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

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

Mapping Facial Expressions Based on Text for Virtual Counseling Chatbot Using IndoBERT Model

Rifki Padilah*[1], Rifki Wijaya[2], Shaufiah[3]

[1,3]Informatics, School of Computing, Telkom University, Indonesia
[2] Center of Excellence of Artificial Intelligence for Learning and Optimization (CoE AILO), Telkom University, Indonesia

Email: [1]rifkipdh@student.telkomuniversity.ac.id

Received : Jul 5, 2025; Revised : Aug 4, 2025; Accepted : Aug 6, 2025; Published : Oct 22, 2025

Abstract

Early marriage in Lombok remains a serious issue, with a prevalence rate of 16.59% in 2021, the second highest in Indonesia. Limited access to counseling services, especially in rural areas, poses a significant prevention challenge. This study developed a virtual counseling chatbot system capable of mapping text-based emotions to facial expressions to improve the effectiveness of counseling for early marriage prevention. The methodology involved training an IndoBERT model on a synthetic dataset to analyze conversation texts. The model was designed to classify user input into five functional emotion categories: Enthusiasm, Gentleness, Analytical, Inspirational, and Cautionary. Performance evaluation revealed that the IndoBERT model achieved an outstanding accuracy of 94% in its final phase. This result significantly surpassed other models evaluated, such as CNN (71.6%) and KNN (79%), confirming the superiority of the chosen approach The study concludes that the high-accuracy IndoBERT model is a robust foundation for empathetic virtual agents. This research provides a significant contribution to the fields of Affective Computing and Human-Computer Interaction by demonstrating an effective framework for mapping nuanced, functional emotions from Indonesian text to facial expressions. The proposed system not only offers a scalable technological solution for mental health challenges like early marriage prevention but also highlights the impact of advanced, context-aware NLP models in creating more human-like and empathetic user interactions.

Keywords: *Chatbot, Early Marriage, Emotion Recognition, Facial Expression Mapping, IndoBERT, Virtual Counseling.*

1. INTRODUCTION

Early marriage persists as a significant socio-cultural issue in Lombok, Indonesia, particularly in the contemporary digital era. With a prevalence rate of 16.59% in 2021, the region ranks as having the second-highest incidence of early marriage nationwide. This practice is deeply entrenched in the cultural and traditional norms of the Sasak community and is often legitimized by influential figures, including religious leaders and educators, who view it as a preventative measure against premarital sexual behavior [1]. The primary drivers are a confluence of cultural, economic, and educational factors, which frequently overshadow the long-term adverse impacts on the affected girls and children, especially concerning their reproductive health, educational attainment, and socio-economic well-being.

Counseling services represent a critical intervention strategy for preventing early marriage, particularly for adolescents confronting social and cultural pressures to marry young. Through counseling, adolescents can acquire a comprehensive understanding of the negative consequences of early marriage, thereby facilitating more informed decision-making regarding their futures [2]. However, the efficacy of traditional counseling services is constrained by several limitations, especially in rural areas. Access is impeded by factors such as inadequate infrastructure, prohibitive costs, limited service hours, and inequitable outreach, resulting in a significant portion of the youth in remote regions

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

being deprived of essential education and guidance [3]. These challenges are compounded by a significant treatment gap, social stigma, and a lack of mental health literacy, which are well-documented barriers to accessing mental health services across Indonesia [4]

As an alternative solution, this paper proposes the use of AI-driven counseling chatbots to deliver relevant and responsive emotional support via text-based conversations. These systems operate by analyzing user conversation texts through techniques like sentiment analysis and Natural Language Processing (NLP) to identify emotional states and provide appropriate, empathetic responses [5], [6]. The principal advantages of this modality include 24/7 accessibility, user anonymity which allows individuals to communicate without shame or fear and ease of use, enabling access to services from any location at any time [7]. Chatbots capable of demonstrating empathy can significantly enhance user satisfaction and increase the likelihood of continued use [8]. National and international studies have substantiated the effectiveness of chatbots in providing psychological support, particularly for individuals who face difficulties in expressing emotions directly or who encounter barriers to accessing traditional counseling services [9].

A novel enhancement to these systems is the integration of facial expression mapping based on conversation text analysis, a concept designed to improve the quality of virtual counseling interactions. Leveraging multimodal technology, the system employs language models such as IndoBERT, which is specifically optimized to comprehend the context and nuances of the Indonesian language. This model identifies emotions within the text and maps the analytical results to corresponding virtual facial expressions, aiming to render the counseling interaction more empathetic and anthropomorphic [10], [11].

The development of the IndoBERT model for this purpose was conducted by training it on a synthetic dataset of 2,500 data points, with 500 points allocated to each of five emotion categories: Enthusiasm, Gentleness, Analytical, Inspirational, and Cautionary. The model's performance was quantitatively assessed using standard metrics, including accuracy, precision, recall, and F1-score, which are conventional benchmarks in AI model evaluation for emotion classification and analysis tasks [12]. Related research confirms that the application of IndoBERT to Indonesian language datasets achieves high accuracy and performance in sentiment and emotion analysis [13].

A comparative analysis indicates that IndoBERT exhibits superior performance metrics over other methods. For instance, while a CNN-based approach achieved 71.56% accuracy for classifying six standard emotion labels, it required complex adjustments [14]. Similarly, a KNN model reached 79% accuracy but was limited to a binary sentiment task (positive/negative) and struggled in multi-class contexts [15]. This highlights a key limitation of prior work: a focus on either standard emotion sets or overly simplistic sentiment analysis.

However, a gap remains in the existing literature, particularly within the Indonesian context. While many studies focus on standard sentiment analysis (positive, negative, neutral), few have explored a more nuanced, function-based emotional classification scheme tailored for specific domains like counseling. Furthermore, the integration of such advanced NLP outputs with multimodal feedback, such as facial expressions, to create truly empathetic virtual agents is still an emerging area of research in Indonesia. This study aims to address this gap directly.

The primary novelty of this research, therefore, lies in three key areas. First, the development and application of a functional five-emotion classification framework ('Enthusiasm', 'Gentleness', etc.) that moves beyond simple sentiment analysis to better suit the counseling domain. Second, the implementation of a strategic two-phase training methodology for the IndoBERT model, using both balanced and imbalanced synthetic datasets to achieve high accuracy and robustness. Third, the creation of a direct mapping framework from these classified emotions to corresponding virtual facial

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

expressions, providing a foundational step towards more empathetic and human-like chatbot interactions in the Indonesian language.

This research contributes to improving accessibility to psychological counseling services in Lombok, offering a scalable and effective technological solution to prevent early marriage. By bridging the gap between counseling needs and available resources, this innovative effort helps adolescents develop emotional resilience and informed decision-making skills, especially in underserved communities [16],[17],[18].

2. METHOD

In this phase of the system design, we focus on creating a plan to analyze the implementation of an emotion detection system using the IndoBERT model. This plan maps out the steps needed to build and evaluate the system. The research methodology is structured into several distinct phases, as illustrated in the block diagram in Figure 1. This systematic approach encompasses key stages: data collection, preprocessing, implementation model, mapping facial expression, and evaluation. Each stage is integral to the development of a system capable of accurately detecting emotions from text and mapping them to corresponding facial expressions.

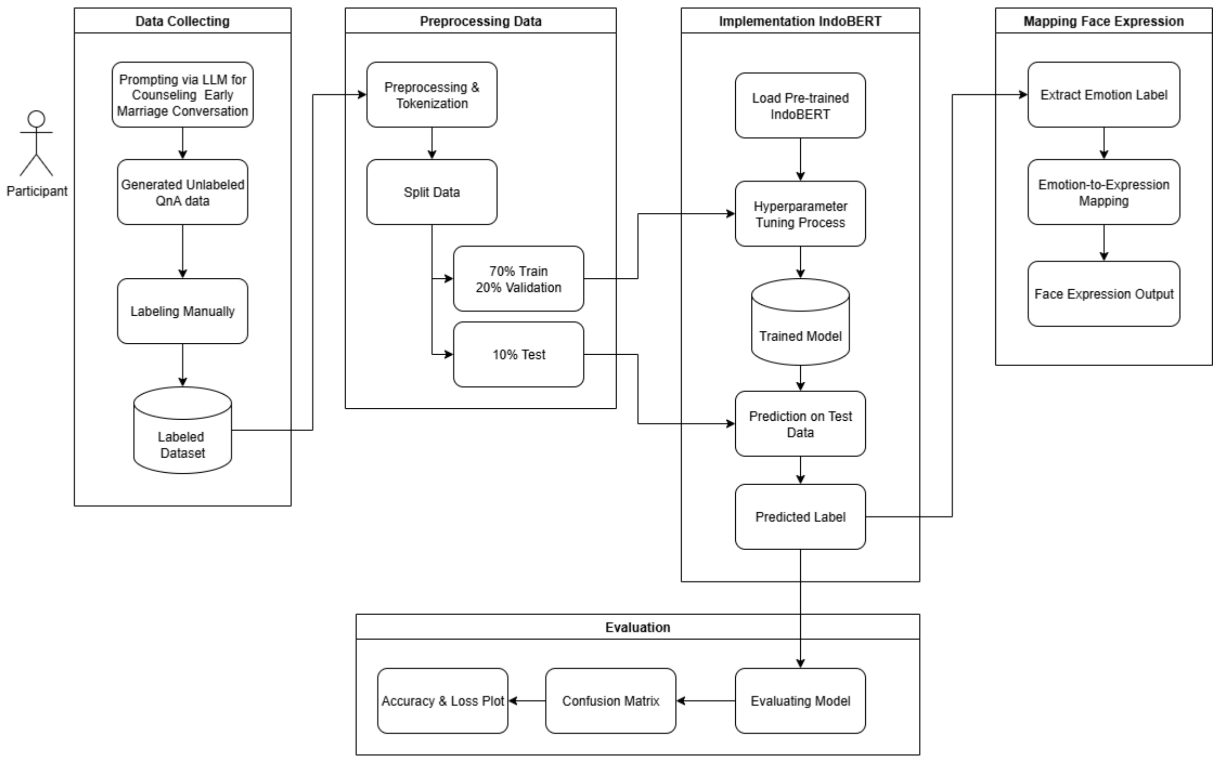

Figure 1. Data Collection, Preprocessing, Implementation, Mapping Face Expression and Evaluation

2.1. Data Collection

In this study, the “Data Collecting” part focuses on using a large language model (LLM) to generate conversations related to early marriage counseling. You can see how it works in the flowchart shown in Figure 1. The process involves a few key steps: generating sample conversations using the LLM, manually reviewing and labeling the responses, and then building a clean, labeled dataset for the next phase of the research.

The process commenced by providing the LLM with a series of prompts or scenarios to generate question-and-answer (QnA) conversations between an adolescent and a counselor. An example of a

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

prompt used was: "Simulate a conversation between a teenager who is hesitant about early marriage and a counselor providing inspirational and reassuring advice."

The resulting QnA data was initially unlabeled. Therefore, manual annotation was performed by the researchers on each counselor's response. Each response was classified into one of five functional emotion categories based on its communicative purpose within the counseling context. The specific criteria for each category Enthusiasm, Gentleness, Analytical, Inspirational, and Cautionary are detailed in Table 3.

This meticulous manual annotation process was crucial to ensure the final dataset was accurate, consistent, and truly reflective of the functional emotional nuances in counseling conversations, forming the basis for the subsequent model training phase.

2.2. Preprocessing

Before the IndoBERT model can be used in this study, the raw text goes through a few important cleanup steps. These steps known as preprocessing are essential because the quality of this early work has a big impact on how well the model performs later on. If the data is handled properly from the start, the model is more likely to give accurate and meaningful results.The entire preprocessing workflow is divided into two main stages: (1) Text Cleaning and Normalization, and (2) Tokenization and Input Formatting.

The first stage, Text Cleaning and Normalization, aims to standardize the text and reduce noise from the raw conversational data. The specific steps performed in this stage are illustrated in Figure 2.

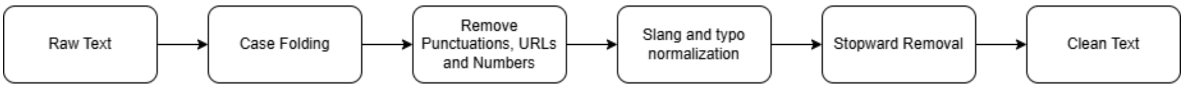

Figure 2. The preprocessing flow

The process begins with Case Folding to convert all text to lowercase, ensuring uniformity. Following this, Remove Punctuations, URLs, and Numbers is applied to eliminate characters that generally do not contribute to the emotional context. A critical step for conversational data is Slang and Typo Normalization, where non-standard words are converted to their formal equivalents using a predefined dictionary. Finally, Stopword Removal is performed to filter out common words that carry little semantic weight. It is noteworthy that traditional steps like stemming or lemmatization are deliberately omitted, as advanced models like IndoBERT can interpret different word forms through their contextual embeddings, making such normalization unnecessary. The output of this stage is clean, normalized text.

In the second stage, Tokenization and Input Formatting, the clean text is converted into a numerical format that the IndoBERT model can understand. The initial data preprocessing step involves tokenization using the IndoBERT native WordPiece tokenizer, which segments text into sub-word units. Special tokens ([CLS], [SEP]) are appended to each sequence before they are converted into numerical sequences, or input_ids. To accommodate variable sequence lengths, padding and truncation are applied to enforce a uniform length across each batch. Concurrently, an attention mask is generated to distinguish substantive tokens from padding, thereby directing the model's focus to meaningful content.

Following this formatting, the dataset is partitioned into training (70%), validation (20%), and test (10%) subsets. This division is a standard convention in modern Natural Language Processing (NLP) [19] and is critical for robust model evaluation and the mitigation of overfitting. The validation set is specifically utilized for hyperparameter tuning, while the test set is reserved for the final, unbiased assessment of model performance. This comprehensive preprocessing stage is fundamental for enhancing data quality and preparing the data for optimal processing by the IndoBERT model.

P-ISSN: 2723-3863
https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

2.3. IndoBERT

The model implementation phase is a central component of this research, focusing on the classification of early marriage counseling texts. This task is accomplished by employing a transfer learning architecture based on the IndoBERT model, which is trained to predict emotion labels from the preprocessed data.

The process commences with loading the pre-trained IndoBERT model, a state-of-the-art language representation model selected for its specific design for the Indonesian language [20]. Its efficacy stems from its pre-training on an extensive and diverse corpus of Indonesian data, which provides it with a sophisticated understanding of the language's context, semantics, and syntax. By leveraging transfer learning, this research taps into the existing knowledge of IndoBERT, which is far more efficient and effective than training a model from scratch, especially given the limited data available in a specialized dataset like counseling conversations [20]. The model loaded is the basic IndoBERT architecture, indobenchmark/indobert-base-p1, with weights from its pre-training, ready to be fine-tuned for the emotion classification task in this study.

Once the pre-trained model is loaded, the next step is fine-tuning, where the model's parameters are re-trained using a more specific dataset. The objective of fine-tuning is to adapt the model's generalized linguistic understanding to the specific task of emotion classification within the counseling domain. For this study, the model is adapted with 70% training data and 20% validation data. A crucial part of this process is hyperparameter tuning, where careful selection of each parameter ensures that the model learns effectively. As shown in Table 1, the study uses the AdamW optimizer, which is reliable in preventing overfitting in Transformer models. The learning rate is set to 2e-6, a small rate that ensures stable weight adjustments without disturbing the knowledge acquired during pre-training. The model is trained for 5 epochs with a batch size of 8, a configuration that balances validation performance with training time.

Table 1 Hyperparameter Settings

Hyperparameter	Value
Base Model	indobenchmark/indobert-base-p1
Optimizer	AdamW
Learning Rate	2e-6
Number of Epochs	5
Batch Size	16
Early Stopping	Patience = 1 epoch

The implementation protocol leverages established transfer learning principles, the efficacy of which has been demonstrated in recent studies using fine-tuned IndoBERT for Indonesian emotion classification [21], [22]. These studies confirm that the model can achieve robust generalization, even on specialized or low-resource datasets. Accordingly, following the fine-tuning stage, the model's generalization performance is formally evaluated on a previously unseen hold-out test set (10% of the data). In this inference phase, the model generates emotion label predictions for the test inputs, which are then compared with the corresponding ground-truth labels to provide a quantitative assessment of classification accuracy.

2.4. Evaluation Phase

The performance of the fine-tuned IndoBERT model on the emotion classification task is evaluated using four standard metrics: accuracy, precision, recall, and F1-score [23], [24]. These metrics, whose formulas are defined in Equations (1) through (4), collectively provide a comprehensive assessment of the model's effectiveness in distinguishing between the designated emotion categories.

2.4.1. Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It provides a summary of prediction results by comparing the labels predicted by the model against the actual (ground truth) labels. This allows for a detailed analysis of the model's ability to differentiate between classes and helps in identifying specific types of misclassification errors.

Table 2. Explaining the confusion matrix

		Actual Value	
		Positive	Negative
Predictive Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

As illustrated in Table 2, a confusion matrix is composed of four fundamental components: True Positive (TP), denoting the number of instances correctly identified by the model as belonging to the positive class; True Negative (TN), indicating the number of instances accurately classified as negative; False Positive (FP), referring to negative instances that the model incorrectly classified as positive commonly associated with Type I error; and False Negative (FN), which captures positive instances that were mistakenly identified as negative, corresponding to a Type II error.

2.4.2. Accuracy

Accuracy measures how often the model's predictions are correct compared to the total number of predictions it makes. It gives a clear snapshot of the model's overall performance, and it's particularly useful when the class distribution in the data is balanced. This metric is simple but effective, helping to quickly gauge how well the model is doing in general [25], [26]. he formula for accuracy is expressed in Equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2.4.3. Precision

Precision is a performance metric that quantifies the proportion of true positive instances among all instances the model has classified as positive[23]. This measure becomes particularly critical in applications where the consequences of a false positive error are substantial. For example, in a context where cautionary messages are misclassified as inspirational, the imperative is to reduce such errors. Accordingly, precision serves as a direct indicator of a model's ability to avoid these specific types of misclassifications. The metric is formally defined in Equation (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

2.4.4. Recall

A model's effectiveness can also be measured by its ability to capture all true positive cases, a concept quantified by the recall metric [23]. This measure is of paramount importance in contexts where

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

overlooking an instance carries significant risk, such as in early marriage counseling. In such scenarios, a failure to register important emotional signals (a false negative) can critically undermine the system's perceived empathy and utility. Consequently, recall is employed to assess the model's success in minimizing these omissions, thereby ensuring a thorough detection of all relevant data points. Its formal definition is presented in Equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

2.4.5. F1-Score

The F1-score provides a method to reconcile the inherent trade-off between precision and recall by computing their harmonic mean[23], [24]. This single, integrated metric offers a more holistic assessment of model efficacy, which is particularly valuable when dealing with imbalanced datasets or when the costs of both false positives and false negatives are substantial. By its formulation, the F1-score ensures that strong performance in one metric does not obscure poor performance in the other, thus yielding a more balanced evaluation of the model's overall performance. The metric is mathematically defined in Equation (4).

$$F1 - Score \; = 2\, x \frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

2.4.6. Accuracy and Loss Plot Curve

The evolution of a model's performance during training is critically assessed using accuracy and loss plots, which respectively quantify the rates of correct classification and prediction error over epochs. A decreasing loss function indicates that the model's parameters are successfully converging towards an optimal data representation through methods like gradient descent. Simultaneously, an increasing accuracy curve demonstrates a growing proficiency in mapping inputs to correct outputs.

A key aspect of this analysis is the comparison between training and validation performance to evaluate generalization. While concurrent improvement in both training and validation metrics signifies effective learning, a notable divergence characterized by high training accuracy and low validation accuracy signals the onset of overfitting. In this state, the model has memorized noise and artifacts from the training set at the expense of its ability to perform on new data.

Therefore, the combined interpretation of these plots is a crucial diagnostic practice. Suboptimal learning patterns, such as stagnating loss and accuracy, can point to underlying problems like learning rate issues or significant class imbalance. The latter condition is particularly important, as accuracy alone can be a deceptive indicator of performance on imbalanced data without a corresponding analysis of the loss function[27], [28]. By carefully interpreting these visualizations, practitioners can make informed decisions regarding model adjustments, such as implementing regularization methods, to ensure the final model is both accurate and robustly generalizable.

2.5. Mapping Emotion Label to Face Expression

In this study, the journey of data doesn't end when the IndoBERT model finishes classifying emotions. There's one final and essential step: "Face Expression Mapping." This step acts as a bridge, turning abstract text analysis into a visual representation that people can directly feel and understand, giving life to the data.

Once the IndoBERT model has completed its task, the process begins. Take, for example, a sentence from a counseling session about early marriage: "I'm sure we can get through this together." After analysis, it's labeled as Inspirational. This label is then extracted and serves as the key to the next step, unlocking the visual dictionary within the system.

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

The mapping process itself is where the magic happens. The system uses predefined rules to translate each emotion label into a specific facial expression. For a clearer picture, these mappings are summarized in Table 3.

Table 3. Guide to textualize emotion to Face Expression.

Emotion	Face Expression	Description
Enthusiasm	Displaying a wide, open smile, with eyes sparkling with energy.	Create a positive and enthusiastic atmosphere, showing optimism and active support for the client's thoughts.
Gentleness	Showing a sincere but subtle smile, with warm and soothing eyes.	Establish a sense of security and show empathy. This expression communicates that the psychologist is listening attentively.
Analytical	Has a neutral but focused facial expression, with a slightly furrowed brow.	Indicates complete concentration and deep thought, as if considering information objectively.
Inspirational	Give a confident smile accompanied by a straightforward and reassuring gaze.	Show confidence and hope to empower clients and encourage them to see their future potential.
Cautionary	Show a serious expression, with furrowed eyebrows and a sharp gaze.	Communicate concern or seriousness about a risk in a firm but caring manner.

At its core, the entire process of mapping facial expressions aims to humanize interactions with technology. By transforming the analytical outputs of artificial intelligence into empathetic and familiar visual cues, this approach enhances user experiences. For instance, when communicating with a virtual counselor, these visual cues make the interaction feel more alive, responsive, and genuinely meaningful. This transformation bridges the gap between raw data and human connection, ensuring that technology resonates on an emotional level.

3. RESULT

This section presents the empirical findings of the study, structured to mirror the methodological stages outlined in Chapter 2. The presentation begins with an overview of the dataset, followed by the results from the two sequential training phases of the IndoBERT model, and concludes with the implementation of the facial expression mapping framework. The model was subjected to a sequential fine-tuning protocol, initially using a 2,500-sample dataset, followed by a more extensive 9,000-sample dataset. Its performance was quantified using a suite of standard classification metrics: accuracy, precision, recall, and F1-score. This statistical assessment is complemented by visualizations of the training dynamics, including accuracy and loss plots, and confusion matrices that provide a detailed breakdown of the model's per-class classification accuracy.

Building upon these classification capabilities, this chapter also puts forth a novel schema to translate the model's categorical outputs into corresponding facial expressions. The framework’s key innovation lies in its direct semantic mapping, where each emotion label is programmatically linked to a facial configuration that represents its intrinsic meaning. This proposed methodology establishes a foundational basis for developing visually expressive virtual agents, enabling them to generate more empathetic and nuanced interactions within a counseling context.

3.1. Dataset Overview

A specialized corpus was constructed for this study to reflect the dialogic patterns of early marriage counseling. Structurally, the dataset comprises simulated question-and-answer (QnA)

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

interactions between adolescents and psychological counselors. The development of this corpus adhered to a structured protocol involving three key phases: an initial content generation stage, a semi-automated procedure for labeling the data, and a final preprocessing step designed to optimize the text for subsequent model training.

The construction of the dataset began with the generation of raw conversational data using a Large Language Model (LLM) to produce unlabeled QnA pairs. To establish a high-quality foundation, an initial set of 2,500 samples was manually annotated. These samples were distributed evenly across five predefined emotional categories: ‘Enthusiasm’, ‘Gentleness, ‘Analytical’, ‘Inspirational’, and ‘Cautionary’, with each class containing exactly 500 instances. Subsequently, an IndoBERT model was fine-tuned on this manually labeled subset and then utilized to predict labels for the remaining 9,000 QnA pairs. This semi-supervised approach effectively yielded a total of 11,500 labeled data points for the experiment. Table 4 presents illustrative examples from the annotated dataset. This table demonstrates how complex counselor responses were assigned a single functional emotion label, which formed the basis for the model's classification task.

Table 4. Example of Q&A in English

Input	Response	Label
I feel very confident that I can handle married life, even at a young age!	That's an amazing spirit! It's true, you can face it if you feel ready. However, remember that there are many things you can achieve before taking that big step. Focus on yourself first to build a strong foundation!	Enthusiasm
I am concerned that if I delay marriage, my standards will become higher and it will be difficult to find a partner.	As you grow older and gain more experience, it is natural for your standards to evolve. This does not mean that you become more difficult, but rather more selective for your own good. Improving your quality will attract a suitable partner.	Gentleness
People say that getting married young makes life stable. Is that true?	A stable life comes from readiness and planning, not just from getting married young. If you're not ready, marriage can actually add to your burdens. Focus on building your emotional and financial foundations first!	Analytical
	Postponing marriage gives you the opportunity to focus on your education and personal development. By improving your skills and knowledge, you will be better prepared for marriage and achieve your life goals more effectively.	Inspirational
I feel that getting married young could give me stability, but I am concerned about my desire to change over time.	Be careful, as time goes by, we often grow and change. Getting married young can be challenging if the couple is not ready to deal with each other's changes over time, which can affect the relationship.	Caution

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

Table 5. Example of Q&A in Bahasa Indonesia

Input	Response	Label
Aku merasa sangat yakin bahwa aku bisa menghadapi kehidupan pernikahan, bahkan di usia muda!	Itu semangat yang luar biasa! Memang benar, kamu bisa menghadapinya jika kamu merasa siap. Namun, ingatlah bahwa ada banyak hal yang bisa kamu capai sebelum kamu mengambil langkah besar itu. Fokus pada dirimu terlebih dahulu untuk membangun fondasi yang kuat!	Enthusiasm
Saya khawatir jika menunda menikah, standar saya akan semakin tinggi dan sulit dapat jodoh.	Seiring bertambahnya usia dan pengalaman, wajar jika standarmu berkembang. Ini bukan berarti jadi lebih sulit, tapi lebih selektif untuk kebaikanmu. Kualitas diri yang meningkat justru akan menarik pasangan yang sepadan	Gentleness
Orang bilang nikah muda bikin hidup stabil. Bener nggak?	Hidup yang stabil datang dari kesiapan dan perencanaan, bukan cuma dari nikah muda. Kalau belum siap, nikah malah bisa tambah beban. Fokus dulu bangun pondasi emosional dan keuangan, ya!	Analytical
Kenapa saya harus menunda pernikahan dan lebih fokus pada pendidikan dan pengembangan diri?	Menunda pernikahan memberi kamu kesempatan untuk fokus pada pendidikan dan pengembangan diri. Dengan meningkatkan keterampilan dan pengetahuan, kamu akan lebih siap menghadapi pernikahan dan mencapai tujuan hidup secara lebih efektif.	Inspirational
Saya merasa menikah muda bisa memberi saya kestabilan, tetapi saya khawatir tentang keinginan untuk berubah seiring waktu.	Hati-hati, seiring berjalannya waktu, kita sering kali berkembang dan berubah. Menikah muda bisa menjadi tantangan jika pasangan tidak siap untuk menghadapi perubahan masing-masing seiring waktu, yang dapat mempengaruhi hubungan.	Caution

The examples in Table 4 and Table 5 illustrate the practical application of the defined emotional categories to the conversational data. Each counselor's response was carefully evaluated to determine its primary emotional function, ensuring that the manual labels were consistent and reliable. This meticulous annotation process was crucial for establishing a robust ground-truth dataset. The balanced distribution of this initial dataset is detailed in Table 6. This table confirms that each of the five emotion categories contained an equal number of samples (500), ensuring that the model in the first phase would not be biased toward any majority class.

Table 6. Label Distribution of the 2,500 Manually Annotated Samples

Emotion Label	Number of Samples
Enthusiasm	500
Gentleness	500
Analytical	500
Inspirational	500
Cautionary	500
Total	2.500

Following the creation of the initial set, the dataset was expanded by an additional 9,000 QnA pairs. An IndoBERT model, fine-tuned on the initial 2,500 balanced samples, was used to automatically predict labels for this larger, unlabeled portion. The distribution of the predicted labels for these 11,500 samples is detailed in Table 7. In contrast to the initial dataset, Table 7 reveals a significant class imbalance, with the 'Analytical' and 'Inspirational' categories being the most dominant. This imbalanced distribution more closely reflects a real-world data scenario and was used to test the model's robustness in the second training phase.

Table 7. Distribution of Predicted Labels for the 11,500 Automatically Labeled Samples

Emotion Label	Number of Samples
Analytical	3.485
Inspirational	3.485
Gentleness	2.358
Enthusiasm	1.153
Cautionary	818
Total	11.500

Prior to the training phase, the entire text corpus underwent an essential preprocessing stage, which included text normalization and tokenization. Once prepared, the complete dataset was partitioned into training, validation, and testing subsets using Stratified Sampling. The final allocation was 70% for the training set, 20% for the validation set, and 10% for the test set.

3.2. Training Phase 1: Fine-Tuning on 2,500 Labeled Data

In the initial training phase, the IndoBERT model was fine-tuned using a focused, foundational dataset. This dataset consists of 2,500 counseling dialogue samples that were manually labeled with one of five function-based emotion labels: ‘Enthusiasm’, ‘Gentleness’, ‘Analytical’, ‘Inspirational’, and ‘Cautionary’. Each category was represented by exactly 500 samples, creating a balanced dataset that allowed the model to learn without bias toward any single emotion class.

The fine-tuning process employed a specific set of hyperparameters, chosen to balance computational efficiency with model performance. As previously detailed in Table 1, the training was conducted for a maximum of 5 epochs, utilizing the AdamW optimizer with a learning rate of 2e-6 and a batch size of 16. An early stopping mechanism with a patience of 1 epochs was implemented to retain the best-performing model checkpoint based on validation loss.

The learning progress throughout the training phase is visualized in Figure 3, which plots the model's accuracy and loss on both the training and validation sets. As shown in the plots, the accuracy curve (left) for both the training and validation data exhibits a consistent and convergent increase, indicating that the model learned effectively without significant overfitting. Similarly, the loss plot (right) shows a steady decrease for both sets, confirming a healthy training process.

As shown in the plots, the model demonstrated consistent improvements, with performance on the validation set peaking at the best epoch. At this point, the validation accuracy is high, and the validation loss is at its minimum before starting to plateau. This trend indicates that model represents the optimal balance between learning and generalization. Therefore, this optimal checkpoint was selected as the final trained model for subsequent inference tasks. The detailed performance metrics from this best-performing epoch are summarized in the classification report in Table 8. This report

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

shows a high overall accuracy of 92% and strong F1-scores across all categories, validating the model's ability to distinguish between the five emotion labels on a balanced dataset.

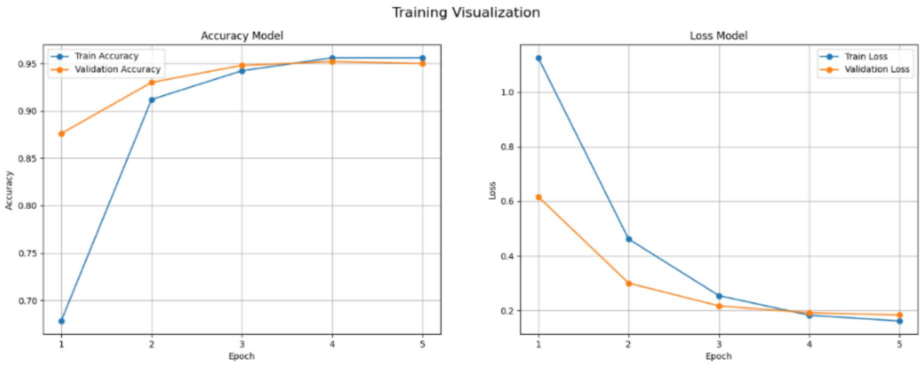

Figure 3. Plot Accuracy and Loss Model

Table 8. Classification Report from the Best Performing Epoch

Emotion Label	Precision	Recall	F1-Score	Support
Enthusiasm	0.96	1.00	0.98	50
Gentleness	0.90	0.90	0.90	50
Analytical	0.95	0.82	0.88	50
Inspirational	0.94	0.92	0.93	50
Cautionary	0.88	0.98	0.92	50

The results from the best epoch show a high overall accuracy of 92% and consistently strong F1-scores, ranging from 0.88 to 0.98 across all categories. To further analyze the model's classification performance on a class-by-class basis, a confusion matrix is presented in Figure 4. This matrix reveals a strong performance, characterized by a dominant diagonal that indicates a high number of correct predictions across all categories. Specifically, the 'Enthusiasm' category was classified with perfect accuracy, while minor misclassifications occurred for the 'Analytical' category, suggesting a linguistic overlap with the 'Cautionary' and 'Gentleness' categories.

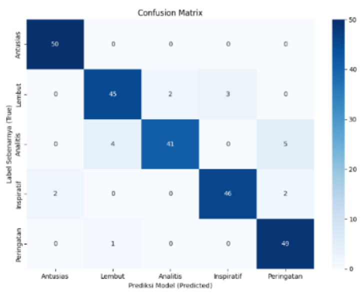

Figure 4. Confusion Matrix Plot for Phase 1

For a more in-depth analysis of the model's classification performance, a confusion matrix is presented in Figure 4. The matrix reveals a strong performance, characterized by a dominant diagonal which indicates a high number of correct predictions across all categories.

Specifically, the 'Enthusiastic' category was classified with perfect accuracy, while the 'Cautionary' and 'Inspirational' classes also demonstrated near-perfect performance. Minor misclassifications were observed, most notably with the 'Analytical' class, which was occasionally mispredicted as 'Cautionary' (5 instances) and 'Gentle' (4 instances), suggesting an overlap in linguistic

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

features. Nevertheless, the overall results confirm the model's robust capability to distinguish between most of the complex emotional nuances.

3.3. Inference and Data Expansion

Following the initial fine-tuning phase, the best-performing IndoBERT model was utilized to perform inference on an expanded, unlabeled dataset of 9,000 new question-and-answer pairs. The primary objective of this stage was to automatically assign emotion labels to this larger corpus, thereby creating a comprehensive, fully annotated dataset for a second, more advanced training phase. To ensure the integrity of the model, the inference process was conducted strictly in evaluation mode, meaning no model weights were updated during prediction.

The distribution of the predicted emotion labels resulting from this automated process is summarized in Table 9. A notable class imbalance emerged, with the 'Inspirational' and 'Analytical' categories being the most prevalent, collectively accounting for nearly 62% of the newly labeled data.

Table 9. Distribution of Predicted Labels for the 11,500 Automatically Labeled Samples

Emotion Label	Number of Samples	Percentage (%)
Inspirational	3.686	32.05
Analytical	3.485	30.30
Gentleness	2.358	20.50
Enthusiasm	1.153	10.03
Cautionary	818	7.11

This resulting skew is likely a reflection of the patterns learned by the model from the initial balanced dataset, causing it to exhibit a predictive tendency towards certain emotional tones. No post-inference balancing techniques were applied to this dataset. The intention was to preserve these distributional characteristics to create a more realistic, albeit imbalanced, corpus that mimics potential real-world data distributions for the subsequent training phase.

3.4. Training Phase 2: Fine Tuning on 11,500 Labeled Data

In the second training phase, the IndoBERT model was re-trained on the entire 11,500-sample dataset. This step aimed to enhance the model's generalization capabilities and robustness by exposing it to a larger, imbalanced corpus that mirrors real-world data characteristics. The training commenced from the best-performing checkpoint secured during the initial training phase.

To maintain methodological consistency, this second training process utilized the same hyperparameter configuration as the first phase, as previously detailed in Table 1. The learning progression during this second phase is visualized in Figure 5, which plots the model's accuracy and loss. As depicted, the model exhibited stable performance gains, with the validation accuracy curve rising in tandem with the training accuracy. Concurrently, the validation loss consistently decreased without significant signs of overfitting, which validates the effectiveness of training on a larger, imbalanced dataset.

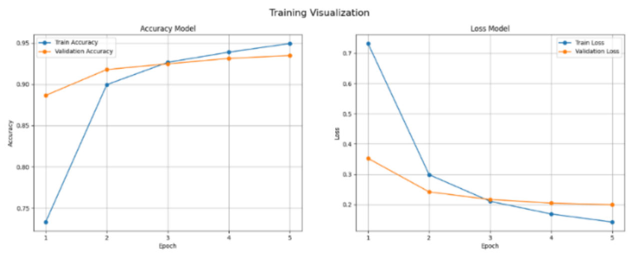

Figure 5. Visualization of the Second Training History Plot

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

As depicted, the model exhibited stable performance gains, with the validation accuracy curve rising in tandem with the training accuracy. Concurrently, the validation loss consistently decreased without significant signs of overfitting. Based on this positive trend, the model checkpoint from the final epoch was selected as the definitive model.

A final evaluation was conducted on the test set, comprising 1,150 samples. The performance results of the final model are summarized in the classification report in Table 10. The report indicates an excellent overall accuracy of 94%. The strong performance across all categories, despite the data imbalance, validates the efficacy of the two-phase training methodology.

Table 10. Final Model Classification Report on the Test Set

Emotion Label	Precision	Recall	F1-Score	Support
Enthusiasm	0.92	0.90	0.91	115
Gentle	0.95	0.93	0.94	236
Analytical	0.94	0.98	0.96	349
Inspirational	0.94	0.93	0.94	368
Cautionary	0.96	0.93	0.94	82

The final evaluation yielded an overall accuracy of 94%. The model demonstrated particularly strong performance on the majority classes, 'Analytical' (F1-score of 0.96) and 'Inspirational' (F1-score of 0.94). This robust performance across all categories, despite the inherent data imbalance, validates the efficacy of the two-phase training methodology.

To further analyze the model's classification be havior, a confusion matrix from the test set evaluation is presented in Figure 6. This matrix confirms the model's high accuracy, with strong diagonal values indicating the correct identification of most samples. The most notable confusion occurs between the 'Inspirational' and 'Analytical' classes, suggesting a degree of semantic overlap in the responses for these two categories.

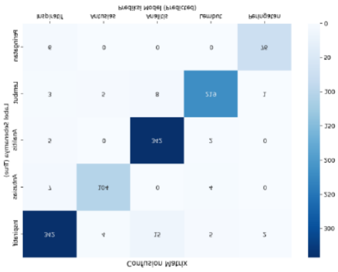

Figure 6. Confusion Matrix for the Final Model

The confusion matrix reveals that the model performed exceptionally well on the majority classes, correctly identifying 342 'Inspirational' samples and 342 'Analytical' samples. The strong diagonal values confirm the model's high accuracy. The most notable confusion occurs between the 'Inspirational' and 'Analytical' classes, where 15 'Inspirational' samples were misclassified as 'Analytical'. This suggests a degree of semantic overlap in the responses for these two categories. Despite these minor misclassifications, the matrix confirms the model's overall robustness and its ability to effectively handle an imbalanced, multi-class classification task.

3.5. Facial Expression Mapping

The final stage of this research involves translating the model's emotion classification output into a visual representation suitable for implementation in a virtual agent or chatbot. The objective of this mapping is to foster a more human-like and empathetic interaction, where the counselor's response is conveyed not only textually but also through corresponding facial expressions.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 5, October 2025, Page. 3750-3768
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

This process utilizes a direct mapping approach, where each emotion label generated by the IndoBERT model is specifically associated with a set of facial expression parameters. This framework is grounded in the psychological interpretation of each emotional function within the counseling context. For instance, an 'Inspirational' label would trigger an expression of encouragement and optimism, such as a sincere smile and bright eyes. Conversely, a 'Cautionary' label would elicit a more serious and attentive expression. The mapping between emotion labels and their visual implementation is detailed in Figure 7. This figure provides a concrete visual representation of how the analytical output from the NLP model can be translated into empathetic non-verbal cues, bridging the gap between text analysis and human-like interaction.

Figure 7. Mapping of Emotion Labels to Visual Implementation Examples. The illustrative images were generated by the author using a generative AI model.

This mapping framework serves as a foundation for the future development of virtual agent interfaces, enabling the outputs from the natural language processing model to directly animate a visual avatar, thereby creating a richer and more interactive counseling experience.

4. DISCUSSIONS

The primary findings from this research confirm that the two-phase fine-tuned IndoBERT model is highly effective at identifying emotional functions within early marriage counseling dialogues. As summarized in Table 11, this approach demonstrates highly competitive performance. In the first phase, the model, trained on a balanced 2,500-sample dataset, achieved a superior validation accuracy of 92%. This high performance can be attributed to the clean and evenly distributed dataset, which enabled the model to optimally learn the distinguishing features of each emotion category without bias. Conversely, the second phase saw an expected decrease in accuracy to 94% after the model was retrained on the entire 11,500-sample dataset. This is a crucial finding, as it demonstrates the model's robust generalization capabilities when confronted with a larger, more varied, and imbalanced dataset that better represents real-world conditions.

Table 11. Accuracy Comparison Across Training Phases and Other Studies

Model	Accuracy (%)	Emotion Label
IndoBERT Phase 1	92.0	5 Functional Emotion
IndoBERT Phase 2	94.1	5 Functional Emotion
CNN	71.6	6 Emotion Labels
KNN	79.0	2 Emotion Labels

The final accuracy of 94% also represents a significant advantage over other approaches for emotion classification in Indonesian text. The superiority of IndoBERT over methods like CNN (71.6%) and KNN (79%) lies not only in its higher accuracy figures but also in its architectural foundation. This success can be attributed to IndoBERT's Transformer-based architecture, which is capable of capturing long-range dependencies and contextual nuances in the Indonesian language far more effectively than

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

the manual feature engineering-dependent KNN or the localized feature detection of CNN. IndoBERT, with its self-attention mechanism, can holistically understand the relationships between words in a sentence, which is critical for complex emotion classification tasks [29].

This success is largely underpinned by the strategic two-phase training methodology. The first phase served to build a strong conceptual foundation, while the second phase tested and reinforced the model's generalization capabilities. This combination ensures the model is not only accurate in a controlled environment but also reliable when faced with more complex data.

4.1. Implications for Informatics and Computer Science

Beyond accuracy, the primary contributions of this research lie in several key areas of computer science. First, this study presents the development of a five-part functional emotion framework ('Inspirational', 'Analytical', etc.) for the Indonesian counseling context, offering a more nuanced analysis than standard sentiment classification. This constitutes an important contribution to the fields of Natural Language Processing (NLP) and Affective Computing, particularly for low-resource languages. Second, the integration of this workflow with facial expression mapping presents an innovative step toward the implementation of empathetic multimodal virtual agents, an area not extensively explored in research within Indonesia, thereby contributing to the field of Human-Computer Interaction (HCI).

In practical terms, the resulting model has significant implications for technological development in the mental health sector, serving as a core component for empathetic counseling chatbots or as a training aid for aspiring counselors. However, several limitations must be acknowledged. The dataset, while extensive, was generated by an LLM and not derived from actual counseling transcripts, and the five-emotion scheme may not encompass the full spectrum of emotional functions. For future research, validating the model on real-world transcript data is a priority. Exploring methods such as multi-label classification and implementing this framework into a virtual agent prototype for user studies would provide empirical validation of its effectiveness in creating more human-like interactions.

5. CONCLUSION

This study has successfully developed and validated the IndoBERT model for the task of classifying five categories of functional emotions in early marriage counseling dialogues. Through the application of a strategic two-phase training methodology starting with a balanced dataset to build a foundation and continuing with an unbalanced combined dataset to improve generalization the final model demonstrated high efficacy, achieving an overall accuracy of 94%. This performance significantly outperforms other classification methods such as CNN and KNN in the context of Indonesian text.

The main contribution of this research lies in the development of a new functional emotion framework specifically designed for the counseling context, as well as the demonstration of an effective training approach for imbalanced data. Ultimately, this research not only offers a promising technological solution to enhance counseling services but also provides a significant contribution to the field of Computer Science. By demonstrating the effective integration of text classification and facial expression mapping, this study provides a strong foundation for the development of more empathetic multimodal virtual counseling agents and delivers a replicable methodology for creating more nuanced and human-like AI systems, particularly in the fields of Affective Computing and Human-Computer Interaction (HCI).

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

CONFLICT OF INTEREST

All authors declare that they have no financial or non-financial relationships, affiliations, or involvement with any organizations or entities that could be perceived as having a potential interest or influence in the content, subject matter, or materials presented in this manuscript.

REFERENCES

[1] M. D. H. Rahiem, "COVID-19 and the surge of child marriages: A phenomenon in Nusa Tenggara Barat, Indonesia," *Child Abuse Negl*, vol. 118, p. 105168, Aug. 2021, doi: 10.1016/j.chiabu.2021.105168.

[2] F. A. Purnami, D. M. Maula, A. A. Nisa, R. Cahyaningtyas, A. R. Jundan, and F. Fitri, "Sosialisasi Dampak Pernikahan Dini dan Penguatan Mental Remaja sebagai Strategi Pencegahan Pernikahan Dini," *Welfare : Jurnal Pengabdian Masyarakat*, vol. 2, no. 4, pp. 698–703, Dec. 2024, doi: 10.30762/welfare.v2i4.1979.

[3] A. Sampurna, H. J. Ritonga, and A. R. Matondang, "Integration of Media Literacy in Religious Counseling for Preventing Early Marriage in Nias Barat," *International Journal of Islamic Education, Research and Multiculturalism (IJIERM)*, vol. 6, no. 3, pp. 1205–1218, Dec. 2024, doi: 10.47006/ijierm.v6i3.392.

[4] L. Munira, P. Liamputtong, and P. Viwattanakulvanid, "Barriers and facilitators to access mental health services among people with mental disorders in Indonesia: A qualitative study," *Belitung Nurs J*, vol. 9, no. 2, pp. 110–117, Apr. 2023, doi: 10.33546/bnj.2521.

[5] C. D. Putra, R. A. S. Prayoga, M. Cinthya, R. Basatha, M. S. Akbar, and E. A. Elfaiz, "Mengembangkan Chatbot Empatik untuk Dukungan Kesehatan Mental: Solusi Inovatif dalam Pendampingan Psikologis," *Jurnal Ilmu Komputer dan Multimedia*, vol. 1, no. 2, pp. 7–12, Dec. 2024, doi: 10.46510/ilkomedia.v1i2.19.

[6] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Lebouché, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *J Med Internet Res*, vol. 26, p. e56930, Jul. 2024, doi: 10.2196/56930.

[7] I. A. Assegaf *et al.*, "PENGEMBANGAN CHATBOT KONSULTASI KESEHATAN MENTAL KESEHATAN MENTAL BERBASIS OPEN AI MODEL GPT-3.5 TURBO MENGGUNAKAN MEDIA WHATSAPP," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 6, no. 4, pp. 785–793, Nov. 2024, doi: 10.51401/jinteks.v6i4.4749.

[8] G. Park, J. Chung, and S. Lee, "Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model," *Current Psychology*, vol. 42, no. 32, pp. 28663–28673, Nov. 2023, doi: 10.1007/s12144-022-03932-z.

[9] H. Chin *et al.*, "The Potential of Chatbots for Emotional Support and Promoting Mental Well-Being in Different Cultures: Mixed Methods Study," *J Med Internet Res*, vol. 25, p. e51712, Oct. 2023, doi: 10.2196/51712.

[10] J. Sun, "Research And Application Analysis of Multimodal Emotion Recognition Methods Based on Speech, Text, And Facial Expressions," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 293–297, Mar. 2024, doi: 10.54097/agvjvq19.

[11] Y. Zhang, "Computer-Assisted Human-Computer Interaction in Visual Communication," *Comput Aided Des Appl*, vol. 18, no. S1, pp. 109–119, May 2020, doi: 10.14733/cadaps.2021.S1.109-119.

[12] L. D. Cahya, A. Luthfiart, J. I. T. Krisna, S. Winarno, and A. Nugraha, "Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 9, no. 3, pp. 290–298, Jan. 2024, doi: 10.25077/TEKNOSI.v9i3.2023.290-298.

[13] W. M. Baihaqi and A. Munandar, "Sentiment Analysis of Student Comment on the College Performance Evaluation Questionnaire Using Naïve Bayes and IndoBERT," *JUITA : Jurnal Informatika*, vol. 11, no. 2, p. 213, Nov. 2023, doi: 10.30595/juita.v11i2.17336.

[14] M. Y. Baihaqi, E. Halawa, R. A. S. Syah, A. Nurrahma, and W. Wijaya, "Emotion Classification in Indonesian Language: A CNN Approach with Hyperband Tuning," *Jurnal Buana Informatika*, vol. 14, no. 02, pp. 137–146, Oct. 2023, doi: 10.24002/jbi.v14i02.7558.

[15] A. Zamsuri, S. Defit, and G. W. Nurcahyo, "Classification of Multiple Emotions in Indonesian Text Using The K-Nearest Neighbor Method," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 4, no. 2, pp. 1012–1021, Jun. 2023, doi: 10.37385/jaets.v4i2.1964.

[16] S. Elysia and Herianto, "Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah," *Journal TIFDA (Technology Information and Data Analytic)*, vol. 1, no. 2, pp. 52–58, Dec. 2024, doi: 10.70491/tifda.v1i2.52.

[17] F. R. Fatonah, D. S. Maylawati, and E. Nurlatifah, "Chatbot Edukasi Pra-Nikah berbasis Telegram Menggunakan Bidirectional Encoder Representations From Transformers (BERT)," *Jurnal Algoritma*, vol. 21, no. 2, pp. 29–40, Nov. 2024, doi: 10.33364/algoritma/v.21-2.1657.

[18] A. Ardhiyansyah, C. Bakker, and S. G. Sijabat, "Dampak Teknologi Digital terhadap Kesejahteraan Mental: Tinjauan Interaksi, Tantangan, dan Solusi," *Jurnal Psikologi dan Konseling West Science*, vol. 1, no. 04, pp. 181–188, Sep. 2023, doi: 10.58812/jpkws.v1i04.651.

[19] M. Nesca, A. Katz, C. K. Leung, and L. M. Lix, "A scoping review of preprocessing methods for unstructured text data to assess data quality.," *Int J Popul Data Sci*, vol. 7, no. 1, p. 1757, 2022, doi: 10.23889/ijpds.v6i1.1757.

[20] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[21] S. William, Kenny, and A. Chowanda, "EMOTION RECOGNITION INDONESIAN LANGUAGE FROM TWITTER USING INDOBERT AND BI-LSTM," *Communications in Mathematical Biology and Neuroscience*, vol. 2024, 2024, doi: 10.28919/cmbn/7858.

[22] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc Netw Anal Min*, vol. 15, no. 1, p. 22, Mar. 2025, doi: 10.1007/s13278-025-01439-6.

[23] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[24] R. J. L. Melatisudra, S. Utomo, S. Sutjiningtyas, and H. Hernawati, "Implementasi Pengenalan Ekspresi Wajah dengan Menggunakan Metode Convolutional Neural Network dan OpenCV Berbasis Webcam," *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 1, pp. 339–348, Nov. 2024, doi: 10.47065/josyc.v6i1.6114.

[25] K. Wisnudhanti and F. Candra, "Image Classification of Pandawa Figures Using Convolutional Neural Network on Raspberry Pi 4," *J Phys Conf Ser*, vol. 1655, no. 1, p. 012103, Oct. 2020, doi: 10.1088/1742-6596/1655/1/012103.

[26] D. M. Aprilla, F. Bimantoro, and I. G. P. S. Wijaya, "The Palmprint Recognition Using Xception, VGG16, ResNet50, MobileNet, and EfficientNetB0 Architecture," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 1065, Apr. 2024, doi: 10.30865/mib.v8i2.7577.

[27] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artif Intell Rev*, vol. 58, no. 7, p. 195, Apr. 2025, doi: 10.1007/s10462-025-11198-7.

[28] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sens (Basel)*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.

Vol. 6, No. 5, October 2025, Page. 3750-3768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5039

[29] Rajesh Ediga, "Understanding the Technical Foundations of Large Language Models: Architectures, Training, and Applications," *Journal of Computer Science and Technology Studies*, vol. 7, no. 7, pp. 154–161, Jul. 2025, doi: 10.32996/jcsts.2025.7.7.13.