Journal of Information Systems and Informatics
Vol. 7, No. 2, June 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i2.1079 Published By DRPM-UBD

Implementation of a Telegram-Based Child Consultation Chatbot Using IndoBERT

Gusti Ayu Wahyu Whurapsari[1], I Made Agus Dwi Suarjaya[2], Wayan Oger Vihikan[3]

[1,2,3]Departement of Information Technology, Udayana University, Bali, Indonesia
Email: [1]wahyuwhurapsari@student.unud.ac.id, [2]agussuarjaya@it.unud.ac.id, [3]oger_vihikan@unud.ac.id

Abstract

Children's health and development are crucial aspects that require proper attention from parents. However, many parents lack easy access to immediate consultation regarding their child's health and well-being. To address this issue, this study develops a child consultation chatbot on Telegram using the IndoBERT model. The chatbot utilizes data from Halodoc and Alodokter, structured into an intent-based format with 227 tags, 5,428 patterns, and 278 responses. The dataset undergoes preprocessing, including lowercasing, text cleaning, normalization, stopword removal, and stemming. Four preprocessing scenarios are tested, including the use of term frequency-based stopwords without applying stemming, the use of NLTK stopwords without stemming, the use of term frequency-based stopwords combined with stemming, and the use of NLTK stopwords combined with stemming. The best model, trained with an 80:20 training-validation split using term frequency-based stopwords without stemming, achieves 98% accuracy, 98.5% F1-score, 98.9% precision, and 98.5% recall. The chatbot successfully classifies user intent and ensures structured interactions through a confidence-based response mechanism. This research demonstrates that an IndoBERT-based chatbot can effectively assist parents in obtaining quick and relevant information regarding their children's health and development.

Keywords: Child Consultation, Chatbot, IndoBERT, Intent Classification

1. INTRODUCTION

Children are the future assets of a nation, and their welfare must be prioritized to build quality human resources. Parents play a crucial role in supporting their children's growth and development, which includes instilling healthy habits from an early age and applying appropriate parenting practices. However, in Indonesia, access to quality pediatric healthcare remains uneven. According to the Central Statistics Agency (BPS), 28.81% of children in Indonesia experienced health complaints in 2022, marking a 4.13% increase from 24.68% in the previous year [1]. The 2022 Next Generation Indonesia report by the British Council also highlighted that access to medical care is still a pressing issue among the youth [2].

Furthermore, the Indonesian Pediatric Society underscores the persistent inequality in the distribution of pediatric health services across the country [3]. With such a high percentage of health complaints among children, timely and appropriate interventions are crucial to prevent further escalation of health problems.

The role of parents is vital in the development of children. Parents who do not pay attention to their child's developmental conditions will result in children growing in their way and not developing according to the expected pattern [4]. Sometimes, this happens without parents realizing or knowing, and often happens unintentionally. Parents may act like that because they do not understand how to educate children properly, or even though they know, certain conditions that force them to behave like that.

In the era of rapidly developing technology, a platform is needed that makes it easy for parents to consult quickly and easily. The consultations in question to children's health and parenting patterns. Many online consultation platforms can be used, but most of them require waiting for a response from a doctor. One of the artificial intelligence technologies, namely chatbots, can be used to solve these problems. A chatbot is a type of machine designed to interact with humans using natural language, aiming to simplify social activities across various fields [5]. Chatbots can be accessed at any time, without time or time zone restrictions, and can respond to questions in seconds.

Conventional approaches to creating chatbots usually use rule-based systems, but this method is often less effective when dealing with new contexts or changes that are not predicted by existing rules. As a result, such systems require regular updates to adapt to changing scenarios [6]. In addition, rule-based systems have limited flexibility [7], require high development and maintenance costs because they must be manually updated periodically, and are less reliable in handling variations in user input. These systems often have difficulty recognizing user intent, especially in cases such as spelling errors, use of slang, or abbreviations [8].

To overcome this challenge, deep learning techniques emerged, especially transformer-based models such as BERT. One important task in chatbot development is intent classification, which is a task in natural language processing (NLP) that classifies user queries into specific intents. A study entitled "Comparative Analysis of Chatbot Intent Classification Using Deep Learning BERT, RoBERTa, and IndoBERT" conducted a comparison of chatbot intent classification using the deep learning models BERT, RoBERTa, and IndoBERT. The results showed that IndoBERT outperformed BERT and RoBERTa, achieving an accuracy of 94%, compared to 89% for BERT and 84% for RoBERTa [9]. IndoBERT is a transformer-based language model specifically

trained on Indonesian language corpora [10]. By leveraging linguistic patterns and vocabulary specific to Indonesian, IndoBERT is better suited for understanding nuances in user queries written in Indonesian, which contributes to its superior performance in tasks such as intent classification. This shows that IndoBERT can provide more accurate results in understanding user intent and producing relevant responses

While several studies have examined the use of deep learning models in chatbot intent classification, few have focused specifically on consultation systems addressing both child health and parenting topics in the Indonesian language. There have been limited studies concerned with integrating transformer-based chatbots into practical applications for Indonesian-language parenting and child consultation. Additionally, little attention has been paid to the impact of different preprocessing techniques such as stopword removal and stemming on the performance of intent classification models in this domain. Therefore, this research intends to develop a chatbot system using IndoBERT to classify and respond to queries related to both child health and parenting in Bahasa Indonesia. The objectives of this research are to build an intent classification model using IndoBERT tailored to Indonesian-language queries on child health and parenting, compare the effects of various preprocessing techniques (including stopword removal and stemming) on model performance, and evaluate the chatbot's effectiveness in understanding and providing relevant responses for quick, accessible consultations.

2. METHODS

This study consists of eight stages, as illustrated in Figure 1. The process begins with data collection, creating data intent, splitting the dataset, preprocessing data, fine-tuning the IndoBERT model, evaluating the model, creating a telegram bot, and ending with user acceptance test.

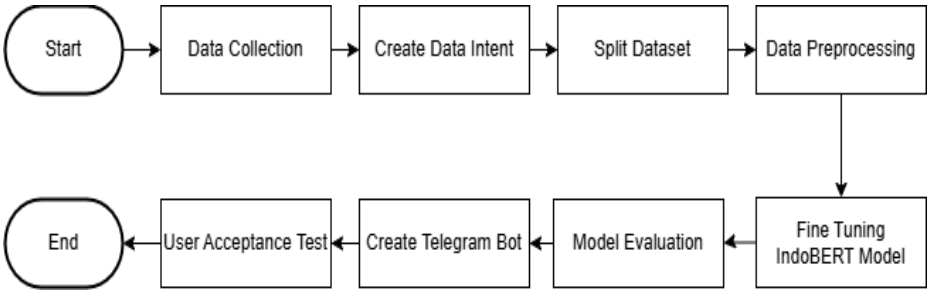

Figure 1. Research process flow

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.1. Data Collection

The dataset for this study was sourced from the Alodokter and Halodoc websites. Alodokter data consists of question-and-answer data taken from 2015 to 2024. Alodokter data topics are taken with topic tags such as children, child psychology consultation, child nutrition, and fever in children. Meanwhile, Halodoc data is in the form of article data with the topic of childcare patterns. Data collection from Alodokter and Halodoc websites using the Python programming language with the Beautiful Soup and Selenium libraries. The results of scraping Alodokter data are in the form of questions and answers, while for Halodoc data it is in the form of article data. The data scraped from Alodokter was structured into several columns, including the short question, the date the question was asked, the date the doctor responded, the doctor's name, the complete question, and the doctor's answer. A data cleaning process was carried out, as some entries had inconsistent structures that made them unsuitable for direct column formatting. The cleaning involved steps such as removing excessive line breaks and formatting inconsistencies. Because Halodoc data is still in the form of articles, the researcher needs to sort the articles by taking only important points to be used as answers, and the researcher creates question data that matches the answers in the Halodoc article. The data from Alodokter scraping is 9,938 data while the Halodoc article data is 135 data. Not all of the scraping data is used, it needs to go through a data selection process that is entered into the intent.

2.2. Data Intent

Intent is a collection of data consisting of input-output pairs. This intent represents the intent or action that the user wants to achieve when interacting with the chatbot [9]. Intent is a collection of data consisting of input-output pairs. This intent represents the intent or action that the user wants to achieve when interacting with the chatbot. The intent structure consists of tags, patterns, and responses. Tag means the topic's name raised in the question (the tag name must be unique where there is no tag with the same name). Tag is also referred to as a label, which represents the category or class associated with the question (pattern) in the dataset. Patterns contain a list of similar questions with the same topic. The response includes the appropriate answer to the question. To group the data into intents, the researcher manually reviewed the complete questions from Alodokter and categorized those with similar purposes under the same tag. For example, under the tag "*tantrum_pada_anak*", the researcher grouped questions related to tantrums that aimed to find solutions for managing them. Most of the questions selected focused on problem-solving by asking for specific solutions. Meanwhile, for Halodoc, the article content was used as answer data by manually extracting only the key parts that addressed the manually listed questions (patterns) created by the researcher. Not all data can be used because of its wide variety, making it

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

difficult to create intent if only using a small amount of data. The number of tags used in this study was 227 tags with a total of 5.428 patterns and 254 responses. Of the total 227 tags, there are 16 tags consisting of 278 patterns and 27 responses that include greetings, thanks, and other interaction elements that can make conversations with chatbots more lively. Each tag has only one response, except for the greetings, thanks, and interaction tags, which have multiple responses. An example of intent data is presented in Table 1.

Tabel 1. Example of intent data

Tag	Pattern	Response
tantrum_pada_anak (tantrums_in _children)	*sy punya keponakan umur 2 tahun 3bln, jika nangis suka menyakiti diri sendiri dengan menjambak rambut hingga botak.krn khawatir maka rambut di botakin.stlh rambut botak kl nangis skr suka mencakar perut,muka dan kepala, yg sy tanyakan knapa hal ini bs terjadi.dan apa penyebabnya* (I have a nephew who is 2 years and 3 months old. When he cries, he likes to hurt himself by pulling his hair until he is bald. Because he is worried, he shaves his hair off. After his hair is bald, when he cries, he likes to scratch his stomach, face and head. What I want to ask is why this can happen and what causes it.)	*Tantrum merupakan suatu kondisi ketika anak meluapkan emosinya dengan menangis kencang, berguling-guling bahkan melempar barang. Jadi bukan sebuah penyakit ya, kondisi tantrum ini umum dialami oleh anak-anak karena luapan emosinya akibat sesuatu yang tidak ia dapat …. Semoga dapat membantu.* (Tantrum is a condition when a child expresses his emotions by crying loudly, rolling around and even throwing things. So it is not a disease, this tantrum condition is commonly experienced by children because of their emotional outburst due to something they cannot get.... Hopefully it can help.
demam_pada_anak (fever_in _children)	*dok, saya mau tanya.. anak saya umur nya 9 bulan.. badan nya agak panas (anget).. bagaimana untuk mengatasi nya ??* (Doc, I want to ask... my child is 9 months old... his body is a bit hot (hot).. how to deal with it??)	*Demam (febris) merupakan respon tubuh yang normal yang terjadi ketika tubuh mengalami infeksi. Untuk dirumah ada berapa hal yang bisa anda lakukan seperti : berikan anak cairan yang banyak seperti air putih …. Jika dirasakan demam tidak kunjung membaik sebaiknya kunjugi dokter anak.* (Fever (febrile) is a normal body response that occurs when the body experiences

	an infection. At home, there are several things you can do, such as: give your child lots of fluids such as water .... If you feel the fever does not improve, you should visit a pediatrician)

2.3. Data Preprocessing

Data pre-processing is a crucial step that aims to clean, transform, and prepare the data, making it easier to analyze and leading to more accurate outcomes. [11]. Data that goes through preprocessing is pattern data or question data. Meanwhile, the response data is not subjected to preprocessing, as it is already in a clean and structured format. The pre-processing stages carried out in this study are explained as follows and the flow can be seen in Figure 2.

1) Lowercasing: changing text to lower case with the aim of standardizing the text so as to minimize any capitalization differences that may exist in the text data [12].
2) Cleaning Text: includes removing mentions, URLs, hashtags, extra spaces, newline characters, punctuation, emojis, and excess whitespace, as these elements do not represent the primary context in a text [12].
3) Normalization: process of text standardization involves correcting typographical errors, standardizing abbreviations, and converting non-standard language into its proper form using a predefined normalization dictionary [13].
4) Stopword Removal: to remove stopwords or irrelevant words because they do not represent the context of a text [14]. There are 2 stopwords used, namely term frequency-based stopwords and NLTK (Natural Language Toolkit) library stopwords.
5) Stemming: process of reducing affixed words to their root form [15]. The stemming method used in this study utilizes the Sastrawi library, which is specifically designed for stemming words in the Indonesian language.

The preprocessing stage includes lowercasing, text cleaning, normalization, stopword removal, and stemming. Stopword removal is applied using two approaches: term frequency-based and the NLTK stopword list, while stemming is tested with and without application to analyze its impact on model performance. To evaluate the effectiveness of these techniques, four preprocessing scenarios are compared: (1) term frequency-based stopwords without stemming, (2) NLTK stopwords without stemming, (3) term frequency-based stopwords with stemming, and (4) NLTK stopwords with stemming. This comparison provides deeper

insights into how preprocessing influences model quality. The preprocessing results can be seen in Table 2.

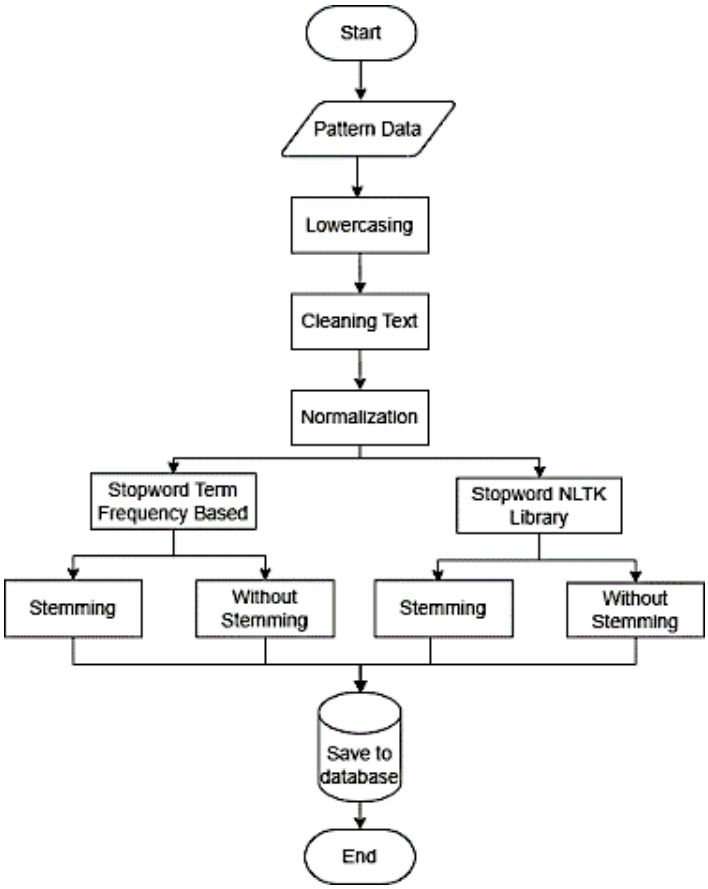

Figure 2. Data preprocessing flow

One approach to constructing a stopword list involves ranking all tokens by their frequency of occurrence, indicating how often each word appears in a document collection. Words with exceptionally high frequencies are removed based on a predefined threshold. Since overly frequent words are often insignificant or irrelevant for document classification, they can be identified as stopwords and excluded [16].

In this study, the term frequency-based stopword removal method eliminated 22 words out of a total of 5,523 unique words in the corpus, accounting for approximately 0.40% of the total. For stopword removal using the NLTK library, the researcher used the default predefined stopword list without any additional

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

technical configuration. However, in both approaches, the words "*tidak*", "*tidakkah*", and "*tidaklah*" were deliberately retained, as they play a crucial role in expressing negation in the Indonesian language. Removing these words could significantly alter the meaning of the sentences.

Tabel 2. Pre-processing steps

Process	Text (Indonesia)	Result (Indonesia)
Lowercasing	Dok,, dr kmrin anak saya muntah2 doank,udh di bawa' ke bidan, tp hari ini malah muntah2 ny disertai mencret, tolong solusi ny dok,,	dok,, dr kmrin anak saya muntah2 doank,udh di bawa' ke bidan, tp hari ini malah muntah2 ny disertai mencret, tolong solusi ny dok,,
Cleaning Text	dok,, dr kmrin anak saya muntah2 doank,udh di bawa' ke bidan, tp hari ini malah muntah2 ny disertai mencret, tolong solusi ny dok,,	dok dr kmrin anak saya muntah2 doank udh di bawa ke bidan tp hari ini malah muntah2 ny disertai mencret tolong solusi ny dok
Normalization	dok dr kmrin anak saya muntah2 doank udh di bawa ke bidan tp hari ini malah muntah2 ny disertai mencret tolong solusi ny dok	dok dari kemarin anak saya muntah muntah saja sudah di bawa ke bidan tetapi hari ini malah muntah muntah nya disertai mencret tolong solusi nya dok
Stopword Removal (Term Frequency-Based)	dok dari kemarin anak saya muntah muntah saja sudah di bawa ke bidan tetapi hari ini malah muntah muntah nya disertai mencret tolong solusi nya dok	dari kemarin muntah muntah saja bawa ke bidan malah muntah muntah nya disertai mencret tolong solusi nya
Stopword Removal (NLTK Library)	dok dari kemarin anak saya muntah muntah saja sudah di bawa ke bidan tetapi hari ini malah muntah muntah nya disertai mencret tolong solusi nya dok	dok kemarin anak muntah muntah bawa bidan muntah muntah nya disertai mencret tolong solusi nya dok
Stemming After Stopword Removal Term Frequency-Based	dari kemarin muntah muntah saja bawa ke bidan malah muntah muntah nya disertai mencret tolong solusi nya	dari kemarin muntah muntah saja bawa ke bidan malah muntah muntah nya serta mencret tolong solusi nya
Stemming After Stopword Removal (NLTK Library)	dok kemarin anak muntah muntah bawa bidan muntah muntah nya disertai mencret tolong solusi nya dok	dok kemarin anak muntah muntah bawa bidan muntah muntah nya serta mencret tolong solusi nya dok

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.4. IndoBERT

IndoBERT (Indonesia Bidirectional Encoder Representations from Transformers) is a BERT-based language model tailored for the Indonesian language. It was trained on Indo4B, a vast dataset comprising around 4 billion words from various sources, including formal and informal texts like news articles, social media posts, and Wikipedia. [17]. The key difference between BERT and IndoBERT is that BERT supports English or multiple languages, while IndoBERT is specifically designed for Indonesian. BERT is trained on datasets like English Wikipedia and BooksCorpus, whereas IndoBERT is trained exclusively on Indonesian texts from sources such as Indonesian Wikipedia and other local datasets.

BERT leverages the Transformer architecture, which learns word relationships through the attention mechanism. While the original Transformer includes both an encoder and a decoder, BERT uses only the encoder component to process input text and convert it into word vectors. This transformation relies on three types of embeddings: token embeddings where special tokens such as [CLS] indicate the start of a sentence and [SEP] mark the end or separate two sentences, segment embeddings which differentiate between paired input sentences, and positional embeddings which encode the position of each token within the sequence. These embeddings enable BERT to construct a language model applicable to various Natural Language Processing (NLP) tasks [18]. BERT's classification process involves two key stages, pre-training and fine-tuning. During pre-training, the model learns contextual representations from large amounts of unlabeled text, while in fine-tuning, it uses the pre-trained parameters and adjusts them further using labeled data for specific downstream tasks. Although both stages share the same architecture, they differ mainly in the output layer that is tailored for each task. The special tokens [CLS] and [SEP] play a crucial role in enabling BERT to transfer the learned representations during pre-training into fine-tuning for task-specific performance improvements [10]. Figure 3 illustrates the overall workflow of BERT's pre-training and fine-tuning stages by showing how the model is first trained on general language understanding objectives and later adapted to specific applications such as sentence classification, question answering, and named entity recognition.

For this study, the researcher utilized the pretrained model "indobenchmark/indobert-base-p1", which consists of 124.5 million parameters. As detailed in Table 3, the model was trained using the AdamW optimizer with a learning rate of 5e-5 and run for 25 epochs. The maximum token (sequence) length was set to 249, adjusted according to the maximum input length in the dataset to ensure no truncation of relevant information. Additionally, the batch size was set to 64 per device for training and 32 per device for evaluation, balancing training efficiency and computational constraints.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

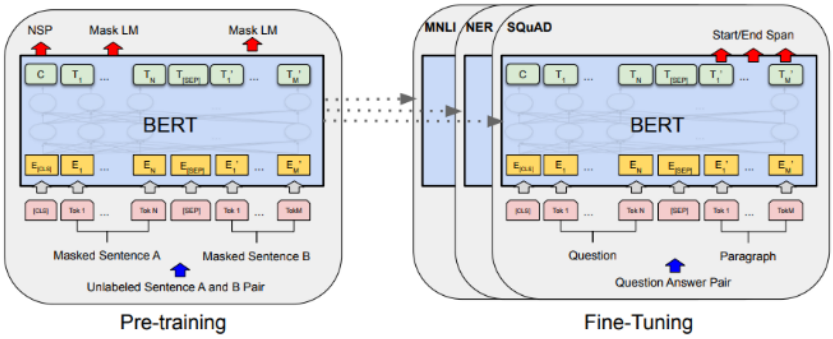

Figure 3. Illustration of BERT pre-training and fine-tuning process

Tabel 3. Model configuration

Pretrained Model	Hyperparameter	Values
indobenchmark/indobert-base-p1	Token Length	249
	Batch Size (Training)	64
	Batch Size (Validation)	32
	Epoch	25
	Learning Rate	5e-5
	Optimizer	AdamW

2.5. Telegram Bot

Telegram is an internet-hosted chat application that prioritizes rapid performance and data protection, allowing users to share text, audio, video, images, and stickers swiftly and securely [19]. BotFather is a bot account used to create bots on Telegram. The @BotFather account is the official bot creator account from Telegram and manages Telegram bots from this account. BotFather serves as the main controller for all bots created on the Telegram platform [20].

For the implementation of the bot, the best model from the previous experiments was chosen. Figure 4 illustrates the workflow of the child consultation chatbot. The user submits a question through the Telegram interface. The input is then classified by the IndoBERT model into a specific intent (tag) that reflects the topic of the question. There are two threshold conditions applied. First, the chatbot will provide an answer if the model's accuracy is above 0.1. Second, if the accuracy difference between the first and second classes meets the threshold (a maximum of 0.2), as well as the difference between the first and third classes (a maximum of 0.2), and so on, then all responses from the detected intent will be sent to the user. This threshold mechanism is intended to maintain response relevance and ensure

that users receive comprehensive answers when multiple intents are similarly probable.

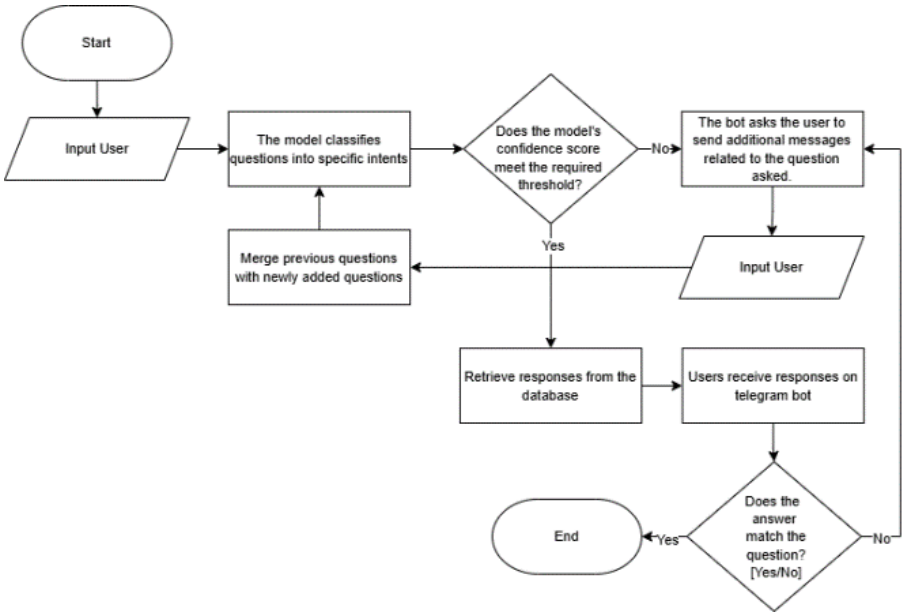

Figure 4. Bot system flow

2.6. User Acceptance Test (UAT)

User Acceptance Testing (UAT) is a process conducted to validate whether a system meets user needs and requirements in real-world usage scenarios. In this study, UAT was carried out to evaluate the chatbot's performance and usability from the users' perspective. The evaluation questions are divided into three main aspects: the user aspect, which measures user experience and ease of use, the interaction aspect, which examines how well the chatbot engages with users, including the relevance and accuracy of its responses, and the system aspect, which assesses the overall functionality and stability of the chatbot [21]. The questionnaire results are scored based on predefined indicators and point values, as detailed in Table 4.

Tabel 4. List of weighted UAT responses

Code	Description	Weight
VG	Very Good	5
G	Good	4
N	Neutral	3
PG	Pretty Good	2
NG	Not Good	1

3. RESULTS AND DISCUSSION

3.1. Model Evaluation

Table 5 presents the evaluation results of the IndoBERT model for the child consultation chatbot, using accuracy, precision, recall, and F1-score as metrics. The model was tested under four different preprocessing scenarios, as described in the Method section. Ten percent of the total dataset was reserved as a test set, while the remaining 90% was used for training and validation. To assess the effect of data split ratios on model performance, four training-validation splits were applied: 90:10, 80:20, 70:30, and 60:40. This approach helped evaluate the model's learning and generalization under varying data availability.

Tabel 5. Model evaluation on the test dataset

Scenario	Split Data	Accuracy	F1 Score	Precision	Recall
Stopword Term Frecuency-Based Without Stemming	90:10	97.5%	97.9%	98.3%	98.2%
	80:20	**98%**	**98,5%**	**98,9%**	**98,5%**
	70:30	97.3%	97.8%	98.3%	97.9%
	60:40	96.6%	96.8%	97.6%	97.2%
Stopword NLTK Library Without Stemming	90:10	96.6%	96.2%	97.1%	96.5%
	80:20	96.3%	96.1%	97.0%	96.5%
	70:30	96.3%	96.0%	97.1%	96.2%
	60:40	95.4%	95.6%	96.6%	95.9%
Stopword Term Frecuency Based and Stemming	90:10	97.5%	98.1%	98.4%	98.2%
	80:20	97.1%	97.8%	98.2%	98.1%
	70:30	97.3%	97.6%	98.4%	97.6%
	60:40	96.5%	97.0%	97.5%	97.4%
Stopword NLTK Library and Stemming	90:10	95.6%	95.6%	96.6%	96.1%
	80:20	96.5%	96.3%	97.2%	96.6%
	70:30	96.0%	95.7%	96.7%	96.0%
	60:40	95.8%	95.8%	96.8%	96.2%

Based on the evaluation results presented in Table 5, the stopword term frequency-based method without stemming performed the best, achieving an accuracy of 98%, an F1-score of 98.5%, a precision of 98.9%, and a recall of 98.5% in the 80:20 data split. This approach outperformed NLTK stopword methods due to its context-specific filtering based on word frequency. Stemming was found to reduce performance as it overly simplified words, losing important semantic distinctions. Therefore, the best-performing model will be implemented in the child consultation Telegram chatbot to enable accurate intent classification. To illustrate the model's evaluation results, several examples of misclassification observed in the test data are presented in Table 6.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Tabel 6. Examples of misclassified test dataset

Text	Actual Tag	Prediction Tag
dok anak saya berusia 2th 6 bulan. di badan anak saya tumbuh bintik2 merah seperti biang keringat,tp badan anak saya panas dok . dan dibibir sama lidah nya tumbuh seperti sariawan. tapi setelah saya minumi obat panas nya turun,untuk bintik2 nya apabila badan berkeringat dia merasa gatal.apa kah iini gejala campak dok? (Doc, my child is 2 years and 6 months old. On my child's body, red spots like prickly heat have grown, but my child's body is hot, Doc. And on his lips and tongue, there are growths like canker sores. But after I gave him medicine, his fever went down. As for the spots, when his body sweats, he feels itchy. Is this a symptom of measles, Doc?)	*campak_ pada_ anak* (measles_ in_children)	*biang_keringat _pada_anak* (prickly_heat_ in_children)
Dok anak saya umur 7.3bln mengeluh sakit perut bagian atas pusar.sebelum nya mengalami panas dulu dan kulit bentol bentol.tapi sekarang panas nya sudah turun yg d keluhkan sekarang perut bagian atas terasa sakit (Doc, my 7.3 month old child is complaining of pain in the upper abdomen in the navel. Previously he had a fever and his skin was covered in bumps. But now the fever has gone down, what he is complaining about now is that his upper abdomen feels sore.)	*nyeri_perut _pada_anak* (abdominal _pain_in _children)	*biduran_ pada_anak* (hives_in_ children)
anak sya umur 6.tahun..sdh 3 hari ini demam.sdh d bwa berobat sampai ganti2 dokter tp ttp sja blm ada perubahan.sebentar panas sebentar dingin..klo obat d minumkan demam turun tp habis itu panas lg..pnya riwayat amandel sma types..mohon bantuannya dok.. (My 6 year old child has had a fever for 3 days. I have taken him to different doctors for treatment but there is still no change. He is hot for a while and then cold. If he is given medicine, the fever goes down but then he is hot again. He has a history of tonsillitis and typhoid. Please help, doc.)	*demam_ pada_anak* (fever_in _children)	*amandel_pada _anak* (tonsils_in_ children)

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The misclassifications presented in Table 6 indicate that while the model can identify certain symptoms, it sometimes misinterprets the primary concern of the user's message. In the first case, although the user clearly suspects measles due to the combination of red spots, fever, and sores in the mouth, the model incorrectly predicts "prickly heat" as the tag, likely influenced by the mention of itchy red spots. In the second example, the user highlights abdominal pain as the main complaint after fever subsided, yet the model misclassifies it as "hives," possibly due to the mention of previous skin bumps. Lastly, in the third case, even though the main concern is persistent fever, the model predicts "tonsils in children," which may be due to the presence of the word "tonsils" in the child's medical history, even though it is not the main issue. These errors indicate that the model relies heavily on specific keywords rather than fully understanding the context or accurately identifying the primary health complaint described. Additionally, the misclassification can also be attributed to the similarity of symptoms between certain conditions for example, measles and hives both may present with rashes, while tonsillitis is often accompanied by fever, which could lead to confusion in intent classification.

3.2. Telegram Bot

The child consultation chatbot on Telegram is shown in Figure 5. The screenshot displays a chat with the bot, including a greeting, a question, and a response confirmation. The left screenshot shows an example of a greeting chat and a thank you response, the middle screenshot shows an example of a user's question to the chatbot, and the right screenshot shows the response confirmation provided by the chatbot.

Figure 5 Child consultation chatbot on Telegram (1)

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 6 below shows a screenshot of an interaction with the child consultation bot. The left and middle images display the bot confirming a response with "No," requiring additional input from the user. Meanwhile, the right image shows the model failing to generate a response due to not meeting the threshold, prompting the user for further input.

Figure 6. Child consultation chatbot on Telegram (2)

Figure 7 shows a screenshot of an interaction with the child consultation bot. The bot provides multiple possible answers because the model's predictions have a small difference and meet the set threshold, resulting in three responses instead of just one. For example, if a user sends a question like "my child has a cough," the model might be uncertain because there are several types of coughs in the dataset, such as productive cough, dry cough, and flu-related cough. Therefore, a threshold is needed to display all responses when the model's prediction scores are close to each other. This is where the role of the threshold becomes crucial it allows the bot to present several potentially relevant answers when prediction scores are close, ensuring that users receive more comprehensive and contextually appropriate responses.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 7. Child consultation chatbot on Telegram (3)

3.3. UAT Result

The child consultation Telegram chatbot was evaluated using User Acceptance Testing (UAT), conducted through a questionnaire form. Participants were required to interact with the chatbot first before completing the UAT questionnaire. A key requirement for respondents was that they must have prior experience in caring for children. A total of 30 qualified respondents successfully participated in the testing. The results of the UAT calculation are presented in Table 7 and Table 8.

Tabel 7. Question list with response counts per answer choice

No	Question	Usability Aspect				
		VG	G	N	PG	NG
User Aspect						
P1	Is the chatbot easy to use?	16	14	0	0	0
P2	Does the chatbot provide clear user instructions?	14	14	2	0	0
P3	Is the chatbot easy to access and use via Telegram without any issues?	15	14	1	0	0
Interaction Aspect						
P4	Is the information provided by the chatbot as expected?	9	20	1	0	0

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

No	Question	Usability Aspect				
		VG	G	N	PG	NG
P5	Is it easier to get information about child consultation via chatbot than via Web Browser?	17	9	4	0	0
System Aspect						
P6	Do you agree that the chatbot response time is fast enough?	20	10	0	0	0
P7	Are you satisfied with the performance of the chatbot?	13	17	0	0	0
P8	Does the chatbot function properly without frequent glitches or errors?	16	14	0	0	0

Table 7 presents a question list with response counts for each answer choice. The table includes a total of 8 questions asked to respondents, each with 5 answer options. Table 8 below shows the results of the UAT score calculations.

Tabel 8. UAT value calculation

No	Usability Aspect					Qty	Qty/ Resp.	%	AVG
	VG x 5	G x 4	N x 3	PG x 2	NG x 1				
User Aspect									
P1	80	56	0	0	0	136	4.4	90.7%	
P2	70	56	6	0	0	132	4.4	88%	89.3%
P3	75	56	3	0	0	134	4.47	89.3%	
Interaction Aspect									
P4	45	80	3	0	0	128	4.27	85.3%	87%
P5	85	36	12	0	0	133	4.43	88.7%	
System Aspect									
P6	100	40	0	0	0	140	4.67	93.3%	
P7	65	68	0	0	0	133	4.43	88.7%	90.9%
P8	80	56	0	0	0	136	4.53	90.7%	

Table 8 displays the UAT score calculations. The calculation is done by multiplying the number of responses by the weight of each aspect, then summing the total score from all aspects (Qty). Next, the average score for each question is obtained by dividing the total score by the number of respondents (30). This average score is then divided by 5 (the number of aspects) and multiplied by 100% to convert it into a percentage. Finally, the average (avg) is calculated to determine the mean score for each aspect in the UAT evaluation. Based on the usability level, the results of the User Acceptance Test (UAT) indicate a usability score above 80%, reflecting excellent performance across all evaluated aspects. The user aspect achieved an

average score of 89.3%, the interaction aspect scored 87.3%, and the system aspect reached 90.9%.

3.4. Discussion

Based on the comparison in Table 2, the stopword term frequency–based method without stemming achieved the best performance in the 80:20 data split scenario, reaching 98% accuracy, a 98.5% F1-score, 98.9% precision, and 98.5% recall. The method using term frequency with stemming showed slightly lower results, while both NLTK stopword approaches whether with or without stemming performed less effectively overall. The higher effectiveness of the custom stopword method is likely due to its context-specific nature, as it removes words based on their frequency in the dataset, eliminating extremely common or rare terms that may carry little value for classification. In contrast, the default NLTK stopword list may not fully align with the characteristics of the domain-specific data used. Additionally, evaluation results showed a decline in model performance when stemming was applied. This may be due to the aggressive nature of some stemming algorithms, which can reduce words to forms that do not reflect their original meaning accurately. For instance, "*belakangan*" becomes "*belakang*," "*kasian*" becomes "*kasi*," and "*perawatan*" becomes "*awat*." Such transformations can result in the loss of important semantic distinctions, reducing the richness and precision of features used in classification.

Then for the child consultation Telegram chatbot, researchers used the best-performing model, which was the stopword term frequency-based scenario without stemming and with an 80:20 data split. The bot successfully classified user intents and was also able to respond appropriately to greeting messages, expressions of gratitude, and other general inputs. The threshold set for intent classification helped guide users effectively by providing varied responses when the confidence scores between the top predicted classes were close. This threshold was carefully chosen to balance accuracy and flexibility in response generation, ensuring that the chatbot avoids producing irrelevant replies when prediction confidence is low. By applying a well-calibrated threshold, the chatbot was able to handle ambiguity and offer alternative yet contextually relevant responses, thereby improving user satisfaction and trust.

To evaluate usability, a User Acceptance Test (UAT) was conducted with 30 respondents experienced in child caregiving. After interacting with the chatbot, they completed a questionnaire covering user experience, interaction quality, and system performance. The chatbot received high usability scores: 89.3% for the user aspect, 87.3% for interaction, and 90.9% for system performance, indicating strong overall effectiveness. However, a few misclassifications were observed in the test data, which can be attributed to the limited size and diversity of the dataset. The

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

lack of variation in some intent categories reduced the model's ability to generalize to unseen queries. Expanding the dataset with more varied examples would likely enhance the chatbot's robustness and its performance across a broader range of user inputs.

4. CONCLUSION

This study presents a novel application of IndoBERT for intent classification in a Telegram-based Indonesian-language chatbot focused on child health and parenting addressing a gap where few transformer-based consultation systems exist for this domain. The primary contributions include development of an intent classification model tailored to Bahasa Indonesia parenting and child-health queries, and a systematic comparison of preprocessing techniques, demonstrating that term-frequency–based stopword removal outperforms general-purpose libraries while stemming degrades semantic richness. Achieving 98% accuracy, 98.5% F1-score, 98.9% precision, and 98.5% recall affirms the model's reliability. To assess usability, a User Acceptance Test (UAT) involving 30 respondents with child caregiving experience was conducted. The chatbot achieved high usability scores 89.3% for user experience, 87.3% for interaction quality, and 90.9% for system performance indicating strong user satisfaction. Despite these strengths, the system relies on static responses without multi-turn context tracking and is constrained by a dataset drawn solely from Halodoc and Alodokter. Future work should integrate dialog-level models for sustained multi-turn interactions, implement user feedback loops for continuous learning, and expand data coverage to a wider range of parenting and pediatric topics. Practically, this chatbot offers a scalable, rapid-response consultation tool that can empower Indonesian parents especially in underserved areas to make timely, informed decisions about their children's health.

REFERENCES

[1] R. K. Sari, S. P. Astuti, M. Sari, and R. N. Syari'ati, *Profil Kesehatan Ibu dan Anak 2022*. Jakarta: Badan Pusat Statistik, Jakarta - Indonesia, 2022.

[2] Guy Allison *et al.*, "Next Generation Indonesia," *British Council*, vol. 1, no. 1, p. 63, 2022.

[3] Soenarto, Y., Trisnantoro, L., and Fuad, A., "Penyebaran Spesialis Anak di Indonesia Tahun 2004: Implikasinya Terhadap Kebijakan Kesehatan dan Pendidikan," *Sari Pediatri*, vol. 8, no. 2, pp. 94–99, 2016.

[4] G. Achmad Marzuki and A. Setyawan, "Peran Orang Tua Dalam Pendidikan Anak," *JPBB : Jurnal Pendidikan*, vol. 1, no. 4, 2022.

[5] R. A. Sekarwati, A. Sururi, R. Rakhmat, M. Arifin, and A. Wibowo, "Survei Metode Pengujian Chatbot pada Media Sosial untuk Mengukur Tingkat Akurasi," *Sisfotenika*, vol. 11, no. 2, p. 172, 2021.

[6] N. Shahin and L. Ismail, "From Rule-Based Models to Deep Learning Transform-ers Architectures for Natural Language Processing and Sign Language Translation Systems: Survey, Taxonomy and Performance Evaluation," 2024. doi: 10.1007/s10462-024-10895-z.

[7] D. Griol, Z. Callejas, J. M. Molina, and A. Sanchis, "Adaptive dialogue management using intent clustering and fuzzy rules," *Expert Syst*, vol. 38, no. 1, 2020.

[8] Ahmet Birim and Mustafa Erden, "Robustness to Spelling Errors for Intent Detection," *2022 30th Signal Processing and Communications Applications Conference (SIU)*, Aug. 2022.

[9] A. Dwiyono, M. Fachrurrozi, J. Palembang-Prabumulih, K. Ogan Ilir, and S. Selatan, "Analisis Perbandingan Klasifikasi Intent Chatbot Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT," *Journal of Information System Research*, vol. 6, no. 1, pp. 605–616, 2024, doi: 10.47065/josh.v6i1.6051.

[10] P. Sayarizki and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Journal on Computing*, vol. 9, no. 2, pp. 61–72, 2024, doi: 10.34818/indojc.2024.9.2.934.

[11] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, pp. 97–100, 2023.

[12] H. Hendiana, A. Irma Purnamasari, and I. Ali, "Analisis Sentimen Komentar Berita Detik.Com Menggunakan Algoritma Suport Vektor Machine (Svm)," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 3175–3181, 2024, doi: 10.36040/jati.v8i3.8421.

[13] P. A. Setiawati, I. M. A. D. Suarjaya, and I. N. P. Trisna, "Sentiment Analysis of Unemployment in Indonesia During and Post COVID-19 on X (Twitter) Using Naïve Bayes and Support Vector Machine," *Journal of Information Systems and Informatics*, vol. 6, no. 2, pp. 662–675, 2024, doi: 10.51519/journalisi.v6i2.713.

[14] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[15] J. Pardede and D. Darmawan, "Perbandingan Algoritma Stemming Porter , Sastrawi , Idris, dan Arifin & Setiono Pada Dokumen Teks Bahasa Indonesia," vol. 12, no. 1, 2025, doi: 10.25126/jtiik.2025128860.

[16] N. Rajkumar, T. S. Subashini, K. Rajan, and V. Ramalingam, "Tamil Stopword Removal Based on Term Frequency," *Advances in Intelligent Systems and Computing*, p. 21, 2020.

[17] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *arXiv preprint arXiv:2009.05387*, 2020.

[18] Budi Juarto and Yulianto, "Indonesian News Classification Using IndoBert," 2023. [Online]. Available: www.ijisae.org

[19] D. Abimanto and I. Mahendro, "Penggunaan Aplikasi Telegram Untuk Kegiatan Pembelajaran Jarak Jauh pada Mata Kuliah Bahasa Inggris Materi Speaking pada Mahasiswa Universitas Maritim AMNI Semarang," *Prosiding Kemaritiman*, pp. 245–256, 2021.

[20] R. P. Yuwan, R. Soelistijadi, and E. Zuliarso, "Implementasi Chatbot Telegram untuk Meningkatkan Kualitas Layanan Jaringan Internet Pada Layanan ICONNET Menggunakan Penerapan Metode Action Research (AR)," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, pp. 40–48, 2024, doi: 10.35870/jtik.v8i1.1431.

[21] Ni Putu Utari Dyani Laksmi, Oka Sudana, and Agung Cahyawan, "Innovative Learning Model for Dharmagita Based on Telegram Chatbot," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 2, pp. 248–257, 2024, doi: 10.23887/janapati.v13i2.78535.