Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

A Dental Chatbot Based on IndoBERT with Next Sentence Prediction and Intent Classification

Nadhief Athallah Isya[1*], Rasim[2], Ani Anisyah[3]
[1,2,3]Universitas Pendidikan Indonesia, Indonesia
[1]nadhiefathallahi@upi.edu, [2]rasim@upi.edu, [3]anianisyah@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 21-07-2025
Accepted: 28-07-2025
Published: 02-08-2025

Keywords:
Dental Health Chatbot; IndoBERT; Intent Classification; Natural Language Processing; Next Sentence Prediction

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Low public awareness regarding the importance of dental health remains a significant issue in Indonesia. This situation is exacerbated by limited access to consultation services that are easy, fast, affordable, and available at any time. As a result, many dental diseases go undetected at an early stage. Additionally, the tendency to delay dental check-ups is often caused by time constraints and the distance to healthcare facilities, leading many people to avoid consulting with dentists. To address this problem, this research developed a dental health chatbot based on Natural Language Processing (NLP) using IndoBERT. The model was pretrained with the Masked Language Model (MLM) approach and fine-tuned using Next Sentence Prediction (NSP) and intent classification tasks. The dataset was compiled from Indonesian-language dental health articles, symptom–disease sentence pairs, and follow-up questions, all validated by certified dentists. The system was implemented as a web application using React JS for the frontend, Express JS and MySQL for the backend, and integrated with the NLP model through a Flask API. Evaluation results show that the chatbot can provide relevant dental health information, including lightweight consultations to assist in early symptom detection, answer follow-up questions, and generate digital medical records. Expert validation produced an average score of "Good" across the aspects of clarity, relevance, medical accuracy, and completeness, with Likert scale scores ranging from 3.53 to 3.67. This research is expected to contribute as an accessible 24-hour online dental health information service aimed at increasing public knowledge and awareness.

INTRODUCTION

The rapid development of information technology has significantly impacted various sectors, including healthcare (Prasanti and Indriani 2018). In Indonesia, this digital transformation continues to expand alongside the growing number of internet users, which reached 221.5 million in 2024, approximately 79% of the total population (Tri Haryanto 2024). This presents an opportunity to deliver faster, more efficient, and digitally accessible health services.

Unfortunately, access to dental consultation services that are easy, fast, affordable, and available anytime remains limited. Many individuals are reluctant to consult a dentist due to the significant time and effort required to reach healthcare facilities, especially among those with busy routines (Yunianto 2023). As a result, numerous dental problems go undetected in their early stages. The 2023 Indonesian Health Survey reported that although 56.9% of the population experienced dental complaints within the past year, only 11.2% sought medical consultation. Similar findings have been observed in areas such as Bantul, Cimahi, and Tegal, where people struggle to identify early symptoms of dental disease and face limited access to credible dental health education (Ramadhan et al. 2023; Himawati 2023; Sabrina Putri and Sofiani 2023).

These conditions indicate that the main challenge lies not only in public awareness but also in the lack of accessibility to early consultation services and trustworthy health information. Therefore, a solution is needed to bridge this gap using flexible digital technologies that are available anytime and anywhere.

Chatbots present a promising digital solution to address these issues. However, generative models like ChatGPT still face challenges such as hallucination delivering inaccurate yet convincing responses and data privacy concerns (Brown et al. 2020). Hence, a more controlled, domain-specific chatbot approach using Natural Language Processing (NLP) is essential.

This approach enables the model to better understand inter-sentence relationships, maintain response relevance, and reduce the risk of hallucination (Devlin et al. 2018; Kumar 2024). In addition, the chatbot is equipped with an intent classification module to detect the user's question intent more accurately, thereby generating more relevant and informative responses.

This study develops a dental consultation chatbot based on IndoBERT, pretrained using a Masked Language Model (MLM) on medical corpora and fine-tuned through two downstream tasks: Next Sentence Prediction (NSP) to match user complaints with disease descriptions, and intent classification to interpret follow-up questions. The system

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

integrates a knowledge base validated by dental experts and includes a digital medical record generation feature. The application is deployed as a lightweight web platform accessible 24/7.

The objective of this study is to design and develop a dental health chatbot based on IndoBERT using MLM, NSP, and intent classification approaches, and to evaluate its performance and medical relevance through functional testing and expert validation.

LITERATURE REVIEW

The development of healthcare chatbots has progressed rapidly alongside advancements in Large Language Models (LLMs) such as GPT, LLaMA, and Medbot. These generative models excel at producing natural and flexible conversational responses; however, they face notable challenges in ensuring the factual correctness of medical information provided to users. (Brown et al. 2020) demonstrated GPT-3's capability to generate coherent and context-aware responses across domains, including health-related dialogues. While several applications leverage GPT-3 for preliminary medical consultations, its fully generative nature is prone to hallucination—producing fluent yet factually incorrect or medically unsupported answers. (Ouyang et al. 2022) introduced Reinforcement Learning from Human Feedback (RLHF) to improve response alignment with human expectations, but such approaches still lack direct integration with validated medical knowledge, making them risky in clinical contexts.

(Bharti et al. 2020) developed *Medbot*, a conversational AI for telehealth services, employing a hybrid approach that combines intent classification with rule-based responses to improve response consistency and reduce medical risks. Similarly, open-source models like LLaMA (Touvron et al. 2023) demonstrate potential for interactive applications, yet are rarely optimized for local languages or specific domains such as dental health in Bahasa Indonesia. (Devlin et al. 2018) introduced BERT, a transformer-based language model that enables contextual understanding through Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Subsequent studies, such as (Ji et al. 2022), fine-tuned BERT for mental health screening and achieved strong performance in intent classification and sentence pairing tasks. However, the integration of MLM, NSP, and intent classification in a single framework remains underexplored, particularly for dental health applications in the Indonesian language. (Kumar 2024) emphasized the future need for combining LLMs with structured and verifiable knowledge bases, particularly in medical domains where precision and safety are paramount. Nonetheless, most existing chatbots rely either on purely generative responses or simple classification mechanisms, without the benefit of a domain-validated knowledge base to enhance safety and accuracy.

In summary, while LLMs hold great promise in conversational AI, the literature reveals a significant gap in domain-specific NLP models for Indonesian dental health chatbots. This study addresses that gap by developing a chatbot that integrates domain-adaptive MLM pretraining, NSP, and intent classification—supported by a curated knowledge base validated by dental professionals. This hybrid approach is expected to improve response accuracy, reduce hallucination risk, and provide medically accountable early consultation support.

METHOD

This study follows a system development model using the structured and systematic Waterfall methodology. This model is chosen because it is suitable for software development processes that involve sequential and interdependent stages, ranging from requirements analysis to system maintenance (Sommerville, 2011). The implementation flow of this methodology is illustrated in Figure 1, which outlines the logical sequence of phases from requirement gathering to system testing and refinement.

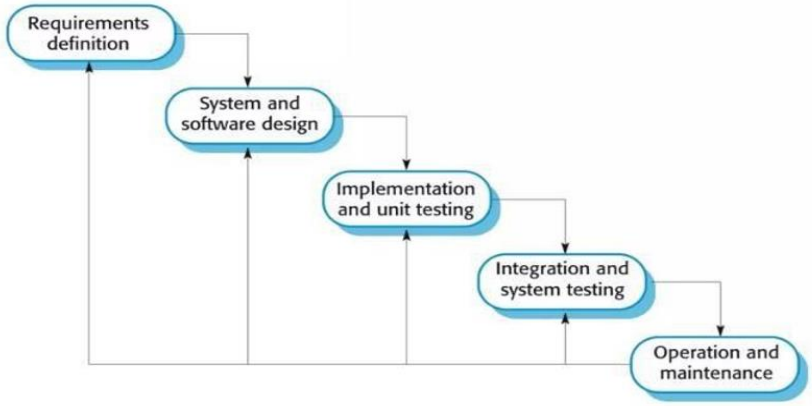


Fig 1. Waterfall Method (Sommerville 2011)

Each stage in the Waterfall model plays a crucial role in ensuring that the system is developed consistently, starting from user requirement identification to system testing and refinement. This approach ensures that the chatbot system is built on a solid foundation and validated at every phase. These stages are described in more detail in the following sections, beginning with the problem identification that forms the basis of the system requirements.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

Identification of Problems

The main problems underlying this study are the low public awareness of dental health, the common habit of delaying dental check-ups, and the limited access to accurate and reliable information. To address these issues, a dental health chatbot based on Natural Language Processing (NLP) was developed to help users recognize early symptoms, receive basic treatment suggestions, and obtain quick and relevant answers to follow-up questions.

Data Collection

The data were obtained from two main sources. Primary data were collected through interviews with dentists to construct a knowledge base containing responses related to symptoms, first aid, medications, prevention, complications, and other subtopics. Secondary data, in the form of Indonesian-language dental health articles, were used for the pre-training stage of the Masked Language Model (MLM). These articles were organized as raw text, with each complete sentence placed on a separate line to facilitate tokenization and bidirectional contextual modeling. Additionally, the dataset for Next Sentence Prediction (NSP) was compiled as sentence pairs consisting of user complaints and disease descriptions, while the intent classification dataset consisted of variations of follow-up questions labeled with intent categories such as symptoms, causes, first aid, and prevention. The data underwent normalization, tokenization, and were then split into training and testing sets.

Table 1. Example of Sentence Pairs in the Next Sentence Prediction (NSP) Dataset

sentence_a	sentence_b	label
Gusi saya bengkak dan nyeri saat mengunyah	Pasien mengalami abses periapical pada area gigi molar	1
Gigi saya terasa ngilu ketika makan manis	Pasien memiliki karies tahap awal pada email gigi	1
Saya ingin tahu cara merawat rambut rontok	Pasien mengalami abses periapical pada area gigi molar	0

Table 1 presents an example of training data for the Next Sentence Prediction (NSP) task. The dataset consists of sentence pairs representing user complaints and disease descriptions. A label of 1 indicates that the two sentences are semantically related, while a label of 0 indicates contextual mismatch. This dataset is used to train the model to recognize logical relationships between sentences in early diagnosis scenarios.

Table 2. Example of Questions in the Intent Classification Dataset

text	intent
Apa penyebab gusi saya sering berdarah?	Penyebab
Bagaimana cara meredakan nyeri pada gigi?	Tindakan awal
Obat apa yang aman untuk sakit gigi berlubang?	Obat
Apa saja komplikasi jika karies tidak ditangani?	Komplikasi

Table 2 presents example inputs for intent classification, which is the process of categorizing follow-up questions based on the type of information requested by the user. Each text is labeled with an intent such as Cause, First Aid, Medication, and Complication, which guides the model in generating appropriate responses. This dataset serves as the foundation for training the system to recognize the contextual intent of questions and deliver targeted answers.

System Requirements Analysis

The system requirements were formulated based on previously identified problems and strengthened through observations of user behavior and needs in the context of online health consultations. These insights were synthesized into a structured list of requirements categorized into two main types: functional and non-functional. The functional requirements define the core features that the system must provide to ensure a seamless and meaningful user experience. These include user registration and login, authentication, input of dental complaints, initial diagnosis using a Next Sentence Prediction (NSP) model, follow-up question handling using intent classification, storage of conversation history, generation of digital medical records, and the ability to download those records in PDF format.

On the other hand, the non-functional requirements specify performance-related attributes necessary for system reliability and user trust. These include a response time of less than five seconds, 24/7 system availability, data privacy and security, scalability to support multiple users simultaneously, and compatibility across various devices. This comprehensive set of requirements forms the foundation for system design, ensuring that both functionality and usability are addressed in the development of the dental health chatbot.

System Architecture Design and NLP Model

After the data preparation stage, the next step is to design the system workflow and the architecture of the main components involved in the development of the chatbot. To understand how the chatbot operates from the user side to the backend system and NLP model, a process flow diagram was created. This diagram helps illustrate how data is processed step by step and interactively, starting from the user's input to the final output in the form of a response or

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

medical record, as shown in Figure 2.

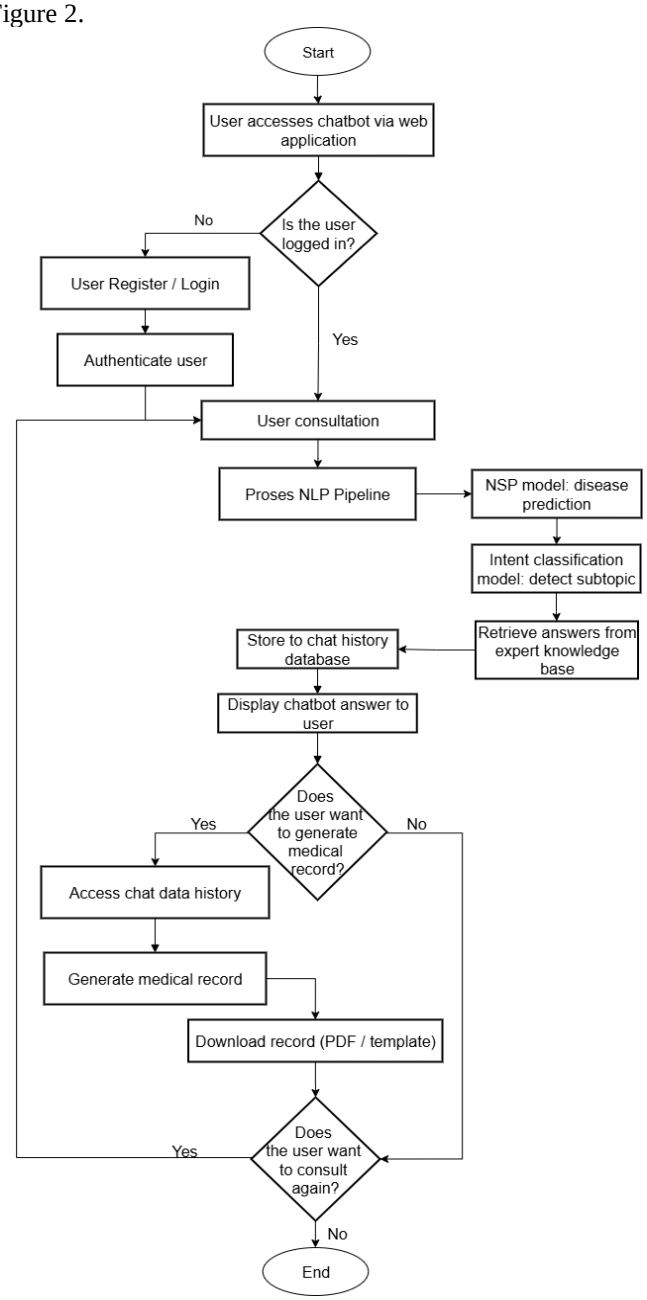


Fig 2. Process Flow Diagram of the Dental Health Chatbot System

The NLP model was based on IndoBERT-base, which was initially pre-trained using a general Indonesian-language corpus. To adapt the model to the specific domain of dental health, additional fine-tuning was conducted in three stages using domain-specific datasets.

First, domain-adaptive pretraining was performed using the Masked Language Modeling (MLM) objective. A corpus of Indonesian dental health articles was compiled and preprocessed into sentence-level plain text, with one complete sentence per line to support bidirectional context modeling. During this stage, training and validation loss were used to monitor learning performance. No classification metrics were required, as MLM is an unsupervised objective focusing on language understanding.

Second, the model was fine-tuned on the Next Sentence Prediction (NSP) task. The dataset consisted of sentence pairs representing user complaints (sentence_a) and disease descriptions (sentence_b). Positive samples represented semantically coherent complaint disease pairs, while negative samples involved mismatched topics. Evaluation was conducted using classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix to assess the model's ability to distinguish between relevant and irrelevant sentence pairs.

Third, an Intent Classification module was trained to identify the type of follow-up question asked by the user. This task utilized labeled data covering intents such as symptoms, causes, treatment, medication, complications, and

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

prevention. Model performance was evaluated using standard classification reports and confusion matrices to ensure accurate intent prediction.

By combining these three fine-tuning stages, the IndoBERT model was adapted not only to understand medical language related to dentistry, but also to classify sentence relations and detect user intent accurately. This NLP pipeline formed the core of the chatbot's reasoning and response-generation capabilities.

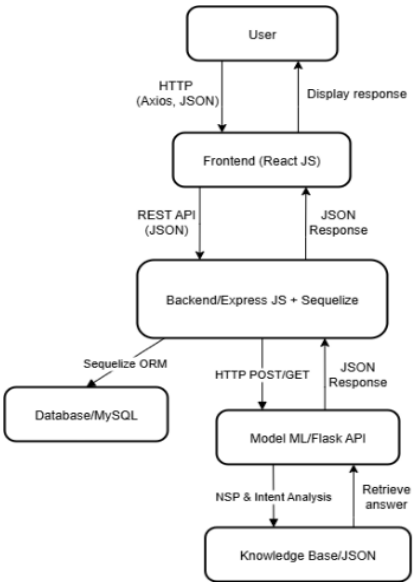


Fig 3. System Architecture of the Dental Health Chatbot

As illustrated in Figure 3, this architecture shows the distribution of responsibilities among system components, enabling a more structured approach to development, testing, and maintenance. The integration with the Flask API and the knowledge base ensures that the system can deliver real-time NLP-based responses. One of the key elements in the chatbot system is data storage, particularly for conversation history and consultation results. To address this, a relational database structure was designed to efficiently represent the relationships between entities.

To ensure data consistency and support core features of the chatbot system, such as user login, complaint recording, and medical record generation, a relational database schema was designed. The structure of these table relationships is depicted in Figure 4, which illustrates how entities are connected and how data flows between them.

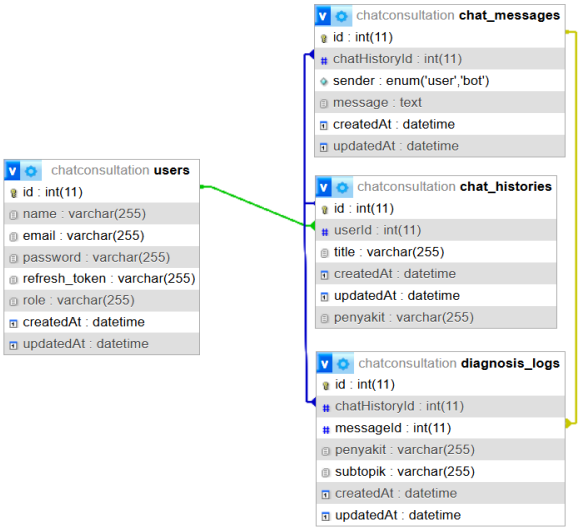


Fig 4. Dental Health Chatbot Database Table Relationship Diagram

The table relationship design supports the functional requirements of the system, such as user login, complaint recording, and medical record generation. These relationships ensure data integrity is maintained and that information can be easily traced during retrieval or further analysis.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

Implementation

The frontend application was developed using React JS, Vite, and Material UI to support interactive conversations, responsiveness, and consultation history display. The backend was built with Node.js and Express JS, connected to a MySQL database through the Sequelize ORM. The IndoBERT model was implemented via a Flask API, which is connected to the backend through HTTP REST API requests. All components were integrated to enable users to log in, input complaints, engage in conversations, store history, and download digital medical records within a unified workflow.

Testing

Functional Testing (Black-Box Approach)

Functional testing was carried out using a black-box testing approach to ensure that each system feature operates in accordance with user expectations, without examining the internal code structure. This testing involved a series of key scenarios that simulate user interactions with the system interface to validate end-to-end behavior. The registration feature was tested by simulating a user filling out the registration form with their full name, email, password, and password confirmation. Upon successful registration, the system displayed a "Register Success" notification, and the form fields were automatically cleared to indicate completion. Login functionality was tested by entering valid credentials corresponding to a previously registered user. When the email and password matched, the system responded with a "Login Successful" message and redirected the user to the main chatbot interface.

The chatbot interaction feature was validated by inputting a dental complaint, such as "I have a small black spot on my tooth that hasn't gone away for two days, what could be the cause?" The system successfully processed the input and returned a diagnostic prediction along with a relevant response based on the identified subtopic. For the consultation history feature, after completing a session, users were able to view stored conversation records. Selecting one of the records successfully displayed the full transcript of that session. Lastly, the medical record generation feature was tested by selecting the "View Medical Record" option after a consultation. The system generated a downloadable PDF containing the relevant consultation details, structured based on the recorded dialogue. This black-box testing confirmed that the key functionalities of the system—user management, complaint handling, session recording, and document generation—operate reliably and align with the intended design.

Non-Functional Testing (Performance Evaluation)

To evaluate the performance of the system, non-functional testing was conducted with a focus on response speed, stability, and scalability when accessed by multiple users simultaneously. This testing aimed to ensure that the chatbot remained responsive and reliable under high-demand conditions through both load testing and stress testing. The testing was carried out using Apache JMeter, a widely used tool for simulating concurrent user access and measuring system behavior under pressure. The test environment was configured on a high-performance server equipped with a 16-core CPU, an H100 SXM NVLink 80GB GPU, 200GB of RAM, and an SSD, running Ubuntu 22.04 LTS.

The simulation involved 100 virtual users accessing the chatbot system concurrently, with a ramp-up period of 100 seconds to gradually increase the load. The deployed NLP components included the Next Sentence Prediction (NSP) model with a size of 1.4 GB and the Intent Classification model sized at 500 MB. The results of the performance testing confirmed that the system maintained a stable response time and did not experience significant degradation under load. This validated the system's readiness for real-world deployment, particularly in scenarios where multiple users may engage in consultations simultaneously.

Expert Validation

Expert validation was conducted to assess the quality of responses generated by the dental health chatbot in terms of accuracy, relevance, and medical appropriateness. This validation employed a questionnaire instrument using a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), as suggested by previous literature (Joshi et al. 2015; Budiaji 2013). The evaluation involved five independent dental professionals who served as expert respondents. Each expert assessed the chatbot responses based on four key aspects: Clarity of Sentence Structure, Relevance of Response, Medical Appropriateness, and Completeness of Information.

A total of 13 evaluation items were distributed across the four aspects. The Relevance of Response aspect included four indicators, such as the chatbot's ability to provide answers aligned with user intent, distinguish between complaints and follow-up questions, and avoid giving dental diagnoses when the input was outside the domain. The Clarity aspect assessed the readability and linguistic naturalness of the responses, as well as the avoidance of unexplained medical jargon. The Medical Appropriateness aspect evaluated the consistency of the responses with standard dental knowledge, avoidance of misleading or dangerous content, and the absence of high-risk suggestions. Lastly, the Completeness aspect examined whether the chatbot provided sufficient elaboration, offered initial treatment suggestions, and guided users toward understanding their condition and practical solutions.

This validation process ensured that all critical dimensions of the chatbot's output were reviewed in detail. The results served not only as a benchmark for current system performance but also as a foundation for future content

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

refinement and continuous improvement of the chatbot's medical communication quality.

Maintenance

The final stage of the system development process is maintenance, which ensures the long-term functionality, adaptability, and reliability of the chatbot after deployment. This phase includes activities such as monitoring user interactions, collecting feedback from both users and experts, correcting identified errors, and updating the system to align with evolving user needs or advancements in dental health knowledge. To support sustainable system operation, a feedback mechanism is planned to capture user-reported issues or suggestions through the application interface. Additionally, periodic reviews will be conducted to update the knowledge base and retrain the NLP models when necessary, especially in response to significant changes in medical guidelines or language patterns. This ongoing maintenance approach aims to keep the chatbot responsive, medically accurate, and contextually relevant over time.

RESULT

This section presents the results of the research and discusses their implications in relation to the research objectives. The evaluation focuses on three main aspects: (1) the performance of the NLP model in each training stage, including domain-adaptive pretraining, Next Sentence Prediction (NSP), and intent classification; (2) the functional and non-functional testing of the web-based chatbot system; and (3) expert validation of the chatbot's responses. Each sub-section provides an in-depth explanation of the methods used, the evaluation metrics applied, and the interpretation of the outcomes obtained during testing and validation.

NLP Model Evaluation

Pretrained Masked Language Model

The initial stage of system development was conducted by retraining the IndoBERT model using the Masked Language Modeling (MLM) approach. The pretraining dataset was compiled from Indonesian-language dental health articles, which were processed so that each sentence was placed on a separate line. After preprocessing, the number of lines increased from 1,981 to 2,349, indicating that sentence segmentation successfully enhanced the granularity of the data. The model was trained for 30 epochs using an L4 GPU on Google Colab. The training process yielded a final training loss value of 1.079, suggesting that the model successfully learned the linguistic structure and contextual patterns of the medical domain. The training took approximately 17 minutes and 47 seconds. The resulting model was then utilized for the subsequent fine-tuning stages.

Fine-tuning for Next Sentence Prediction (NSP)

In the NSP fine-tuning stage, the model was trained to recognize the relationship between user complaints and disease descriptions. The dataset consisted of 680 sentence pairs, which were split into 80% for training and 20% for validation. Evaluation on the validation set yielded an accuracy of 93%, with an F1-score of 0.93 for the "Not Next" label and 0.92 for the "Is Next" label. The detailed classification results are presented in the following Table 3:

Table 3. NSP Classification Report

label	precision	recall	F1-Score	Support
Not Next	0.93	0.93	0.93	73
Is Next	0.92	0.92	0.92	63
Accuracy			**0.93**	**136**

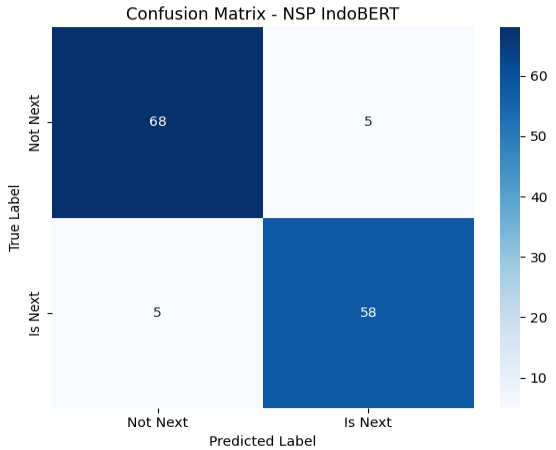


Fig 5. Confusion Matrix NSP

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

Figure 5 displays the confusion matrix for the NSP IndoBERT model on the validation set. The matrix shows that the model correctly classified 68 out of 73 "Not Next" sentence pairs and 58 out of 63 "Is Next" sentence pairs. Each misclassification amounted to only 5 instances for both labels. These results indicate that the model performs well in identifying semantically related sentence pairs. The balanced distribution of predictions and minimal misclassifications demonstrate that the model generalizes well in distinguishing whether a given user complaint logically matches a disease description. This capability is essential for ensuring that the chatbot can provide accurate early diagnosis based on user input.

Fine-tuning for Intent Classification

The intent classification model was designed to recognize the user's follow-up question intentions, such as "what is the cause?", "what is the medication?", or "do I need to see a dentist?". The dataset consisted of 452 labeled questions, which were split into 80% for training and 20% for validation. Evaluation of the model on the validation set showed an overall accuracy of 93%, with high F1-scores across most intent categories. The detailed results are presented in the following Table 4:

Table 4. Intent Class Classification Report

label	precision	recall	F1-Score	Support
recovery_duration	0.91	1.00	0.95	10
symtoms	0.80	0.80	0.80	10
complications	0.77	1.00	0.87	10
medication	1.00	1.00	1.00	10
prevention	1.00	1.00	1.00	10
definition_of_disease	0.91	1.00	0.95	10
causes	0.90	0.90	0.90	10
need_to_see_dentist	0.86	0.60	0.71	10
things_to_avoid	1.00	0.80	0.89	10
Accuracy			**0.93**	**136**

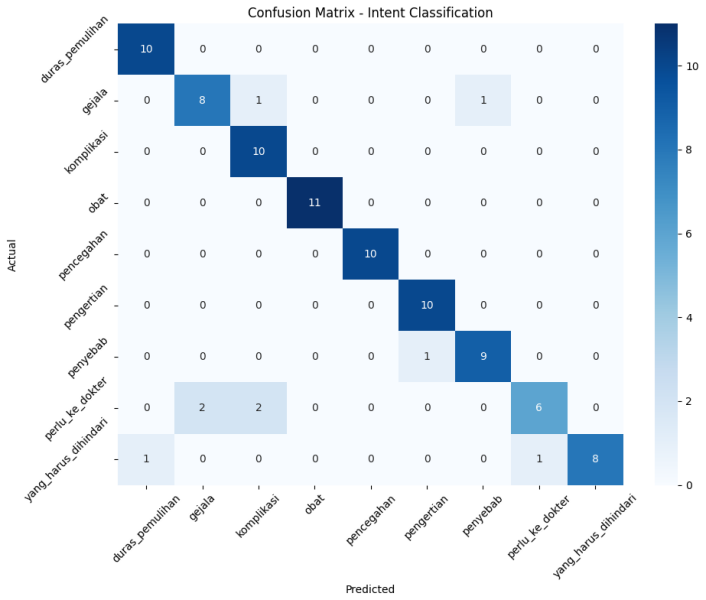


Fig 6. Confusion Matrix Intent Class

The results presented in Table 4 and Figure 6 indicate strong performance of the model in classifying users' follow-up question intents. The model achieved an accuracy of 93% with high F1-scores across most intent categories. Several labels such as *obat* (medication), *pencegahan* (prevention), and *pengertian* (definition) attained perfect F1-scores of 1.00, likely due to their clear lexical patterns and contextual distinctiveness. The confusion matrix further supports this result, with most predictions aligning along the diagonal—indicating accurate classification. However, misclassifications were observed, particularly in the *perlu_ke_dokter* (need_to_see_dentist) category, which was frequently confused with *gejala* (symptoms) and *komplikasi* (complications), suggesting semantic overlap among these intents. Similar issues occurred with *gejala* and *yang_harus_dihindari* (things_to_avoid), although less frequently. These findings suggest the need for further refinement in distinguishing semantically similar categories to ensure more

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

accurate responses in real-world chatbot interactions.

In summary, the model exhibits excellent classification performance across most intents, though refinement may be needed for closely related categories to enhance interpretability and ensure precise responses in real-world chatbot interactions.

Functional Testing (Black-Box)

Functional testing was conducted using a black-box testing approach. All core features including user authentication, complaint input, and the generation of digital medical records were tested according to real usage scenarios. The results showed that the system operated stably without errors and was able to respond to each input with the appropriate output. All interaction flows with the chatbot, including the initial diagnosis process and intent-based follow-up responses, functioned properly. This indicates that the system's logic and functional implementation performed according to the specified design requirements.

Non-Function Testing (Performance)

Non-functional testing was conducted to evaluate the system's performance when accessed concurrently by multiple users. The testing was performed using JMeter with a configuration of 100 virtual users and a ramp-up period of 100 seconds. The simulation followed a user activity flow that reflected actual usage scenarios, including logging into the system, accessing the chatbot homepage, and submitting dental complaints through the chat interface. Each virtual user repeatedly executed this sequence of actions to assess the system's stability and resilience under high load. The test results showed that the system delivered an average response time of 2.83 seconds without encountering any errors, timeouts, or performance issues. All HTTP requests were successfully processed with a 200 OK response status, and the system's throughput remained stable throughout the testing period. Therefore, the system can be categorized as responsive and reliable, and is considered suitable for use at a medium scale in the context of web-based health consultation services.

Expert Validation

Expert validation was conducted to evaluate the quality of responses generated by the dental health chatbot, focusing on four key aspects: sentence clarity, medical appropriateness, answer relevance, and informational completeness. Each aspect was rated using a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) by five independent dental practitioners. The average scores for each aspect are summarized in Table 5 below:

Table 5. Intent Class Classification Report

Evaluation Aspect	Average Score	Interpretation
Sentence clarity	3,67	Good
Medical appropiateness	3,60	Good
Answer relevance	3,55	Good
Informational completeness	3,53	Good

The results indicate that all evaluated aspects fall within the "Good" category. This suggests that the chatbot is capable of delivering responses that are generally clear, contextually relevant to user complaints, medically appropriate, and sufficiently informative. Although the scores have not yet reached the "Very Good" category, these findings demonstrate that the chatbot is already suitable for use as an initial consultation tool for dental health issues particularly in areas where access to medical professionals is limited.

DISCUSSION

The results of this study demonstrate that the dental health chatbot system based on IndoBERT is capable of generating responses that are generally relevant and contextually appropriate to both initial complaints and follow-up questions. The multi-stage NLP approach—comprising Masked Language Model (MLM) pretraining, Next Sentence Prediction (NSP) fine-tuning, and intent classification—proved effective in modeling semantic relationships and understanding user intent in simple consultation scenarios, as evidenced by the high accuracy scores obtained during model evaluation.

Nevertheless, several technical limitations were identified during the implementation and testing phases. A key issue lies in the quality and diversity of the training dataset. Due to the relatively limited dataset size, the model's ability to generalize remains suboptimal, particularly when handling long, complex, or indirect complaint sentences. While the NSP model achieved high validation accuracy, its performance declined when processing complaints containing multiple symptoms or non-standard sentence structures. This presents a notable challenge in the dental health domain, where similar symptoms may correspond to multiple potential diagnoses, and sentence-level matching alone may not be sufficient to ensure diagnostic precision.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

In the intent classification module, some labels—such as *perlu_ke_dokter* (need to see a doctor) and *gejala* (symptoms)—yielded lower F1-scores compared to other intent categories. This issue can be attributed to semantic overlap between certain labels and insufficient data representation for specific intents. As a result, the system occasionally produced responses that did not align with the user's actual follow-up question, especially in cases of ambiguous phrasing.

On the other hand, the IndoBERT-based approach still demonstrated strong potential for domain adaptation and contextual understanding in Bahasa Indonesia, particularly in dental health contexts. The use of a structured knowledge base as the chatbot's response source also contributed significantly to reducing hallucination risks commonly observed in generative models. Expert validation supported these findings, with average scores categorized as "Good" in terms of sentence clarity, relevance, medical appropriateness, and completeness.

Despite these positive results, the expert reviewers emphasized ethical considerations and limitations regarding AI usage in healthcare. The chatbot must not provide definitive diagnoses or medication recommendations without a physical examination by qualified professionals. Therefore, the system was designed to include disclaimers and avoid generating responses containing sensitive or high-risk medical information.

Future work should focus on expanding and diversifying the training dataset, particularly for long-form complaints, ambiguous questions, and overlapping symptoms. Additionally, the integration of reasoning capabilities or multi-task learning architectures could enhance the chatbot's ability to interpret more complex clinical contexts. This system also has the potential to be extended to other medical domains and linked with national health record systems using NIK (Indonesian national ID) for more personalized and centralized health services.

CONCLUSION

This study successfully developed a dental health chatbot based on IndoBERT using a multi-stage NLP approach, including Masked Language Model (MLM) pretraining, Next Sentence Prediction (NSP) fine-tuning, and intent classification. The system was implemented as an interactive web-based application capable of providing initial symptom analysis, answering follow-up questions, and generating digital medical records.

The model evaluation showed strong performance, achieving 93% accuracy on the NSP task and consistently high F1-scores across various intent categories. Functional testing confirmed that all system features operated according to expected user scenarios, while performance testing using JMeter demonstrated that the system maintained an average response time of under 3 seconds even when accessed by up to 200 simultaneous users. Expert validation resulted in an overall "Good" rating across four key aspects: clarity, relevance, completeness, and medical appropriateness. Expert feedback also emphasized the importance of addressing medical ethics, which was incorporated into the system's maintenance phase through the refinement of responses, inclusion of disclaimers, and continuous updates to the knowledge base. As a result, the chatbot presents a practical, informative, and medically responsible alternative for early-stage dental consultation. Furthermore, the system has potential for future adaptation in other healthcare domains.

This research is not without limitations. The current dataset lacks broad variation, and the scope of diseases is restricted to common cases. Future studies should aim to expand the medical domain, incorporate multilingual NLP models including local dialects, integrate image-based diagnostic support through computer vision, and adopt reinforcement learning to enhance response quality through continuous user feedback. In addition, experimental studies can be conducted to assess the chatbot's impact on public health literacy, and voice-based interaction features can be developed to improve accessibility across diverse user groups.

REFERENCES

Bharti, Urmil, Deepali Bajaj, Hunar Batra, Shreya Lalit, Shweta Lalit, and Aayushi Gangwani. 2020. "Medbot: Conversational Artificial Intelligence Powered Chatbot for Delivering Tele-Health after Covid-19." *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020* (Icces):870–75. doi: 10.1109/ICCES48766.2020.09137944.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 2020-Decem.

Budiaji, Weksi. 2013. "The Measurement Scale and The Number of Responses in Likert Scale." *Jurnal Ilmu Pertanian Dan Perikanan Desember* 2(2):125–31.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A. I. Language. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Naacl-Hlt 2019* (Mlm):4171–86.

Himawati, Marlin. 2023. "Upaya Peningkatan Kesadaran Menjaga Kesehatan Gigi Dan Mulut Di Wilayah Kerja Puskesmas Leuwigajah Dengan Program Penyuluhan." *Jurnal Abdimas Kartika Wijayakusuma* 4(2):130–36. doi:

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6620

10.26874/jakw.v4i2.304.

Ji, Shaoxiong, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare." *2022 Language Resources and Evaluation Conference, LREC 2022* (June):7184–90.

Joshi, Ankur, Saket Kale, Satish Chandel, and D. Pal. 2015. "Likert Scale: Explored and Explained." *British Journal of Applied Science & Technology* 7(4):396–403. doi: 10.9734/bjast/2015/14975.

Kumar, Pranjal. 2024. *Large Language Models (LLMs): Survey, Technical Frameworks, and Future Challenges*. Vol. 57. Springer Netherlands.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35.

Prasanti, Ditha, and Sri Seti Indriani. 2018. "Pengembangan Teknologi Informasi Dan Komunikasi Dalam Sistem E-Health Alodokter.Com the Use of Information and Communication Technology in E-Health System Alodokter.Com." *Jurnal Sosioteknologi* 17(1):93–103.

Ramadhan, Akhmad, Rahmad Arifin, Isnur Hatta, Riky Hamdani, and Nurdiana Dewi. 2023. "HUBUNGAN PENGETAHUAN KESEHATAN GIGI DAN MULUT DENGAN KEHILANGAN GIGI DI WILAYAH KERJA PUSKESMAS SEMANGAT DALAM." *Dentin* 7. doi: 10.20527/dentin.v7i3.10746.

Sabrina Putri, Saskia, and Erma Sofiani. 2023. "Peningkatan Pengetahuan Kesehatan Gigi Dan Mulut Melalui Improving Dental and Oral Health Through Counseling on Students of Madrasah Ibtida ' Iyah." 8(6):947–53.

Sommerville, I. 2011. *Software Engineering (9th Ed.; Boston, Ed.). Massachusetts: Pearson Education.*

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. "LLaMA: Open and Efficient Foundation Language Models."

Tri Haryanto, Agus. 2024. "APJII Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang." *APJII*. Retrieved (https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang).

Yunianto, Tri Kurnia. 2023. "Tertinggi Di Asia, Kenaikan Biaya Medis RI 13,6% Pada Tahun 2023." *Marketeers*, 1.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.