An Indonesian Mental Health Chatbot Model Based on A Sequence to Sequence LSTM

Nia Ekawati[1,2*], Imam Riadi[3], Herman Yuliansyah[4]

[1,3,4]*Doctoral Program of Informatics, Universitas Ahmad Dahlan, Indonesia*
[2]*Department of Informatics, Politeknik TEDC, Indonesia*

*corr-author: 2436083020@webmail.uad.ac.id

Abstract – This research proposes an Indonesian-language mental health chatbot model based on the LSTM Sequence-to-Sequence (Seq2Seq) architecture as an adaptive initial support solution. Unlike static classification models, this generative approach aims to capture emotional dependencies and conversational context through context vectors. The research methodology utilizes the public PSYCHIKA dataset, which includes 5,667 conversation pairs a significant volume for a low-resource language. Evaluation was conducted by comparing 80:20 and 70:30 data split schemes. Experimental results showed the best performance with the 80:20 split, achieving a BLEU-1 score of 0.137, compared to the 70:30 split, which only reached 0.043. The model achieved stable convergence at 15–16 epochs via an early-stopping mechanism without any signs of overfitting. Although training stability was maintained, the low BLEU score confirms that the use of a pure Seq2Seq LSTM without an attention mechanism is not yet sufficient to generate highly fluent responses. These findings provide a reproducible technical baseline for the development of mental health dialogue systems in Indonesia, while also emphasizing the urgency of more advanced architectures to improve the quality of empathy in the future.

Keywords: Dialogue response generation, long short-term memory, low-resource language, mental health chatbot, sequence to sequence model.

I. INTRODUCTION

Mental health challenges are often overlooked when anxiety and cognition shift in extreme ways, even though these changes can seriously disrupt daily life [1, 2] . Because many people cannot access timely treatment, mental health chatbots have emerged as a practical solution for immediate online consultation and more accessible support [3, 4]. Chatbots are widely used to improve human-computer interaction across many sectors because they can understand natural language and generate relevant responses [5, 6]. In healthcare, they help with patient monitoring, appointment scheduling, and the delivery of medical information, while their reliability depends heavily on conversation modeling that combines AI and NLP [7, 8] .

Related research indicates that mental health and medical chatbots have steadily advanced in terms of functionality, accessibility, and personalization, enabling broader reach and more tailored user interactions; nevertheless, persistent ethical concerns (such as privacy, informed consent, and potential misdiagnosis) and technical limitations (including data biases and robustness) continue to constrain their safe deployment [9, 10]. The digitalization of mental health services in Indonesia has demonstrated significant progress through Transformer-based classification models such as IndoBERT and DNN-BERT, which are capable of mapping user inputs into predefined categories with high accuracy [11]. However, this purely classification-based approach is inherently static because it only retrieves preselected responses, and therefore it is less capable of accommodating truly personalized dialogue. On the other hand, traditional LSTM-based conversational models often fail to capture the subtle emotional dynamics present in clinically relevant dialogues. Instead, such models tend to generate bland and generic responses that are neither sufficiently empathetic nor therapeutically specific. Consequently, there is an urgent need for architectural innovation such as emotion-aware encoders, attention mechanisms adapted to affective cues, and multimodal or hierarchical designs that can represent and respond to complex emotional states more accurately, thereby improving clinical utility and user trust in mental health chatbots [12].

The realization of such architectural innovation depends not only on model sophistication but also on the availability of high-quality data for training. Therefore, one significant contribution of this research lies in the utilization of an Indonesian mental health conversational dataset with a substantially larger volume than that used in previous studies within the same domain. This research employs the public PSYCHIKA dataset, which consists of 5,667 dialogue pairs (questions and

responses). The use of this larger dataset is intended to mitigate the risk of overfitting that often occurs in deep learning architectures trained on limited data, while also improving the model's ability to capture long-range contextual dependencies in informal Indonesian language. For comparison, most previous mental health chatbot studies in Indonesia have used relatively small-scale datasets. For example, a dataset consisting of 1,544 entries, compiled through interactions during a limited number of counseling sessions, serves as the knowledge base for the chatbot model that was developed [13]. Meanwhile, the CuraBot platform was developed using a more limited dataset, consisting of 1,624 conversation data points [14]. The dataset from the previous research, which consisted of 5,079 entries, focused on mapping single emotion labels from Twitter rather than on generating consultative dialogue. As a result, there is a gap in the availability of datasets capable of training models to generate natural, generative dialogue responses for users [15]. Accordingly, the 5,667 dialogue pairs used in this research provide an advantage in terms of linguistic variation coverage and emotional contextual richness for the proposed Sequence-to-Sequence LSTM model.

The use of deep learning technology has enabled the development of language-specific chatbots that are adaptive to various linguistic contexts [16, 17]. This research presents a deep learning-based conversation system designed as a means of initial mental health treatment for the general public [18]. The presence of mental health chatbots represents a trend in the use of seq2seq architecture in building text interactions aimed at supporting the psychological well-being of users [19]. The main drawback of traditional intent classification models is their inability to maintain dialogue memory, resulting in responses that tend to be static and lack empathetic nuance in lengthy interactions [11]. In contrast to these models, the LSTM Sequence-to-Sequence (Seq2Seq) architecture has the unique ability to process sequential data and maintain long-term dependencies. This is crucial for capturing subtle shifts in the user's emotions throughout a conversation session, ensuring that the chatbot not only responds to the most recent message but also takes into account the context of previous interactions to generate more coherent and clinically relevant dialogue [20]. In general, the main contributions of this research can be summarized as follows:

- a baseline implementation of an Indonesian-language mental health chatbot using a Seq2Seq LSTM architecture [21];
- evaluation of the model using BLEU-based metrics to analyze response quality and contextual limitations [22].

This paper is organized as follows: Section II presents the methodology, experimental setup, and evaluation; Section III discusses the findings and future research directions; and Section IV summarizes the main contributions to chatbot technology.

II. METHOD

This research adopts an experimental approach by using an LSTM-based Seq2Seq encoder-decoder architecture to handle varying sequence lengths and improve context mapping [23]. It begins with collecting and preprocessing an Indonesian conversation-pair dataset to train the model in generating meaningful vector representations for natural language interaction [24]. The model is designed for mental health support, and its performance is evaluated using BLEU-based metrics to assess response quality and contextual limitations [25]. The research methodology is shown in Fig. 1.

Fig. 1 presents the research methodology for developing the mental health chatbot model. The process begins with collecting the PSYCHIKA public dataset from Kaggle, followed by processing the sequence-to-sequence data using the LSTM algorithm, and concludes with performance evaluation using the BLEU score. The detailed explanation of Fig.1 is as follows:

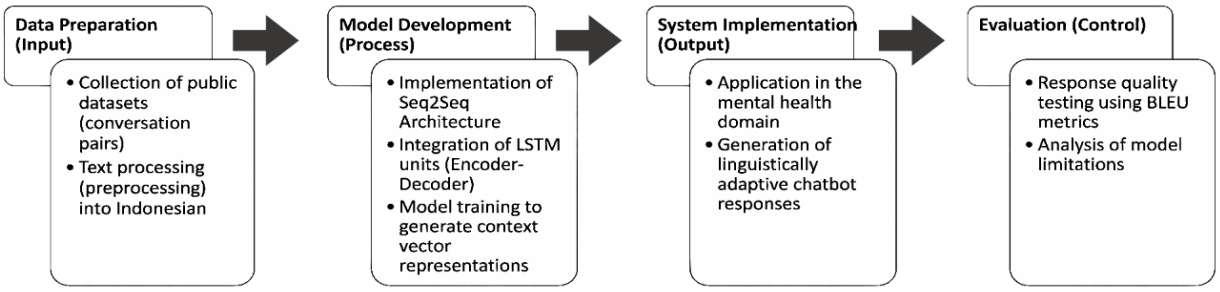


Fig. 1 The research methodology

A. Data Preparation

This stage serves as the fundamental basis of the research. Before model training, the data must be collected and cleaned to ensure that the model can learn effectively. Public conversation-pair datasets are collected, consisting of paired text samples such as user questions or statements and their corresponding responses. Because the target domain is mental health, the dataset is expected to contain counseling dialogues, psychological complaints, or supportive interactions [26]. The raw text data are then processed through several preprocessing steps, including case folding, filtering or cleansing, tokenization, and normalization into consistent and proper Indonesian language. To prevent data leakage, we verified that no duplicate conversation pairs existed between the training, validation, and test sets by hashing each input-output pair and checking for intersections across splits [27].

The research begins by setting up the Python environment with essential libraries for text processing, data preparation, model building, and BLEU-based evaluation. It uses a public XLSX dataset consisting of input-output conversation pairs to train the chatbot and improve its ability to generalize across diverse conversational inputs, the utilization of a publicly available dataset consisting of 5,667 conversational pairs, can be seen in Table I.

Table I illustrates that the workflow starts by importing an Excel dataset into a DataFrame, with the Context column designated as the model input and the Response column serving as the ground-truth reference. After loading, the dataset undergoes structural validation verifying column presence, data types, and the absence of missing or malformed entries to ensure integrity and consistency before it proceeds to subsequent preprocessing and model-training stages. The architecture of LSTM encoder-decoder model can be seen in Fig. 2.

Fig. 2 illustrates LSTM-based Encoder-Decoder Model Architecture designed for natural language processing tasks such as automatic text summarization. The process begins with data retrieval from the Public Dataset, which then enters the Encoder stage to process the input sequence; here, LSTM technology plays an important role in extracting features and maintaining long information contexts. The result of this encoding process is summarized into a Context Vector, which is a dense numerical representation that captures the essence of the input text, which is then translated back by the Decoder into an output sequence in the form of a text summary [28]. Finally, the quality of this model is measured through the Evaluation Performance stage using the BLEU metric, which calculates the accuracy level of the model's prediction results compared to human reference text through precise mathematical calculations [29]. The model's ability to continuously understand emotional context is achieved through an encoder-decoder mechanism. In the encoder stage, the Indonesian language input sequence is processed sequentially to extract linguistic features and affective cues, which are summarized into a Context Vector. This vector serves as a compact numerical representation that captures the emotional "essence" of the user's input. Next, the LSTM uses a gating mechanism—consisting of a forget gate, an input gate, and an output gate—to selectively regulate the flow of information. The forget gate plays a crucial role in discarding irrelevant information, while the memory cells retain significant emotional context to be passed on to the decoder stage to dynamically generate an empathetic response.

B. Model Development

This stage represents the core of the artificial intelligence development process, where the deep learning architecture is designed and trained. The model adopts a Sequence-to-Sequence (Seq2Seq) architecture, which is particularly suitable for chatbot applications because it maps an input sequence of words to an output sequence of words. Within the Seq2Seq framework, two main components use Long Short-Term Memory (LSTM) units: the encoder and the decoder. The encoder reads and interprets the user's input sentence word by word while preserving important information, whereas the decoder uses this information to gradually generate a coherent response. During training, the model learns to compress the meaning of the input sentence into a mathematical representation known as a context vector, which serves as the bridge between the encoder and decoder so that contextual information is not lost [30].

C. System Implementation

After the model has been trained and is capable of generating responses, it is implemented into a real system for user access. In this research, the chatbot is specifically deployed in the mental health domain. As a result, the system functions as a digital listening agent, a provider of psychological first aid, or a conversational companion. The system produces text-based outputs in the form of chatbot responses. These responses are linguistically adaptive, meaning that the language style is natural, empathetic, and grammatically appropriate in Indonesian, enabling users to perceive the interaction as more human-like rather than mechanical [31].

TABLE I
PUBLIC DATASET

No	Context	Responses
1	*Saya menderita ADHD dewasa, gangguan kecemasan, dan depresi. Sulit menemukan dokter di daerah saya dan dokter utama saya tidak akan membantu. Saya menganggur dan kewalahan. Apa yang akan Anda sarankan saya lakukan?*	*Jika hanya konseling yang Anda cari, sejumlah pakaian berbasis agama sangat bersedia untuk mendengarkan dan membantu dengan hal-hal semacam ini, pesan online gratis dan media sosial adalah pilihan sekunder, namun yang satu ini mungkin datang dengan privasi kekhawatiran dan konsekuensi; jika itu adalah saya, saya akan berusaha untuk berbicara dengan manis satu atau dua penasihat yang saya temui untuk melakukan sedikit pekerjaan untuk orang-orang yang tidak mampu membelinya*
	…	…
	…	…
5667	*Saya telah mengalami banyak kecemasan dan serangan panik belakangan ini. Saya baru-baru ini didiagnosis oleh psikiater saya dengan gangguan obsesif-kompulsif. Akhir-akhir ini, saya telah mempertanyakan segalanya dari karier saya hingga hubungan saya. Pacar saya dan saya baru saja pindah beberapa bulan yang lalu. Tiba-tiba, saya tidak merasa nyaman di sekitarnya seperti dulu, meskipun saya tidak bisa menemukan alasan mengapa saya merasa seperti ini.*	*Bagaimana pacar Anda menanggapi ketidaknyamanan Anda di sekitarnya? Apakah dia sudah memberi tahu Anda bahwa dia memperhatikan perubahan pada Anda? Kabar baik tentang suatu hubungan adalah bahwa Anda dapat membicarakan hal-hal dengan orang lain. Melakukan hal itu hanya dapat membantu memperjelas perasaan dan apa yang Anda harapkan dari hubungan itu. Mengingat bahwa Anda berdua baru-baru ini pindah bersama, wajar untuk dinamika, perasaan, dan harapan baru muncul. Itu hanya menambah kewajaran untuk berbicara satu sama lain tentang bagaimana hidup bersama rasanya bagi Anda masing-masing. Psikiatri mencari nafkah dengan mendiagnosis orang dan mengatakan kepada mereka untuk minum pil. Sangat sering, hanya diberi tahu bahwa orang itu memiliki "kondisi" membuat mereka merasa rapuh dan kurang mampu daripada yang sebenarnya. Tanda harus mengatasi banyak atau hal mendalam. Bersabarlah dan berikan diri Anda waktu untuk mempelajari detail hubungan Anda dan apa pun rincian kariernya yang mengganggu Anda. Anda tampaknya sadar diri karena Anda yang menggambarkan situasi bermasalah Anda sendiri. Simpan label "tidak teratur". Label psikiatri lebih baik untuk psikiater daripada yang mereka lakukan untuk orang-orang yang mencoba menjalani hidup mereka.*

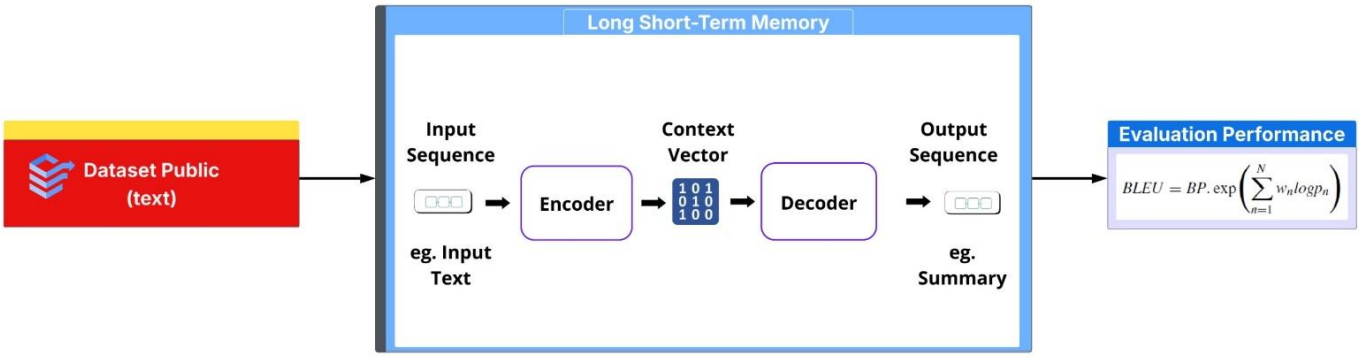


Fig. 2 The architecture of LSTM encoder-decoder model

D. Evaluation

This final stage functions as a quality control mechanism to ensure that the chatbot is safe and suitable for use, particularly because mental health is a highly sensitive domain. The quality of the chatbot responses is evaluated quantitatively using the BLEU (Bilingual Evaluation Understudy) metric. This metric measures the similarity between the chatbot-generated responses and the reference responses prepared by human experts. A higher BLEU score indicates that the chatbot responses are more accurate and natural. In addition, the model limitations are analyzed in depth to identify weaknesses such as repetitive answers, failure to capture complex emotional contexts, or hallucination, in which the model may generate medically inaccurate information. This analysis is essential for guiding future system improvements [32].

Fig. 3 shows the flow of information in the Seq2Seq model: the Embedding layer converts input tokens into dense vectors, the Encoder compresses them into hidden and cell states, and these states are passed to the Decoder to generate responses. During decoding, the model uses either the previous target token or its own last output to produce the next token, allowing it to generate coherent context-based responses [33]. The Seq2Seq model is shown in Fig. 3.

Fig. 3 illustrates the Seq2Seq model architecture used for machine translation or natural language processing tasks, where the process is divided into two main parts, namely Encoder and Decoder. In the Encoder stage (left side), the model receives the input sentence in Indonesian, “Bagaimana Cara Melawan ini ?”, which is processed sequentially through five time steps; each word is converted into a numerical representation through the word2id mechanism and embedding layer before being entered into a cell (purple box) to extract the contextual meaning of the entire sentence into a single context vector. This vector acts as a bridge of information to the Decoder stage (right side), which begins with the trigger token <GO> at time step 6 to generate the prediction of the first word, namely “Terapi”. The process in the Decoder is autoregressive, where the output from one step (for example, the word ‘Terapi’) is fed back as input for the next step to predict the words “Cahaya”, “Sangat,” and “Membantu” in sequence, until the model successfully transforms the original input into a complete sentence on the output side through the id2word layer, which returns the numerical code back to human text. This entire process ultimately produces a calibrated training_model object that is ready to be operated in a real environment to automatically respond to user conversations [34].

The formula in the Fig. 3 explains how LSTM works in the encoder-decoder process. The encoder reads the input sentence sequentially and stores it in a context vector that represents the entire information of the sentence [35]. This context vector is then passed to the decoder as the initial state to generate a response word by word in an autoregressive manner. Each output word is selected from the highest probability produced by the softmax function, then mapped back into text through the id2word process [5, 18].

To provide a comprehensive understanding, the Sequence-to-Sequence (Seq2Seq) architecture in this mental health chatbot is mathematically formulated through two main components: the Encoder and the Decoder.

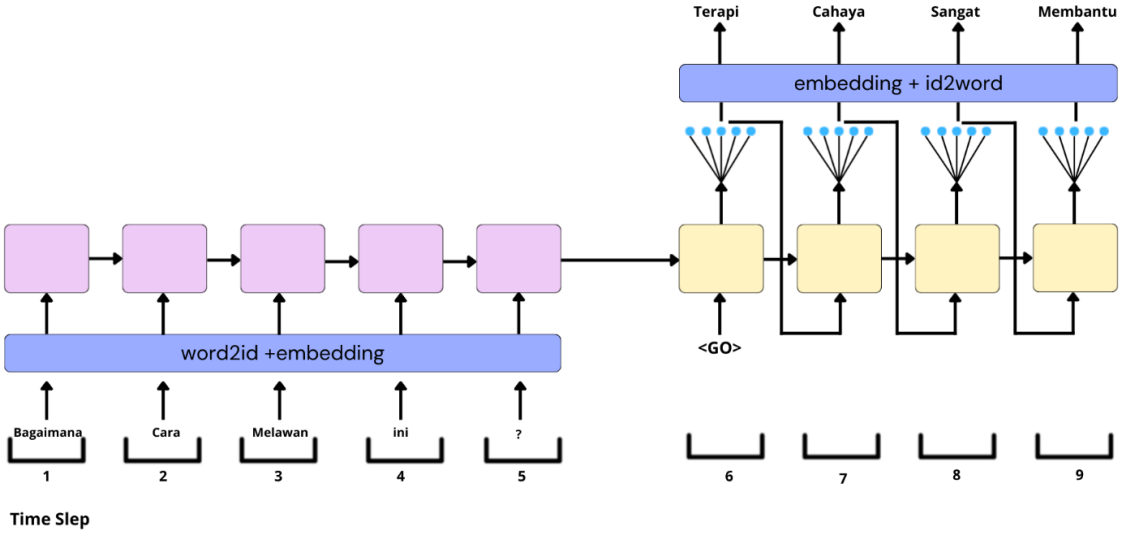


Fig. 3 Seq2seq model encoder-decoder

1) Functions and Equations of the Encoder: The encoder aims to convert the sequence of user input tokens $X = \{x_1, x_2, \ldots, x_n\}$ into a single context vector that represents the essence of the information. Each LSTM cell in the encoder processes one token at a time step while taking into account the previous hidden state h_{t-1} [36]. The key to LSTM's ability to handle long-term dependencies lies in its gate mechanism. The key equation underlying these cells is:

- Forget Gate (f_t): Determines which information should be discarded from long-term memory as in (1).

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

- Input Gate (i_t): Determines which new information will be stored in the cell state as in (2).

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

The context vector is the final hidden state of the encoder, which symbolically summarizes the entire input sequence: $C = h_n$ [36].

2) Decoder Functions and Equations: The decoder predicts the sequence of responses $Y = \{y_1, y_2, \ldots, y_m\}$ based on the context vector *C*. Mathematically, this model operates by maximizing the conditional probability of each response token given the previously generated tokens and the input context vector [36], [37]. The generative probability equation for each time step in the decoder is (3)

$$P(y_t | y_{<t}, C) \quad (3)$$

where $y_{<t}$ represents the sequence of words generated prior to *t* [36]. During the training phase, the model uses the Maximum Likelihood Estimation (MLE) approach to minimize the distance between the prediction and the human reference answer, with the following objective function (4)

$$\max \theta \sum log\pi\theta\,(x_t | s_t) \quad (4)$$

where θ are the model parameters optimized to generate the most relevant responses given the mental health context [37].

3) Evaluation Using BLEU: Once the decoder generates a response, its quality is evaluated using the BLEU metric [36], [38]. This metric works by comparing the $n - gram$ matches between the chatbot's response and the human reference response, while applying a brevity penalty to maintain a balance between precision and sentence length [36].

Based on the explanation in Fig. 3, a detailed description of one pink block for LSTM model can be seen in Fig. 4.

In Fig. 4, the Long Short-Term Memory (LSTM) network is built around core elements known as gating units and memory units. The primary function of this gating mechanism is to carefully regulate which information should be retained and which should be discarded, particularly regarding the features being processed [39]. The conventional LSTM structure includes three crucial types of gates: the input gate, the forget gate, and the output gate. The internal computational process of the LSTM is described through the inference equation [6]. Referring to Figure 2, f_k represents the forget gate, h_{k-1} is the hidden state from the previous time step (*k*-1), and *xk* is the input data at the current time step (*k*). Specifically, the forget gate plays a vital role in deciding which information from the previous hidden state will be passed on to the next stage, according to the presented formula [40].

The final stage in Fig. 2 evaluates model performance using the BLEU metric by comparing the generated responses with reference answers after text normalization and tokenization [41]. It also uses smoothing to handle short chatbot outputs and produce more reliable BLEU scores as in (5). BLEU score is an evaluation metric that measures the quality of text generated by a model by comparing it to human reference text. This measurement standard is widely used in machine translation and various other automated text generation tasks.

$$BLEU = BP.\exp(\textstyle\sum_{n=1}^{N} w_n log p_n) \quad (5)$$

In its calculation, BLEU involves duration penalty factors (brevity penalty), weight (w_n), and n-gram precision (p_n). The scoring scale ranges from 0 to 1; the closer to 1, the more accurate the text is considered to be in relation to the reference. The advantage of this metric lies in its ability to review consistent word sequences (n-grams), there by providing a strong indication of the quality of the model's output [42].

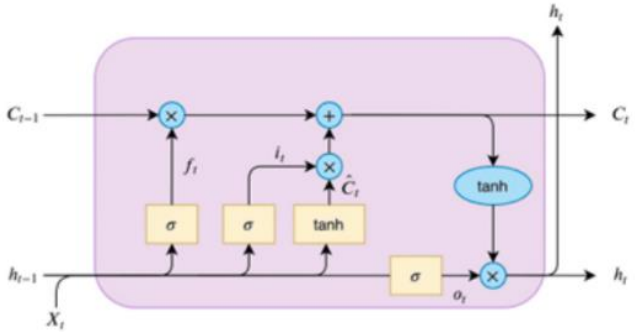


Fig. 4 LSTM model

III. RESULTS AND DISCUSSION

This discussion interprets these findings, linking the technical capabilities of the LSTM architecture to the practical needs of individuals seeking immediate mental health assistance. The experiment was conducted using Google Colab to optimize computing speed. The model development and analysis process relied on Python3 with connect to a hosted runtime A100 to ensure smooth implementation and data visualization [43]. The dataset used in the training and validation process consisted of public dataset.

This research applies two data distribution schemes detailed in Table II, namely the 80:10:10 and 70:20:10 ratios, where each part has the following specific functions:

1) Training Data (80% or 70%): Focuses entirely on the development and construction of the Seq2Seq model structure.

2) Validation Data (10% or 20%): Plays a role in the process of finding the best parameters (hyperparameter tuning) and monitoring the model's generalization ability during the learning process.

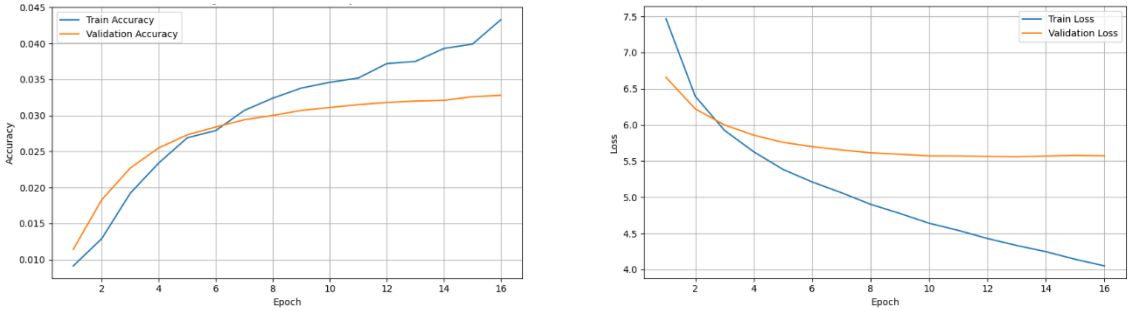
3) Testing Data (10%): Allocated specifically to test the reliability and final performance of the model when faced with new data that has never been processed before.

This model was trained using two data splits of 80:20 and 70:30, with a batch size of 8, an epoch of 100, and a learning rate of 1e-5. In the 80:20 data split, the model experienced early stopping at epoch 16, while in the 70:30 data split, the model also experienced early stopping at epoch 15. Training and validation accuracy and loss for the 80:20 data split can be seen in Fig. 5. In Fig. 5 (a) and (b) present the training and validation accuracy and loss for the 80:20 data split. The training accuracy increases steadily from approximately 0.009 to 0.043, while the validation accuracy improves from around 0.011 to 0.033, showing similar upward trends with a relatively small gap between the curves, indicating effective learning and good generalization to the validation data. At the same time, the training loss decreases substantially from about 7.466 to 4.052, and the validation loss also exhibits a consistent decline from approximately 6.657 to 5.573.

The sentence refers to the pattern seen when monitoring loss metrics during training: both loss curves (training and validation) decrease in parallel as epochs progress, indicating the model is reducing prediction errors on both the training and unseen data; this concurrent decline suggests stable convergence because there are no large fluctuations or noise that would indicate unstable training. Moreover, the absence of a sharp spike in validation loss after the initial downward phase reinforces that the model is not overfitting—meaning it does not merely memorize training patterns but maintains generalization ability on validation data. This parallel pattern is typically supported by mechanisms such as regularization, dropout, and early stopping applied during training, as well as appropriate choices of learning rate and batch size, so the optimization process reaches a good minimum without excessively fitting noise or outliers in the training set. Additionally, a consistent gap of modest size between the training and validation losses implies the model retains capacity to learn (not underfitting) while avoiding complex memorization; monitoring other metrics like validation accuracy, perplexity, or qualitative inspection of generated outputs can further confirm reliable generalization. Finally, repeating experiments with different random seeds or cross-validation and observing similar parallel trends increases confidence that the observed stable convergence is robust rather than a result of a particular initialization or dataset split. Training and validation accuracy and loss for the 70:30 data split is shown in Fig. 6. In Fig. 6 (a) and (b) show the training and validation accuracy and loss for the 70:30 data split. The training accuracy increases from approximately 0.0092 to 0.0392, while the validation accuracy improves from about 0.0088 to 0.0307, indicating a consistent learning trend across epochs. At the same time, the training loss decreases from around 7.4797 to 4.1782, and the validation loss declines from approximately 6.7176 to 5.6765. The simultaneous upward trend in accuracy and downward trend in loss indicate stable convergence and suggest that the model generalizes without overfitting. However, compared to the 80:20 split, the final accuracy and loss values remain slightly lower, which may be attributed to the reduced training data proportion. The effectiveness of the LSTM-based chatbot in maintaining contextual consistency is quantified through the BLEU score evaluation. Table III details the performance comparison across 1-gram through 4-gram levels, highlighting the model's precision in generating high-quality mental health dialogues, can be seen in Table III.

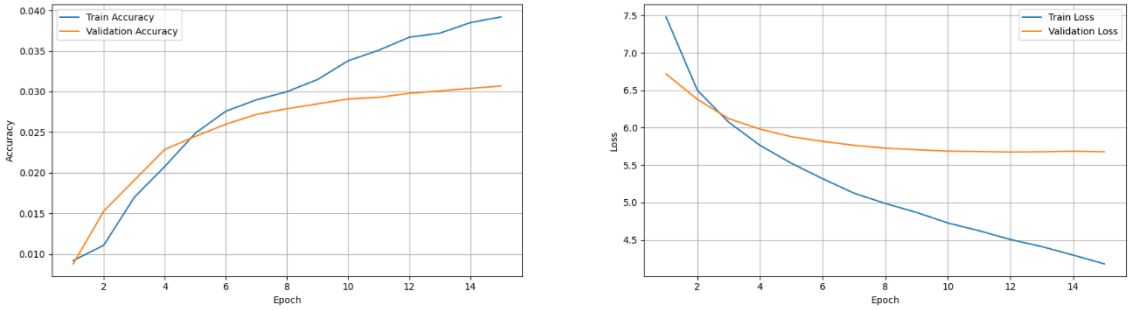
TABLE II
DISTRIBUTION OF RECORDS IN THE PUBLIC DATASETS

Split Rasio	Training	Validation	Test
80:20	2784	324	420
70:30	2387	694	447



(a) (b)

Fig. 5 (a) Training and validation accuracy and (b) training and validation loss for the 80:20 data split



(a) (b)

Fig. 6 (a) Training and validation accuracy and (b) training and validation loss for the 70:30 data split

TABLE III
PERFORMANCE COMPARISON 1-GRAM, 2-GRAM, 3-GRAM AND 4-GRAM BLEU SCORE

Spilt Rasio	Bleu 1-gram	Bleu 2-gram	Bleu 3-gram	Bleu 4-gram
80:20	0.137	0.039	0.014	0.005
70:30	0.043	0.015	0.006	0.002

Based on the BLEU metric test results in Table III, the model shows the best performance in the 80:20 data split scheme. This is evidenced by the 1-gram BLEU score of 0.137, which is significantly higher than the 70:30 scheme, which only reached 0.043. This trend of superiority is also consistent at other n-gram levels (2-gram, 3-gram, and 4-gram), indicating that a larger proportion of training data at a ratio of 80:20 contributes positively to the accuracy of the model's predictions. [44]. The model's convergence stability over a range of 15–16 epochs, with no signs of overfitting, indicates that the Seq2Seq LSTM architecture successfully learned the semantic and emotional patterns from the PSYCHIKA dataset. Achieving a BLEU-1 score of 0.137 on an 80:20 data split proves that using a larger volume of data significantly improves the model's accuracy in mapping context. Technically, this score confirms that the chatbot is capable of generating responses that are not only syntactically correct but also aligned with the user's intent and emotional state as captured in the context vector. This validates the model's role as an initial mental health assistant capable of maintaining the continuity of dialogue without losing therapeutic nuance in the context of informal Indonesian. Building upon these findings regarding data distribution, further evidence suggests that the success of a model does not rely solely on data ratios but also on the strategic application of pre-existing linguistic knowledge.

In a broader context of model enhancement, this research proves that Transfer Learning (TL) strategies based on language structure similarity are highly effective in boosting the performance of machine translation systems for low-resource languages. By utilizing the data ratio from the Tatoeba and TED2020 corpora, the process of knowledge transfer from Turkish-English to Kazakh-English has been proven to significantly improve the BLEU score, namely by +2.19 and +2.76 points above the standard model. These results are further reinforced by the integration of Part-of-Speech (POS) labels, which contribute additionally to the increase in BLEU score (+0.98), chrF (+0.23), and METEOR (+0.48). This achievement indicates that the synergy between the utilization of interlingual linguistic relationships and syntactic information enrichment can overcome the limitations of parallel data, which has been a major obstacle. Thus, this approach not only improves the technical accuracy of translation but also provides an efficient methodological solution for the development of neural translation systems for other minority languages [27].

Beyond linguistic features, the quantitative aspect of data remains paramount, this research confirms that the effectiveness of deep learning models is greatly influenced by the proportion of train/test data distribution, where an increase in training data volume is directly proportional to an increase in accuracy. Empirical data shows that optimal performance in the MobileNetV2, ResNet50v2, and VGG19 architectures is only achieved when the training data exceeds the 70% threshold. This is evident in Dataset2, where the VGG19 model scored perfect results (100%) on the accuracy, sensitivity, and specificity metrics through a 90:10 ratio. Furthermore, the 98.5% accuracy achieved by ResNet50v2 on Dataset1 with an 80:20 ratio emphasizes that a more dominant portion of training data is crucial in sharpening the model's generalization capabilities. These findings indicate that researchers should prioritize the availability of large training data to minimize the risk of underfitting and ensure that the model is able to recognize feature patterns more deeply before being tested on previously unseen data [28, 45].

IV. CONCLUSION

This research has successfully implemented the Seq2Seq LSTM architecture as a baseline model for a mental health chatbot in Indonesian. The experimental results show that the model achieved good technical stability with rapid convergence at 15–16 epochs without any signs of overfitting, as evidenced by the parallel decline in the training and validation loss curves. The use of a larger dataset (5,667 dialogue pairs) was shown to improve the model's generalization ability compared to previous studies. Quantitatively, an 80:20 data split yielded optimal performance with a BLEU-1 score of 0.137. However, the low BLEU scores at higher n-gram levels highlight the limitations of a pure LSTM architecture in generating highly fluent and emotionally complex responses. These findings confirm that while the model is capable of capturing the essence of user intent through context vectors, the integration of attention mechanisms or transformer architectures is essential for future research to enhance empathy and dialogue quality. Overall, this research provides a crucial technical foundation and dataset benchmark for the development of more advanced digital mental health first-aid systems in Indonesia.

ACKNOWLEDGEMENT

This research was supported by the Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) Kementerian Pendidikan Tinggi, Sains, dan Teknologi (Kemendikti-Sainstek) under Penelitian Disertasi Doktor (Doctoral Dissertation Research) with grant number: 284/C3/DT.05.00/PL-BARU/2026 (30 Januari 2026), 0350.12/LL5-INT/AL.04.03/2026 (28 April 2026), 038/PDD/LPPM.UAD/IV/2026 (04 Mei 2026).

REFERENCES

[1] S. T. A. Noor, S. Siddique, O. Das, S. Yeasar, and R. B. Islam, "Exploring mental health disparities in Mozambique: Depression and anxiety symptoms among reproductive-aged women using data from Mozambique Demographic and Health Survey 2022–23," *Glob. Epidemiol.*, vol. 10, no. July, p. 100223, 2025, doi: 10.1016/j.gloepi.2025.100223.

[2] N. Ekawati, I. Riadi, and H. Yuliansyah, "Machine Learning-based Chatbot Model for Healthcare Service: A Bibliometric Analysis," *Aviat. Electron. Inf. Technol. Telecommun. Electr. Control.*, vol. 7, no. 3, p. 347, 2025, doi: 10.28989/avitec.v7i3.3050.

[3] A. Heidari, M. Esmaeilzadeh, and A. H. Mohammadi, "Tourism and mental health: Two raw diamonds awaiting their brilliance through mutual shaping," *Geopsychiatry*, vol. 2, no. August, p. 100020, 2025, doi: 10.1016/j.geopsy.2025.100020.

[4] A. Muhariya, I. Riadi, and Y. Prayudi, "Cyberbullying Analysis on Instagram Using K-Means Clustering," *JUITA J. Inform.*, vol. 10, no. 2, p. 261, 2022, doi: 10.30595/juita.v10i2.14490.

[5] S. F. Handayani, D. Dairoh, and D. I. Af'idah, "Assessment of Retrieval and Generative Chatbots in Tourism Information Service," *JUITA J. Inform.*, vol. 13,

no. 1, pp. 19–27, 2025, doi: 10.30595/juita.v13i1.24182.

[6] B. Yang, Y. Ding, C. Cui, and T. Guo, "A hybrid kernel principal component analysis and long short-term memory approach for EEG signal-based fatigue evaluation," *Alexandria Eng. J.*, vol. 131, no. October, pp. 209–217, 2025, doi: 10.1016/j.aej.2025.09.066.

[7] S. U. Singh and A. S. Namin, "A Survey on Chatbots and Large Language Models: Testing and Evaluation Techniques," *Nat. Lang. Process. J.*, 2025.

[8] H. Yuliansyah, H. A. Raihan, and Murinto, "Sentiment Analysis Model for VTuber Live Stream Chat using Decision Tree and Support Vector Machine," *J. Innov. Inf. Technol. Appl.*, vol. 7, no. 2, pp. 258–268, 2025, doi: 10.35970/jinita.v7i2.2872.

[9] I. Riadi, A. Fadlil, and M. Murni, "Identifying Hate Speech in Tweets with Sentiment Analysis on Indonesian Twitter Utilizing Support Vector Machine Algorithm," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 9, no. 2, pp. 179–191, 2023, doi: 10.23917/khif.v9i2.22470.

[10] C. Lebeuf, M. A. Storey, and A. Zagalsky, "Software Bots," *IEEE Softw.*, vol. 35, no. 1, pp. 18–23, 2017, doi: 10.1109/MS.2017.4541027.

[11] A. Dwiyono, A. Abdiansah, and M. Fachrurrozi, "Analisis Perbandingan Klasifikasi Intent Chatbot Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT," *J. Inf. Syst. Res.*, vol. 6, no. 1, pp. 595–606, 2024, doi: 10.47065/josh.v6i1.6051.

[12] C. Xu, Y. Chen, Y. Tao, and W. Xie, "Deep learning-based detection of depression by fusing auditory, visual and textual clues," *J. Affect. Disord.*, vol. 391, no. July, p. 119860, 2025, doi: 10.1016/j.jad.2025.119860.

[13] A. A. Dzaky *et al.*, "Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students," *J. Appl. Informatics Comput.*, vol. 8, no. 1, pp. 13–21, 2024, doi: 10.30871/jaic.v8i1.7403.

[14] A. Ilham Ardin and A. Salam, "Pengembangan Chatbot Kesehatan Mental Berbasis Web Menggunakan Model Long Short-Term Memory (LSTM)," *Build. Informatics, Technol. Sci.*, vol. 7, no. 1, pp. 320–331, 2025, doi: 10.47065/bits.v7i1.7282.

[15] A. S. Rizky and E. Y. Hidayat, "Emotion Classification in Indonesian Text Using IndoBERT," *Comput. Eng. Appl.*, vol. 14, no. 1, pp. 2252–4274, 2025.

[16] A. Z. Kouzani and M. Nouman, "Detecting indicators of violence in digital text using deep learning," *Nat. Lang. Process. J.*, vol. 12, no. June, p. 100175, 2025, doi: 10.1016/j.nlp.2025.100175.

[17] M. Zulfadhilah, Y. Prayudi, and I. Riadi, "Cyber Profiling Using Log Analysis And K-Means Clustering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 7, pp. 430–435, 2016, doi: 10.14569/ijacsa.2016.070759.

[18] T. Y. Chen, Y. C. Chiu, N. Bi, and R. T. H. Tsai, "Multi-Modal Chatbot in Intelligent Manufacturing," *IEEE Access*, vol. 9, pp. 82118–82129, 2021, doi: 10.1109/ACCESS.2021.3083518.

[19] B. Babayigit, H. Rahmani, and M. Abubaker, "TrBot: A Turkish Deep Learning Chatbot Utilizing Seq2Seq Model," *IEEE Access*, vol. 13, no. February, pp. 49552–49566, 2025, doi: 10.1109/ACCESS.2025.3550852.

[20] Hariani and I. Riadi, "Detection of Cyberbullying on Social Media Using Data Mining Techniques," *Metall. Mater. Eng.*, vol. 31, no. 4, pp. 894–899, 2025, doi: 10.63278/1530.

[21] M. Ghaseminejad Raeini, "The evolution of language models: From N-Grams to LLMs, and beyond," *Nat. Lang. Process. J.*, vol. 12, no. June, p. 100168, 2025, doi: 10.1016/j.nlp.2025.100168.

[22] Z. Liu, Y. Quan, X. Lyu, and M. J. F. Alenazi, "Enhancing Clinical Accuracy of Medical Chatbots With Large Language Models," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 9, pp. 6395–6405, 2025, doi: 10.1109/JBHI.2024.3470323.

[23] M. Clark and S. Bailey, "Chatbots in Health Care: Connecting Patients to Information," *Can. J. Heal. Technol.*, vol. 4, no. 1, pp. 1–22, 2024, doi: 10.51731/cjht.2024.818.

[24] C. B. Nordheim, A. Følstad, and C. A. Bjørkli, "An Initial Model of Trust in Chatbots for Customer Service - Findings from a Questionnaire Study," *Interact. Comput.*, vol. 31, no. 3, pp. 317–335, 2019, doi: 10.1093/iwc/iwz022.

[25] R. Yang, M. Fu, C. Tantithamthavorn, C. Arora, L. Vandenhurk, and J. Chua, "RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice," *J. Syst. Softw.*, vol. 226, no. August 2024, p. 112436, 2025, doi: 10.1016/j.jss.2025.112436.

[26] A. A. Firdaus, A. Yudhana, I. Riadi, and Mahsun, "Indonesian presidential election sentiment: Dataset of response public before 2024," *Data Br.*, vol. 52, p. 109993, 2024, doi: 10.1016/j.dib.2023.109993.

[27] B. K. Yazar and E. Kilic, "Improving Low-Resource Kazakh-English and Turkish-English Neural Machine Translation Using Transfer Learning and Part of Speech Tags," *IEEE Access*, vol. 13, no. January, pp. 32341–32356, 2025, doi: 10.1109/ACCESS.2025.3542491.

[28] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.

[29] H. Guo and T. Erdenebold, "Factors influencing intention to adopt an AI chatbot for learning in higher education: An integrated PLS-SEM, IPMA, and ANN approach," *Comput. Educ. Artif. Intell.*, vol. 9, no. 51, p. 100477, 2025, doi: 10.1016/j.caeai.2025.100477.

[30] G. A. Santos, G. G. de Andrade, G. R. S. Silva, F. C. M. Duarte, J. P. J. Da Costa, and R. T. de Sousa, "A Conversation-Driven Approach for Chatbot

Management," *IEEE Access*, vol. 10, pp. 8474–8486, 2022, doi: 10.1109/ACCESS.2022.3143323.

[31] A. R. Mandli, S. Rajput, and T. Sharma, "COMET: Generating commit messages using delta graph context representation," *J. Syst. Softw.*, vol. 222, no. January 2024, p. 112307, 2025, doi: 10.1016/j.jss.2024.112307.

[32] M. Ahmed, H. U. Khan, and E. U. Munir, "Conversational AI: An Explication of Few-Shot Learning Problem in Transformers-Based Chatbot Systems," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 1888–1906, 2024, doi: 10.1109/TCSS.2023.3281492.

[33] Q. Liu, J. Huang, L. Wu, K. Zhu, and S. Ba, "CBET: design and evaluation of a domain-specific chatbot for mobile learning," *Univers. Access Inf. Soc.*, vol. 19, no. 3, pp. 655–673, 2020, doi: 10.1007/s10209-019-00666-x.

[34] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," *Telemat. Informatics Reports*, vol. 16, no. April, p. 100173, 2024, doi: 10.1016/j.teler.2024.100173.

[35] A. Ohashi and R. Higashinaka, "Adapting language generation to dialogue environments and users for task-oriented dialogue systems," *Nat. Lang. Process. J.*, vol. 11, no. March, p. 100153, 2025, doi: 10.1016/j.nlp.2025.100153.

[36] L. C. Navarro, S. S. Mucciaccia, F. Mutz, T. M. Paixäo, and C. Badue, "Exploring question answering: metric analysis and evaluation framework for enhanced interpretability," *Neural Comput. Appl.*, vol. 38, no. 9, 2026, doi: 10.1007/s00521-026-12053-8.

[37] Y. Zhu, S. Lu, L. Zheng, J. Guo, and W. Zhang, *Texygen: A Benchmarking Platform for Text Generation Models*. 2018.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Ann. Phys.*, vol. 371, no. 23, pp. 437–461, 1922, doi: 10.1002/andp.19223712302.

[39] D. Escobar-Grisales, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Evaluation of effectiveness in conversations between humans and chatbots using parallel convolutional neural networks with multiple temporal resolutions," *Multimed. Tools Appl.*, vol. 83, no. 2, pp. 5473–5492, 2024, doi: 10.1007/s11042-023-14896-y.

[40] C. Luna-Jiménez, M. Gil-Martín, L. F. D'Haro, F. Fernández-Martínez, and R. San-Segundo, "Evaluating emotional and subjective responses in synthetic art-related dialogues: A multi-stage framework with large language models," *Expert Syst. Appl.*, vol. 255, no. PB, p. 124524, 2024, doi: 10.1016/j.eswa.2024.124524.

[41] M. Adam, M. Wessel, and A. Benlian, "AI-based chatbots in customer service and their effects on user compliance," *Electron. Mark.*, vol. 31, no. 2, pp. 427–445, 2021, doi: 10.1007/s12525-020-00414-7.

[42] B. O. S. Miranda, H. Yuliansyah, and M. K. Biddinika, "Machine Translation Indonesian Bengkulu Malay Using Neural Machine Translation-LSTM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 18, no. 3, 2024, doi: 10.22146/ijccs.98384.

[43] F. Benedict, C. V. Mramba, S. Kaaya, J. Kimaro, J. N. Baumgartner, and M. Bachmann, "Developing a regional mental health plan for Dar es Salaam, Tanzania: Results from a situational analysis, qualitative inquiry, and stakeholder engagement process," *SSM - Ment. Heal.*, vol. 8, no. September, p. 100532, 2025, doi: 10.1016/j.ssmmh.2025.100532.

[44] A. Jha, A. Bhatia, K. Tiwari, and H. M. Pandey, "Hierarchical Bayesian Deep Learning for return on advertising spend prediction: A probabilistic approach to e-commerce advertising," *Eng. Appl. Artif. Intell.*, vol. 164, no. PA, p. 113200, 2026, doi: 10.1016/j.engappai.2025.113200.

[45] P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A Mental Health Chatbot with Cognitive Skills for Personalised Behavioural Activation and Remote Health Monitoring," *Sensors*, vol. 22, no. 10, 2022, doi: 10.3390/s22103653.