments [35],[34].

Another key contribution of this work is the implementation of a continuous learning mechanism that records low-confidence queries during system operation. These queries can later be analysed and incorporated into the training dataset to improve future model performance. Continuous learning approaches have been widely recognised as an effective strategy for maintaining the long-term adaptability of conversational AI systems, particularly in dynamic environments where language usage evolves over time [36],[37].

Overall, this study's findings demonstrate that hybrid NLP architectures combined with confidence-aware decision strategies can significantly improve the reliability and adaptability of chatbot systems for public service applications. By leveraging the strengths of both transformer-based contextual modelling and sequential neural networks, the proposed approach provides a practical framework for developing intelligent conversational systems capable of handling diverse and dynamic user queries.

Despite the promising results, this study has several limitations. First, the system's response time is relatively higher than that of conventional real-time chatbot systems, which may affect user experience in time-sensitive applications. Second, the dataset used in this study is domain-specific and limited to public service queries in a particular region, potentially limiting the model's generalizability. Future research should focus on optimising model efficiency and expanding the dataset to include more diverse linguistic variations and multilingual scenarios.

5. CONCLUSION

This study proposed a hybrid Natural Language Understanding architecture that integrates IndoBERT and Bidirectional Long Short-Term Memory (BiLSTM) models to improve intent detection in public service chatbots. The proposed system combines the contextual semantic representation capability of transformer-based models with the sequential learning ability of recurrent neural networks. In addition, a confidence-based decision mechanism was incorporated to evaluate prediction reliability and activate a fallback strategy when the predicted confidence score does not meet a predefined threshold.

An experimental evaluation using a dataset of 53 public service intent categories demonstrated that the proposed hybrid architecture achieved the best classification performance among the evaluated models. The hybrid IndoBERT–BiLSTM model achieved 86.8% accuracy, outperforming the standalone LSTM and BERT models. These results indicate that integrating contextual embeddings with sequential pattern learning can improve the reliability of intent classification in conversational systems.

The proposed system also maintained an acceptable response time, with an average latency of approximately 16.5 seconds when deployed within a chatbot environment using FastAPI and Telegram integration. This response time indicates that the system can support practical public service interactions where response reliability is more critical than instantaneous response speed.

Overall, this study's findings demonstrate that hybrid NLP architectures combined with confidence-aware decision mechanisms can enhance the robustness and reliability of conversational agents in public service environments. Such systems have the potential to support government digital services by providing continuous, automated information assistance to citizens.

Future research may expand the dataset to include a broader range of conversational queries and multilingual expressions. In addition, further investigation could explore integrating larger language models or adaptive learning mechanisms to improve the system's ability to handle unseen user queries and evolving linguistic patterns.

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] A. Gr, "The impact of chatbots on public service provision : A qualitative interview study with citizens and public service providers," vol. 41, no. February, 2024, doi: 10.1016/j.giq.2024.101927.

[2] X. Li and J. Wang, "Should government chatbots behave like civil servants ? The effect of chatbot identity characteristics on citizen experience," *Gov. Inf. Q.*, vol. 41, no. 3, p. 101957, 2024, doi: 10.1016/j.giq.2024.101957.

[3] A. M. Al-ansi, A. Garad, M. Jaboob, and A. Al-ansi, "Heliyon Elevating e-government : Unleashing the power of AI and IoT for enhanced public services," *Heliyon*, vol. 10, no. 23, p. e40591, 2024, doi: 10.1016/j.heliyon.2024.e40591.

[4] G. Papageorgiou, V. Sarlis, and M. Maragoudakis, "Enhancing E-Government Services through State-of-the-Art , Modular , and Reproducible Architecture over Large Language Models," 2024, doi: https://doi.org/10.3390/app14188259.

[5] A. R. Y. S. Prihatmanto and R. Andrian, "Transforming Public Services : A Systematic Review of Smart Government Frameworks , Architectures , and Implementation Challenges," *IEEE Access*, vol. 12, no. September, pp. 135799–135810, 2024, doi: 10.1109/ACCESS.2024.3450907.

[6] S. Hemesath and M. Tepe, "Public value positions and design preferences toward AI-based chatbots in e-government . Evidence from a conjoint experiment with citizens and municipal front desk officers," *Gov. Inf. Q.*, vol. 41, no. 4, p. 101985, 2024, doi: 10.1016/j.giq.2024.101985.

[7] M. H. Miraz, A. Ya, S. Adeyinka-ojo, and J. B. Sarkar, "Intention to use determinants of AI chatbots to improve customer relationship management efficiency," *Cogent Bus. Manag.*, vol. 11, no. 1, p., 2024, doi: 10.1080/23311975.2024.2411445.

[8] 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, doi: DOI:10.61628/jsce.v6i3.2058.

[9] C. Ouaddi, L. Benaddi, A. Jakimi, A. Chehri, and R. Saadane, "ScienceDirect ScienceDirect Assessing Machine Learning Models for Enhancing Intent and Chatbots Detection in Tourism Assessing Machine Learning Models for Enhancing Intent Detection in Tourism Chatbots," *Procedia Comput. Sci.*, vol. 270, pp. 3162–3171, 2025, doi: 10.1016/j.procs.2025.09.441.

[10] R. Aulia and A. Purnama, "Development of an Intent-Classification Chatbot to Support Operational Services at Kadin Indonesia," vol. 5, no. 2, pp. 1171–1180, 2025, doi: https://doi.org/10.47709/brilliance.v5i2.7438.

[11] Y. Zhang and R. Y. K. Lau, "Uncertainty Detection : A Multi - View Decision Boundary Approach Against Healthcare Unknown Intents," pp. 1–17, 2025, doi: https://doi.org/10.3390/app15137114.

[12] I. Dube, "Chatbots: a tool to improve public service delivery and create public value," vol. 9, no. 3, pp. 43–64, 2024, doi: https://doi.org/10.55190/JPADA.2024.341.

[13] M. Saleem and J. Kim, "Intent aware data augmentation by leveraging generative AI for stress detection in social media texts," 2024, doi: 10.7717/peerj-cs.2156.

[14] A. Skuridin and M. Wynn, "Chatbot Design and Implementation : Towards an Operational Model for Chatbots," 2024, doi: https://doi.org/10.3390/info15040226.

[15] S. Senadheera *et al.*, "Understanding Chatbot Adoption in Local Governments : A Review and Framework Understanding Chatbot Adoption in Local Governments :," *J. Urban Technol.*, vol. 32, no. 3, pp. 35–69, 2025, doi: 10.1080/10630732.2023.2297665.

[16] C. Yu, J. Yan, and N. Cai, "ChatGPT in higher education : factors influencing ChatGPT user satisfaction and continued use intention," no. May, pp. 1–11, 2024, doi: 10.3389/feduc.2024.1354929.

[17] U. Informatics, D. Li, S. Wang, B. Zhao, Z. Ma, and L. Li, "A novel model based on a transformer for intent detection and slot filling," *Urban Informatics*, 2024, doi: 10.1007/s44212-

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

024-00056-6.

[18] W. Christian *et al.*, "Leveraging Leveraging IndoBERT IndoBERT and and DistilBERT DistilBERT for for Indonesian Indonesian emotion emotion Leveraging IndoBERT and DistilBERT for Indonesian classification in reviews classification in e-commerce reviews emotion classification in e-commerce reviews ScienceDirect," *Procedia Comput. Sci.*, vol. 269, pp. 321–330, 2025, doi: 10.1016/j.procs.2025.08.284.

[19] S. Shreyashree, P. Sunagar, S. Rajarajeswari, and A. Kanavalli, "BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings," vol. 13, no. 9, pp. 667–675, 2022, doi: DOI:10.14569/IJACSA.2022.0130979.

[20] J. Bae and S. Bum, "Accuracy-informed label smoothing and logit scaling for deep neural network calibration," *Appl. Soft Comput.*, vol. 188, no. November 2025, p. 114410, 2026, doi: 10.1016/j.asoc.2025.114410.

[21] D. Lopez-paz and M. A. Ranzato, "Gradient Episodic Memory for Continual Learning," no. Nips, 2022, doi: https://doi.org/10.48550/arXiv.1706.08840.

[22] N. Brunswick and U. States, "Continual Learning of Large Language Models : A Comprehensive Survey Continual Learning of Large Language Models : A Comprehensive Survey," vol. 58, no. 5, 2026, doi: 10.1145/3735633.

[23] Z. Ke and B. Liu, "Continual Learning of Natural Language Processing Tasks: A Survey," vol. 2, 2023, doi: https://doi.org/10.48550/arXiv.2211.12701.

[24] M. Jbene, G. Jeon, A. Chehri, and R. Saadane, "Intent detection for task-oriented conversational agents : A comparative study of recurrent neural networks and transformer models," no. June 2024, pp. 1–20, 2025, doi: 10.1111/exsy.13712.

[25] A. Souha, C. Ouaddi, L. Benaddi, and A. Jakimi, "Pre-Trained Models for Intent Classification in Chatbot : Comparative Study and Critical Analysis," *2023 6th Int. Conf. Adv. Commun. Technol. Netw.*, pp. 1–6, 2023, doi: 10.1109/CommNet60167.2023.10365312.

[26] A. Tyagi, "Benchmark Text Preprocessing Techniques in Natural Language Processing," *2024 4th Int. Conf. Innov. Sustain. Comput. Technol.*, pp. 1–6, 2024, doi: 10.1109/CISCT62494.2024.11134188.

[27] R. A. Khachfeh, "An Enhanced Hybrid BERT-BiLSTM Learning Model for Arabic News Classification," *2025 Int. Conf. Mach. Intell. Smart Innov.*, pp. 201–206, 2025, doi: 10.1109/ICMISI65108.2025.11115581.

[28] H. Kyerematen-boateng, "Choosing LS-Stat Confidence Scores for Neural Networks Predictions," *2024 Int. Conf. E-mobility, Power Control Smart Syst.*, pp. 1–6, 2024, doi: 10.1109/ICEMPS60684.2024.10559353.

[29] M. Hossain, S. Hossain, M. Safran, S. Alfarhood, M. Alfarhood, and M. F. Mridha, "A Hybrid Attention-Based Transformer Model for Arabic News Classification Using Text Embedding and Deep Learning," *IEEE Access*, vol. 12, no. December, pp. 198046–198066, 2024, doi: 10.1109/ACCESS.2024.3522061.

[30] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, 2024, doi: 10.1186/s40537-023-00842-0.

[31] M. Arevalillo-herráez, R. S. A. Luise, and Y. Wu, "Neurocomputing A non-deteriorating approach to improve Natural Language Understanding in Conversational Agents," vol. 630, no. February, 2025, doi: https://doi.org/10.1016/j.neucom.2025.129652.

[32] B. R. K, B. E. Babu, S. P. Racharla, P. Pavani, Y. Balagoni, and M. Ajmeera, "A Performance Evaluation of Transformer Models and Recurrent Neural Networks Models in Efficient Text Classification Tasks," *2025 8th Int. Conf. Comput. Methodol. Commun.*, pp. 1171–1177, 2025, doi: 10.1109/ICCMC65190.2025.11140627.

[33] W. Hybrid and C. Architecture, "Arsitektur Chatbot WhatsApp Hibrida Rasa - DeepSeek : Desain dan Evaluasi Performa Performance Evaluation," vol. 15, pp. 446–454, 2026.

[34] T. Papadopoulos and C. Alexopoulos, "Evaluating chatbot architectures for public service delivery : balancing functionality , safety , ethics , and adaptability," no. October, 2025, doi: 10.3389/fpos.2025.1601440.

[35] M. Abbasi, F. Cardoso, and P. Martins, "Comparative Performance Analysis of Large Language

Vol. 7, No. 3, June 2026, Page. 2906-2919
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5794

Models for Structured Data Processing : An Evaluation Framework Applied to Bibliometric Analysis," pp. 1–30, 2026, doi: https://doi.org/10.3390/app16020669.

[36] L. Zhang, S. Wang, F. Yuan, B. Geng, and M. Yang, "Lifelong language learning with adaptive uncertainty regularization," *Inf. Sci. (Ny).*, vol. 622, pp. 794–807, 2023, doi: 10.1016/j.ins.2022.11.141.

[37] M. N. Budiyanto and A. Syafebri, "Artificial Intelligence in Public Administration : A Systematic Literature Review on Opportunities and Ethical Challenges," pp. 244–257, 2025, doi: 10.18502/kss.v10i30.20347.