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

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Mapping Gestures Based on Text Emotion Classification for a Virtual Chatbot for Early Marriage Consultation in Lombok Using RoBERTa Model

Adam Zahran Ramadhan*[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]adamzahranr@student.telkomuniversity.ac.id

Received : Jul 4, 2025; Revised : Aug 23, 2025; Accepted : Aug 12, 2025; Published : Oct 22, 2025

Abstract

To address the persistent issue of early marriage among Indonesian adolescents, this study proposes a virtual counseling chatbot that classifies emotional cues in text using a fine-tuned IndoRoBERTa model. The emotion classification framework is designed to support counseling-based prevention efforts by moving beyond basic sentiment analysis and adopting five functional emotional categories such as 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary' to align with psychological counseling styles. Built on fine-tuned IndoRoBERTa architecture, the model was trained in two phases: first with 2,500 manually validated samples yielding 92.8% accuracy, and then with 12,500 auto-labeled entries, resulting in 91.3% accuracy. Performance was assessed using accuracy, precision, recall, and F1-score. A gesture mapping layer was also integrated to enhance empathetic response generation. Each emotion label was paired with a predefined body gesture, grounded in counseling theory, to support future development of multimodal virtual agents capable of expressing emotions both textually and physically. The novelty lies in combining context-aware emotion classification with gesture mapping, enabling future development of expressive, culturally relevant, and empathetic virtual chatbot agents.

Keywords: *Early Marriage, Emotion Classification, Gesture Mapping, IndoRoBERTa, NLP, Virtual Chatbot.*

1. INTRODUCTION

Early marriage remains a long standing and serious concern in Indonesia, particularly in Lombok, located in the province of West Nusa Tenggara (NTB). Although often rooted in cultural customs and social expectations, this practice has wide reaching effects that stretch beyond the individuals directly involved [1]. Early marriage contributes to intergenerational cycles of poverty, restricts access to education, and increases the risk of both physical and mental health problems. Despite numerous efforts from national and local governments as well as non-governmental organizations, the issue remains persistent [2]. According to the Indonesian Central Bureau of Statistics, in 2024, the province of West Nusa Tenggara (NTB) recorded the country's highest percentage of women aged 20 to 24 who had married or lived with a partner before turning 18, reaching 16.23%, the highest rate nationally [3].

Given this context, several studies have explored how digital technologies, particularly those based on Natural Language Processing (NLP) can serve as supportive tools for youth well-being. With progress in digital technology, particularly in Natural Language Processing (NLP), innovative forms of social support have become possible [4]. One emerging approach involves the use of virtual chatbots [5], [6]. These tools are built to mimic human conversation and have been applied in areas such as health

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

counseling, education, and emotional support [6]. Their constant availability and capacity for anonymous interaction make them especially useful for addressing sensitive issues like early marriage.

Numerous studies have examined the application of Natural Language Processing (NLP) models for emotion detection and text classification. Some have explored combining Word2Vec with Long Short-Term Memory (LSTM) networks to capture emotional tone in sequential text, showing strong performance in context recognition [7]. Other research has employed Convolutional Neural Networks (CNN) for emotion classification in Indonesian corpora, demonstrating greater accuracy than traditional methods like Naïve Bayes or Support Vector Machines [8]. Topic classification using Word2Vec with K-Nearest Neighbor (KNN) has also shown promise in grouping similar documents by meaning [9]. More recently, transformer-based models designed specifically for Indonesian have been introduced. IndoBERT, for instance, was developed to address the unique features of Bahasa Indonesia and has outperformed LSTM models in tasks such as hoax detection and multi-class emotion classification [10], [11]. These findings suggest that transformer models are better suited for capturing the complex meanings in Indonesian textual data.

The development of more advanced model architectures has led to IndoRoBERTa, a RoBERTa-based model tailored for the Indonesian language through broader pretraining methods. IndoRoBERTa has been tested on various NLP tasks, including sentiment and emotion classification. Compared to conventional deep learning approaches like CNN and Naïve Bayes, it consistently delivers more context-aware results, particularly in dialogue-heavy settings [12], [13], [14]. These outcomes suggest IndoRoBERTa is well-suited for tasks that demand a deep understanding of emotional tone and contextual nuance in Indonesian texts.

Despite the overall success of such models, few have been adapted to domain-specific applications such as early marriage counseling. Conversations in this setting often involve layered emotional experiences such as fear, anxiety, shame, or social pressure that go beyond what basic sentiment analysis can detect. Generic emotion classifiers tend to miss these subtleties. In addition, most studies have not considered how emotion detection might inform a chatbot's nonverbal features, like expressive gestures or vocal tone changes, which are critical for empathetic communication [11].

Motivated by these gaps, this study proposes a virtual counseling chatbot that classifies emotional cues in text using a fine-tuned IndoRoBERTa model. This study adopts a more targeted emotional labeling scheme rather than relying on broad categories like ‘happy’, ‘sad’, or ‘angry’, it introduces five refined labels such as ‘Enthusiastic’, ‘Gentle’, ‘Analytical’, ‘Inspirational’, and ‘Cautionary’. These better reflect both the emotional state and the communicative purpose in counseling contexts. Recent work in emotion analysis highlights the shortcomings of overly simplified emotion models and encourages the use of layered, context-aware frameworks. For instance, Plaza-del-Arco, et al [15], reviewed over 150 NLP studies and noted that many fail to account for emotional complexity, cultural differences, and situational nuance.

In a related analysis, Koufakou, et al [16], examined 30 emotion annotated corpora introduced since 2018 and noted a growing trend toward the adoption of multi label structures such as the GoEmotions dataset, which maps emotional expressions across 27 distinct categories [16]. These labels mirror the way counsellors or educators speak during sensitive conversations, such as those about early marriage, where the goal might be to reassure, guide, or encourage rather than simply convey emotion.

This study presents the development of an emotion aware chatbot system tailored for early marriage consultation, with IndoRoBERTa, an Indonesian language transformer-based model, at the core of its design. Emotion recognition plays a key role in enabling the system to respond with empathy, especially in conversations involving personal and culturally sensitive topics. The model is fine-tuned using a synthetic yet context-aware dataset of 2,500 labeled text samples designed to mirror interactions between adolescents and counselors. These samples, generated through a large language model, capture

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

the emotional depth often seen in real counseling scenarios. To improve performance, IndoRoBERTa is further applied to a broader set of 12,500 unlabeled samples, enhancing its ability to interpret varied emotional expressions.

The novelty of this research lies in the integration of function-based emotion labels and gesture mapping grounded in counseling theory. Unlike conventional approaches that rely on basic sentiment or emotion categories, this study adopts labels such as 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary' to better reflect the communicative intent found in psychological counseling dialogues. In addition, each classified emotion is linked to specific nonverbal gestures, enabling a more expressive and empathetic response generation. This combined approach, emotion-aware classification and gesture-informed modeling, has not been previously explored in the context of Indonesian counseling applications [17], [18]. As a result, the study offers a novel framework for developing culturally sensitive virtual agents that engage users in more human-centered and emotionally intelligent ways.

2. METHOD

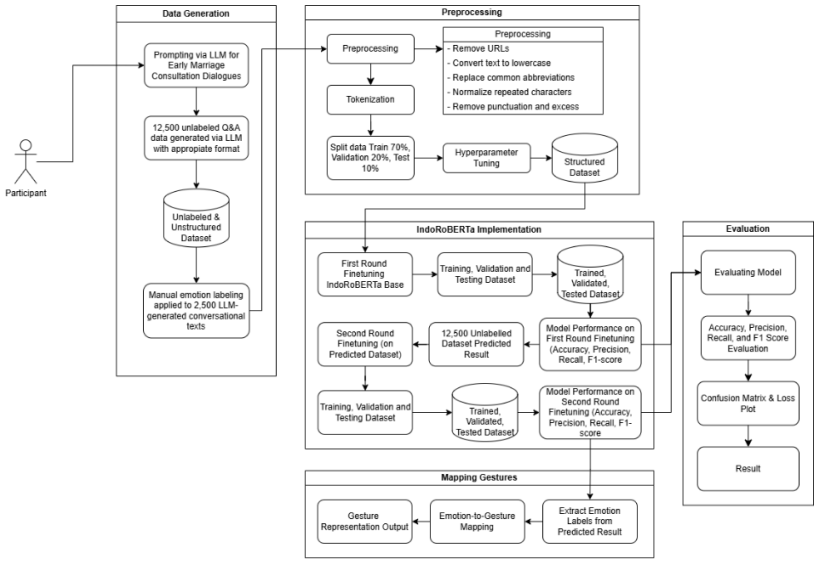

Figure 1. The Proposed Diagram of System Architecture.

The system proposed in this study follows a structured framework for developing an emotion classification model tailored to early marriage consultation. The overall process is outlined in Figure 1, which illustrates a step-by-step workflow from data generation to emotion prediction. Each phase plays a key role in establishing a systematic and replicable approach for model training and evaluation. The first step involves generating 12,500 unlabelled question–answer text samples using a large language model (LLM), crafted to reflect realistic dialogue typically encountered in counseling contexts. From this corpus, a subset of 2,500 samples is manually annotated with emotion labels based on a domain specific classification scheme, forming the foundation for model training. In the subsequent preprocessing stage, the annotated texts are tokenized, hyperparameter tuned and divided into training, validation, and testing sets. Indonesian RoBERTa Base, a pretrained Indonesian language model, is then fine-tuned on this labelled dataset to support multi class emotion classification. This phase allows the model to learn patterns and emotional structures relevant to the counseling domain. Following this phase, the full system integration is depicted in Figure 2.

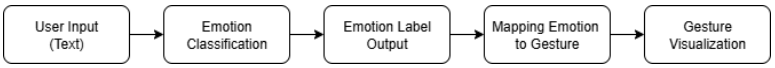

Figure 2. System Flow Diagram.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Figure 2 illustrates the simplified flowchart of the overall system. It outlines the sequential process starting from user text input, followed by emotion classification, label output, gesture mapping, and finally, gesture visualization. While this diagram focuses on functional components, it complements the more detailed visual design that represents character poses based on emotion categories. Each stage in the pipeline is depicted as a text-based block, showing how the system processes input text to produce emotion-appropriate non-verbal gestures. The gesture visualization here refers to static character poses that reflect the classified emotion, without involving 3D animation or real-time rendering. This overview provides a general understanding of how the system integrates emotion recognition and gesture generation for persuasive responses.

After establishing the system pipeline, the focus shifts to model development and evaluation. After training, model performance is assessed using common evaluation metricsaccuracy, precision, recall, and F1-score. Visualization tools such as the confusion matrix and training-validation loss plots provide additional diagnostic insights, highlighting areas of strength and potential misclassification [19], [20]. Once validated, the model is used to predict emotion labels for the full 12,500 sample dataset. These predicted results are not only passed into a gesture mapping module where each emotion label is associated with gesture descriptions but also leveraged for a second round of fine-tuning. In this stage, the model is retrained using the newly labeled 12,500 sample dataset to further stabilize learning, reduce noise from initial annotation inconsistencies, and enhance generalization. This iterative refinement enables more accurate emotion predictions and supports the development of gesture representations for future virtual counseling agents, promoting more expressive and empathetic interactions through multimodal emotion understanding.

2.1. Data Generation

The dataset developed for this study was carefully constructed to reflect the emotional and contextual depth of early marriage consultation dialogue. A total of 12,500 random unlabeled question–answer pairs were generated using a Large Language Model (LLM), prompted to simulate realistic conversations between young individuals and psychological counsellors. These exchanges were designed to capture themes frequently encountered in consultations such as fear, doubt, encouragement, reflection, and social pressure. To ensure cultural and linguistic authenticity, all prompts were written in Bahasa Indonesia and adjusted to reflect local vernacular and socio-cultural dynamics, particularly those characteristics of Lombok and comparable regions.

Due to the lack of publicly available or opensource datasets specifically focused on early marriage consultations in Indonesian contexts, no web scraping or manual data collection from real counseling sessions was conducted. Furthermore, direct collaboration with psychologists or mental health institutions was not pursued in this study, in part due to ethical and privacy concerns regarding sensitive client data, as well as time and access limitations. As a result, synthetic data generation via LLMs was selected as the most practical and scalable approach to simulate a wide variety of representative dialogues. To ensure balanced representation across all emotion types, the manually labeled subset of 2,500 samples was equally distributed, with 500 samples allocated to each of the five emotional categories. This strategy was implemented to prevent class imbalance during model training and allow the classifier to learn each emotional function with equal emphasis.

A subset of 2,500 samples from the corpus was manually annotated using a function-based emotion classification framework designed for the study. The emotional categories ‘Enthusiastic’, ‘Gentle’, ‘Analytical’, ‘Inspirational’, and ‘Cautionary’, were chosen for both their emotional tone and the communicative intent they convey in counseling settings. For example, the ‘gentle’ label may reflect a reassuring response, while ‘Cautionary’ may signal a warning delivered with care. Annotation was carried out using a detailed rubric that defined each emotion’s purpose, linguistic markers, and typical

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

conversational context. This annotated subset was then used to fine-tune the IndoRoBERTa Base model, a language model tailored for Indonesian and suited for context-sensitive emotion recognition.

Although the initial emotion labeling was performed automatically using a Large Language Model (LLM), all annotations underwent a manual review process to prevent misclassification. Each labeled instance was double-checked to ensure it aligned with the predefined emotion categories and communicative functions. This post-labeling validation helped maintain labeling consistency and ensured that the dataset was sufficiently reliable for fine-tuning the emotion classification model.

2.2. Preprocessing

The preprocessing stage standardizes and optimizes the dataset before applying tokenization with the IndoRoBERTa Base model [21]. Rather than using heavy data cleaning, this approach focuses on light but context-aware normalization to retain emotional and conversational subtleties. The process starts by converting all text to lowercase for consistency, so a phrase like "Aku Suka Ini" (“I Like It”) becomes "aku suka ini" (“i like it”). Unnecessary web content is removed by stripping URLs, turning "click url https://example.com" into just "click url." Informal abbreviations common in everyday Indonesian are expanded to their full forms, such as "gk" to "nggak" (meaning “no”) and "udh" to "udah" (meaning “already”), helping maintain meaning. Repeated letters used for emphasis, like "baguuuusss", are simplified to their base form "bagus" (“good”), keeping emotional tone without introducing noise. Extra spaces are trimmed as well, producing a clean, emotionally rich dataset ready for IndoRoBERTa’s tokenizer. The entire preprocessing workflow is shown in Figure 3.

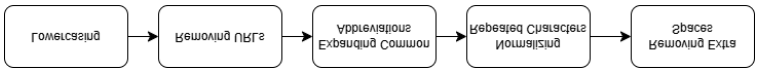

Figure 3. Preprocessing Features.

2.3. Data Splitting and Hyperparameter Tuning

The dataset was divided into training, validation, and testing sets in a 70:20:10 ratio. This setup allows the model to learn effectively while keeping separate data for fine-tuning and fair evaluation. It helps reduce overfitting and improves the model’s ability to handle new, unseen inputs.

Table 1. Hyperparameter Settings for First Phase Finetuning.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0 - 3e-5(with scheduler)
Epochs	5
Batch Size	8

To fine-tune the IndoRoBERTa Base model, standard transformer hyperparameters were used with adjustments for stability and efficiency. As listed in Table 1, the AdamW optimizer was selected to manage weight decay and reduce overfitting. The learning rate was capped at 3e-5, with a scheduler that gradually increased and then decreased the rate to support steady convergence. This helped avoid overshooting and made training more stable, especially critical for large pre-trained models. Training was limited to five Epochs to balance learning and prevent overexposure to the data [22]. Training was run for five Epochs, providing enough iterations for the model to learn meaningful patterns without overexposing it to the same data. Given hardware considerations, a batch size of 8 was used to ensure that training remained feasible on commonly available GPUs while preserving gradient reliability [23]. To further support generalization, only the training set was shuffled between Epochs, ensuring the model encountered a diverse sequence of examples throughout the learning process. These hyperparameter

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

decisions were aimed at achieving a practical balance between performance, training efficiency, and computational constraints and was applied during the first fine-tuning phase, which used a manually balanced dataset of 2,500 annotated samples (500 per emotion category).

Table 2. Hyperparameter Settings for Second Phase Finetuning.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0 - 3e-5(with scheduler)
Epochs	3
Batch Size	8

For the second fine-tuning phase, the same configuration was maintained except for the number of training Epochs, which was reduced to 3 as shown in Table 2. This adjustment is supported by practices in transfer learning, where fewer Epochs are often sufficient when fine-tuning on noisy or weakly labeled data [24]. Since these data were generated randomly without balancing the emotion, the resulting class distribution was naturally imbalanced. This imbalance led to fluctuating validation performance, with less consistent accuracy and validation loss compared to the initial phase. Reducing the Epoch counts in the second training stage helped to mitigate overfitting on dominant emotion classes and allowed the model to stabilize despite the noisier label distribution.

2.4. IndoRoBERTa Implementation

The model used in this study is based on IndoRoBERTa, a transformer-based architecture adapted from the RoBERTa model [12]. IndoRoBERTa shares structural similarities with IndoBERT, which itself is based on the original BERT framework and trained specifically on Indonesian language corpora [25]. However, IndoRoBERTa applies several modifications inspired by the RoBERTa architecture, such as removing the Next Sentence Prediction (NSP) objective, training on longer sequences, and using a larger volume of training data. These changes aim to improve contextual understanding and yield more robust performance compared to the original BERT-based models, especially in downstream tasks like sentiment and emotion classification in Bahasa Indonesia. Given its enhanced capabilities, IndoRoBERTa is well-suited for capturing nuanced emotional patterns in counseling-related texts.

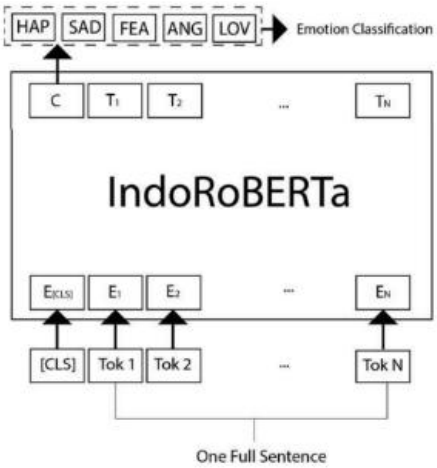

Figure 4. Visualization of Proposed IndoRoBERTa Architecture [13].

Figure 4 illustrates the architecture of the IndoRoBERTa Base model as applied in this study for the task of emotion classification. The process begins with tokenizing the input sentence into sub-word units (Tok1, Tok2, ..., TokN) using a pretrained tokenizer. A special classification token [CLS] is added

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

to the beginning of the sequence, and each token is then mapped to its corresponding embedding (E[CLS], E1, E2, ..., EN) through the model's embedding layer. These embeddings are passed through 12 transformer encoder layers, which generate contextualized representations (C, T1, T2, ..., TN) for each token. The contextual vector of the [CLS] token denoted as "C" serves as an aggregated summary of the input and is subsequently passed to a classification head, which predicts the most appropriate emotion label.

What distinguishes this implementation from standard emotion classification models is its adoption of a function-based emotional framework tailored to the counseling context. While conventional models such as the IndoRoBERTa Emotion Classifier trained on the IndoNLU EmoT dataset from Indonesian Twitter typically rely on basic emotional categories like anger, fear, happiness, love, or sadness, this study takes a different approach. The IndoRoBERTa Base model is fine-tuned to recognize five custom emotional such as 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary'. The emotion labels used in this study were carefully designed to capture the specific emotional roles found in early marriage counseling, such as giving advice, offering comfort, or encouraging motivation. Instead of relying on broad emotional categories, this approach makes use of the general IndoRoBERTa Base model while allowing for custom emotion classes that better reflect the goals and tone of real counseling dialogue. This flexibility is vital in sensitive topics, where emotional intent often goes deeper than surface-level feelings.

Once training and evaluation were complete, the fine-tuned IndoRoBERTa Base model was used to label the full set of 12,500 Q&A samples, including both the 10,000 unlabeled texts and the original 2,500 manually tagged entries. Including the labeled subset allowed for direct comparison between human and model predictions, serving as an added check on the model's accuracy. To further improve reliability and reduce any early annotation errors, the results from this full dataset were used for a second round of fine-tuning. This step reinforced the model's understanding, sharpened classification accuracy, and improved its readiness for follow-up tasks like gesture mapping or integration into virtual counseling tools.

All inputs were preprocessed the same way as during training to keep tokenization consistent. Each sample was then classified into one of five emotion categories: 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', or 'Cautionary'. These predicted labels were added to the dataset, creating a complete emotion-annotated corpus.

To evaluate performance, results were reviewed both quantitatively and qualitatively. Emotion label distribution was analyzed to detect bias, while selected entries were manually checked to confirm whether the model's choices matched the tone and intent of the dialogue. This final round of prediction not only completed the dataset but also helped uncover deeper patterns in the model's behavior, particularly by comparing its output to trusted human labels [26].

2.5. Evaluation Phase

To evaluate the performance of the fine-tuned IndoRoBERTa model in recognizing function-based emotional categories, four standard classification metrics were used: accuracy, precision, recall, and F1-score [27]. Together, these metrics offer a well-rounded view of how effectively the model identifies the five emotion classes 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary', each capturing a distinct communicative purpose within counseling conversations.

2.5.1. Accuracy

Accuracy calculates the ratio of correct predictions to the total number of predictions made. It offers a straightforward snapshot of overall performance, especially valuable when class distributions are relatively balanced [28], the evaluation of accuracy is presented in Equation (1).

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Samples\ (TP+TN+FP+FN)} \quad (1)$$

True Positive (TP): Instances that are positive and correctly identified as such by the model.
True Negative (TN): Instances that are negative and correctly recognized as negative by the model.
False Positive (FP): Negative instances that are incorrectly classified as positive
False Negative (FN): Positive instances that are mistakenly predicted as negative.

2.5.2. Precision

Precision evaluates the proportion of correct positive predictions made by the model out of all positive predictions [29]. This metric is particularly valuable when the cost of false positives is high such as wrongly labeling a cautionary message as inspirational. As shown in the equation (2).

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

A model with high precision ensures that the emotional labels it assigns especially those with sensitive counseling intent are reliable and appropriate for the user context [29].

2.5.3. Recall

In contrast to precision, recall measures the ability of the model to correctly identify all relevant instances of a given emotion class [27]. Computed using the formula in equation (3).

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negative\ (FN)} \quad (3)$$

Recall becomes especially important in applications like early marriage consultation, where missing out on important emotional cues (false negatives) could lead to under-responsive or unempathetic system behavior.

2.5.4. F1-Score

The F1-score is used to strike a balance between precision and recall, serving as their harmonic mean. It's particularly valuable when dealing with uneven class distributions or when it's important to reduce both false positives and false negatives [27]. The F1-score is computed using the equation shown in (4).

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

This balanced metric is particularly appropriate for multi-class emotion classification, where each emotional label may have differing degrees of representation and predictive difficulty.

2.5.5. Confusion Matrix and Loss Plot Curve

In addition to standard metrics, visual tools like confusion matrices and loss curves are essential for understanding how well a deep learning model learns and generalizes. A confusion matrix compares predicted and actual emotion labels, making it easier to see which categories the model struggles to separate. This is particularly helpful in multi-class emotion tasks where categories like "Gentle" and "Inspirational" may share subtle similarities.

Loss curves, on the other hand, track training and validation loss across Epochs. These plots help spot overfitting, when training loss drops but validation loss rises or underfitting, where both losses remain high or inconsistent. Ideally, both curves should show steady decline, indicating that the model is learning properly and generalizing to new data.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Together, these tools offer more than performance monitoring, they support meaningful model refinement. Prior research supports this perspective, for instance, Sathyanarayanan, et al [20], emphasized the value of confusion matrices for improving model accuracy, and Jiang, et al [30], showed how analyzing loss curves can guide hyperparameter tuning in complex tasks.

2.6. Mapping Emotion Label to Gesture

While this study centers on textual emotion classification, the framework is purposefully designed with future multimodal integration in mind particularly through the incorporation of gesture mapping. As illustrated in Figure 1, the overall system includes a dedicated Mapping Gestures module, which translates each emotion label into corresponding non-verbal expressions such as facial cues and body gestures. Importantly, this mapping process is not derived from model predictions but is instead based on a predefined schema developed during the early stages of this study. Each functional emotion label (e.g., Analytical, Enthusiastic, Cautionary) was manually associated with specific gesture patterns, drawing from human-computer interaction literature and psychological insights. These predefined associations serve as a design reference for future implementation in virtual counseling agents that aim to deliver emotionally expressive and context-sensitive feedback through both text and embodied gestures. Table 3 provides a detailed mapping between each emotion label and its corresponding body gestures. These gesture sets are designed to visually reinforce the communicative function of each emotion category within conversational interactions.

Table 3. Guide to textualize emotion to body gestures.

Emotion	Body gestures	Description
Enthusiastic	Broad smile, open eyes, raised eyebrows	Conveys excitement and engagement through expressive facial cues and open emotional stance.
Gentle	Soft smile, relaxed eyes, calm eyebrows	Suggests calmness and comfort, indicating empathy and a soothing presence.
Analytical	Furrowed eyebrows, closed lips, focused gaze	Reflects critical thinking and attention, signaling cognitive processing or problem analysis.
Inspirational	Sincere smile, bright eyes, slightly raised eyebrows	Expresses motivation and optimism, encouraging forward looking and positive interaction.
Cautionary	Slight frown, serious gaze, focused eye contact	Indicates concern or careful warning, communicating attention and a sense of measured caution.

Gesture-based emotion interpretation is supported by earlier research. As noted by Diwan et al [31], found that both facial expressions and body movements play a key role in conveying emotion, as non-verbal behavior often carries essential emotional signals. For example, upright posture and open-handed gestures are typically associated with enthusiasm or inspiration, while slouched or closed body language can reflect caution or emotional withdrawal. These findings support the use of bodily gestures to represent the functional emotional categories used in this study, offering a solid basis for developing virtual agents that can communicate with greater empathy in sensitive contexts like early marriage counseling.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

3. RESULT

This section outlines the results of the experiments, highlighting how well the IndoRoBERTa Base model performed in classifying emotional functions in early marriage counseling dialogues. The evaluation covers a two-stage fine-tuning process, first with 2,500 manually labeled samples, then with 12,500 labeled through model inference. Performance was measured using standard metrics: accuracy, precision, recall, and F1-score. Visual tools such as loss and accuracy plots, along with confusion matrices, were also used to support the analysis.

In addition, this study incorporates a predefined emotion to gesture mapping schema, developed independently from the model output. Each emotion label was manually associated with specific facial expressions and body gestures based on psychological and human-computer interaction literature. This mapping serves as a foundational reference for future work on multimodal virtual agents, enabling more expressive and empathetic dialogue delivery in early marriage counseling contexts.

3.1. Dataset Overview

The dataset developed for this study contains 12,500 question-response pairs that simulate dialogue between adolescents and psychological counselors in the context of early marriage consultations. It was assembled in two stages. First, 2,500 samples were manually labeled using a custom five emotion scheme: ‘Enthusiastic’, ‘Gentle’, ‘Analytical’, ‘Inspirational’ and ‘Cautionary’, with 500 examples assigned to each class for balance. The remaining 10,000 pairs were labeled automatically using the previously fine-tuned IndoRoBERTa emotion classification model.

To support effective model training, all text samples went through a preprocessing phase, which included text normalization and the expansion of common contractions and informal expressions. Table 4 and Table 5 presents an extract of the annotated data for reference. After inference on the full set, Table 6 shows the final distribution of emotion labels across all 12,500 samples. As expected, some imbalance emerged in the auto labeled portion due to the generative nature of the LLM, which tended to favor certain emotional tones such as 'Enthusiastic and Analytical more frequently than others. This is a natural consequence of using LLM-generated responses, where stylistic and semantic tendencies of the model can distort outputs toward more assertive or reflective tones.

Table 4. Example of Q&A pairs in Bahasa Indonesia.

Input	Response	Label
Nikah muda bikin aku lebih jago bikin video, nggak?	Kamu bisa bikin video kece sekarang! Coba edit pakai aplikasi atau bikin konten TikTok, kreasimu pasti viral!	Enthusiastic
Aku merasa jelek dan takut nggak ada yang mau kalau nggak nikah sekarang.	Setiap orang itu berharga, termasuk kamu. Nilai dirimu nggak ditentukan seberapa cepat kamu menikah.	Gentle
Apakah menikah muda bisa membuat hubungan jadi lebih kuat?	Hubungan bisa jadi kuat jika dibangun dengan komunikasi dan pemahaman yang matang, bukan hanya usia muda.	Analytical
Boleh nggak sih aku tunda nikah dulu?	Boleh banget. Masa depanmu penting, dan kamu berhak menjalaninya dengan persiapan matang.	Inspirational
Aku lihat di YouTube banyak yang nikah muda terus bisnis bareng dan sukses.	Yang viral itu kisah sukses, bukan perjuangannya. Jangan tertipu, sukses butuh lebih dari sekadar status menikah.	Cautionary

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Table 5. Example of Q&A pairs in English.

Input	Response	Label
Will getting married young make me better at making videos?	You can totally create awesome videos now! Try editing with an app or making TikTok content, your creativity could go viral!	Enthusiastic
I feel ugly and scared no one will want me if I don't get married soon.	Every person is valuable, including you. Your worth isn't defined by how quickly you get married.	Gentle
Does getting married young make a relationship stronger?	A strong relationship is built on communication and mutual understanding, not just on getting married young.	Analytical
Is it okay if I postpone marriage for now?	Absolutely. Your future matters, and you deserve to live it with proper preparation.	Inspirational
I see a lot of young couples on YouTube who get married and run successful businesses together.	What goes viral is the success story, not the struggle behind it. Don't be fooled, success takes more than just being married.	Cautionary

Table 6. Label distribution on phase 2.

Label	Count	Percentage
Enthusiastic	3,700	29.6%
Gentle	2,019	16.1%
Analytical	2,375	19%
Inspirational	1,958	15.6%
Cautionary	2,448	19.6%

To offer a complete view of dataset evolution, Table 7 previously presented the distribution from the initial 2,500 manually labeled samples, which were deliberately balanced across all five emotion categories. By contrast, Table 6 illustrates the emergent imbalance following automated labeling, revealing the model's preferential tendencies based on the prompt-response dynamics.

Table 7. Label distribution on phase 1.

Label	Count	Percentage
Enthusiastic	500	20%
Gentle	500	20%
Analytical	500	20%
Inspirational	500	20%
Cautionary	500	20%

For model training and evaluation, the full dataset was split into training, validation, and test sets using stratified sampling to preserve the class proportions. This split 70% for training, 20% for validation, and 10% for testing ensured that the model could learn from a diverse and balanced subset of data while being evaluated on unseen examples.

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

In the initial training phase, the IndoRoBERTa Base model was fine-tuned on a focused set of 2,500 counseling dialogue samples, each annotated with one of five function-based emotion labels such

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

as 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational' and 'Cautionary'. These samples were generated using a Large Language Model (LLM), then manually reviewed to ensure the assigned labels matched the intended communicative tone. Each category was equally represented with 500 samples, creating a balanced dataset that allowed the model to learn without bias toward any single emotion class.

The fine-tuning process used the AdamW optimizer and a learning rate scheduler, where the learning rate gradually increased from 0 to a maximum of 3e-5. A batch size of 8 was selected to strike a balance between computational efficiency and model performance, particularly when working within the limits of GPU memory. Training ran for a maximum of five Epochs, with early stopping enabled to retain the best performing checkpoint based on validation loss, although it was ultimately not triggered. These hyperparameter settings, summarized in Table 8, formed the foundation for the model's ability to distinguish subtle emotional cues in subsequent inference tasks.

Table 8. Hyperparameter configuration for Training Phase 1.

Hyperparameter	Value
Base model	IndoRoBERTa Base
Learning Rate	0 - 3e-5(with scheduler)
Optimizer	Adam
Epochs	5
Batch Size	8
Max sequence length	128 tokens
Loss function	CrossEntropyLoss
Early stopping	Patience = 2 Epochs

Throughout training, the model demonstrated consistent improvements in both training and validation sets peaked at Epoch 5. The performance metrics are summarized in Table 9.

Table 9. Performance Metrics from Test Set.

Label	Accuracy	Precision	Recall	F1-Score
Enthusiastic	1.00	1.00	1.00	1.00
Gentle	0.92	0.94	0.92	0.93
Analytical	0.86	0.83	0.96	0.84
Inspirational	0.98	0.96	0.98	0.97
Cautionary	0.88	0.92	0.88	0.90

A confusion matrix was generated and is presented in Figure 5. This visualization highlights how accurately the model distinguished between the five emotion categories in the test set. Each row of the matrix represents the actual label, while each column indicates the predicted label. High values along the diagonal indicate correct predictions, while off-diagonal values represent misclassifications.

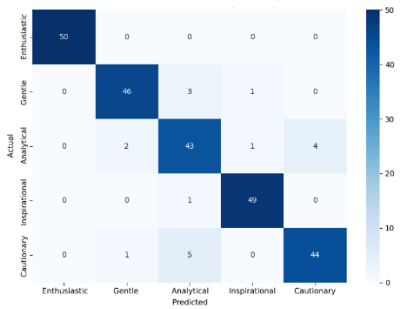

Figure 5. Confusion Matrix on Phase 1.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

The model generally performed well in distinguishing between different emotion categories, showing clear separation with only a few misclassifications. However, some confusion was observed between "Analytical" and "Cautionary." This overlap may stem from the fact that cautionary messages often use structured reasoning to warn or advise, a style that closely mirrors the logical tone of analytical responses. Additionally, there were minor misclassifications between "Gentle" and "Analytical," as well as between "Gentle" and "Inspirational." These categories may share similar emotional qualities, such as warmth or encouragement, which can lead to difficulty in differentiating them based solely on vocabulary and tone.

During this phase, the model showed steady improvements in both training and validation performance. The training dynamics are visualized in Figure 6, demonstrating a consistent drop in loss and increase in accuracy peaked at Epoch 5.

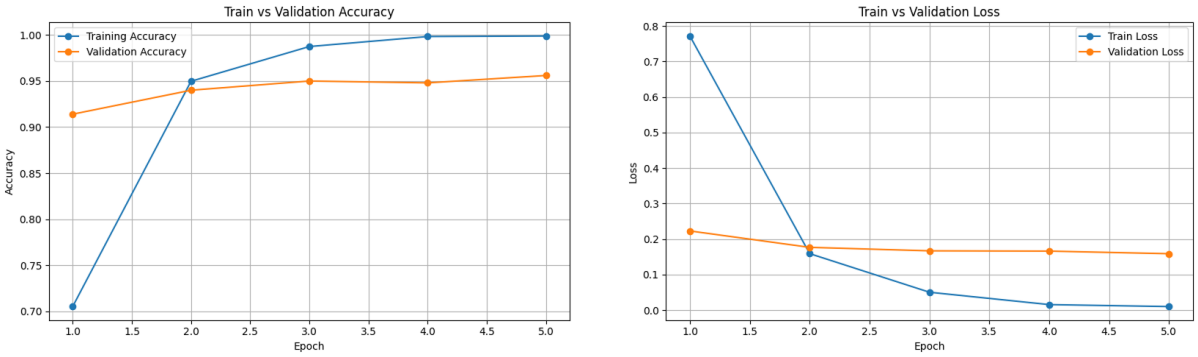

Figure 6. Accuracy And Loss Plots On Phase 1.

The training process showed stable convergence across all Epochs, with improvements in both accuracy and loss values. The performance peaked at Epoch 5, where the model achieved a training accuracy of 99.8%, validation accuracy of 95.6%, test accuracy of 92.8%, training loss of 0.0098, and validation loss of 0.1587. Since the lowest validation loss was reached at Epoch 5, the model from this Epoch was saved as the best performing version and used for further inference and evaluation.

These results suggest that the model is highly capable of learning emotion categories when trained on clean, manually verified data. The perfect classification of the 'Enthusiastic' class indicates clear lexical cues for that emotion, while slightly lower performance on 'Analytical' implies more subtle or overlapping linguistic patterns that challenge the model's ability to distinguish them.

3.3. Inference on 12,500 Unlabeled Data

Following the initial fine-tuning phase, the best performing IndoRoBERTa Base model was used to perform inference on the remaining 12,500 unlabeled question-response pairs. This process aimed to automatically assign emotion labels to the entire dataset, enabling the construction of a fully annotated corpus for further training and analysis. The inference was conducted in evaluation mode to ensure no model weights were updated during prediction.

Table 10. Emotion Label Distribution after Inference on 12,500 Samples

Label	Count	Percentage
Enthusiastic	3,700	29.6%
Gentle	2,019	16.1%
Analytical	2,375	19%
Inspirational	1,958	15.6%
Cautionary	2,448	19.6%

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

The distribution of predicted emotion labels is summarized in Table 10. As expected, a degree of imbalance is observed, with the Enthusiastic and Cautionary categories appearing more frequently. This skew reflects the generative tendencies of the language model used to create the original text samples, which may naturally favor certain emotional tones such as optimism or reflection. No post inference balancing was applied, as the purpose was to preserve the distributional characteristics of real-world-like generated data, which will be leveraged in the second training phase.

The label distribution highlights a noticeable pattern in how the language model tends to respond, it leans toward producing messages that are emotionally rich or reflective in tone. This built-in style preference points to a possible imbalance in the data. If the goal is to develop a model that maintains a more neutral emotional tone or adjusts its emotional expression based on different contexts, then additional steps may be needed. These could include rebalancing the dataset or introducing new examples that promote a wider range of emotional responses.

3.4. Training Phase 2: Fine-Tuning on Auto Labeled Dataset

In the second training phase, the model was fine-tuned on a significantly larger dataset of 12,500 automatically labeled samples, which were generated by the previously trained IndoRoBERTa model. Given the increased data volume and label imbalance in this phase, the hyperparameters were slightly adjusted to ensure training efficiency while preserving generalization. The learning rate, optimizer, and batch size remained consistent with Phase 1. However, the number of training Epochs was reduced from 5 to 3 to avoid prolonged overfitting caused by the imbalanced data. Table 11 summarizes the key hyperparameters used during this phase.

Table 11. Hyperparameter configuration for Training phase 2.

Hyperparameter	Value
Base model	IndoRoBERTa Base
Learning Rate	0 - 3e-5(with scheduler)
Optimizer	AdamW
Epochs	3
Batch Size	8
Max sequence length	128 tokens
Loss function	CrossEntropyLoss
Early stopping	Patience = 2 Epochs

Table 12. Performance Metrics from Test Set.

Label	Accuracy	Precision	Recall	F1-Score
Enthusiastic	0.94	0.98	0.94	0.96
Gentle	0.88	0.88	0.88	0.88
Analytical	0.93	0.90	0.93	0.92
Inspirational	0.90	0.88	0.90	0.89
Cautionary	0.90	0.88	0.90	0.89

Table 12 presents the metrics from Test Set. the model achieved 97.4% training accuracy, 89.9% validation accuracy and 91.3% test accuracy, with a training loss of 0.0798 and a validation loss of 0.2771. Since Epoch 3 yielded the best validation performance, it was selected as the final checkpoint for downstream tasks. To support detailed analysis, Figure 7 displays the training and validation accuracy or loss progression across Epochs.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

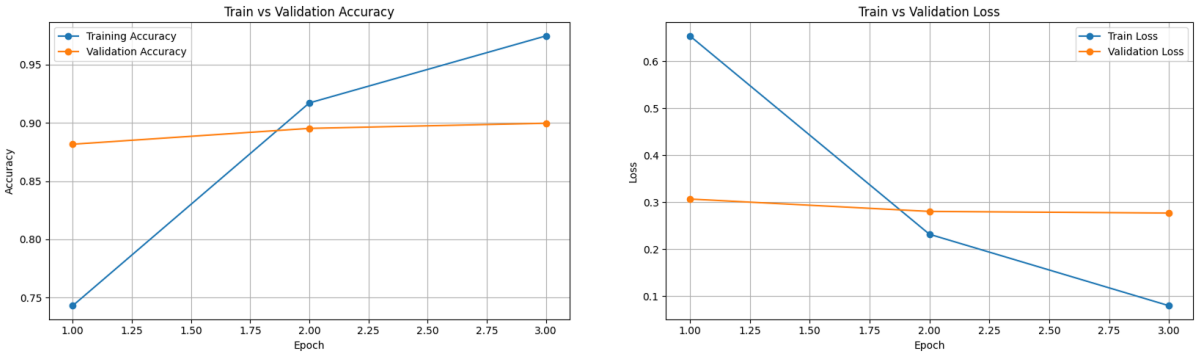

Figure 7. Accuracy And Loss Plots of Proposed Model.

To further assess classification behavior during testing, a confusion matrix was generated for the test set in Phase 2, as shown in Figure 8.

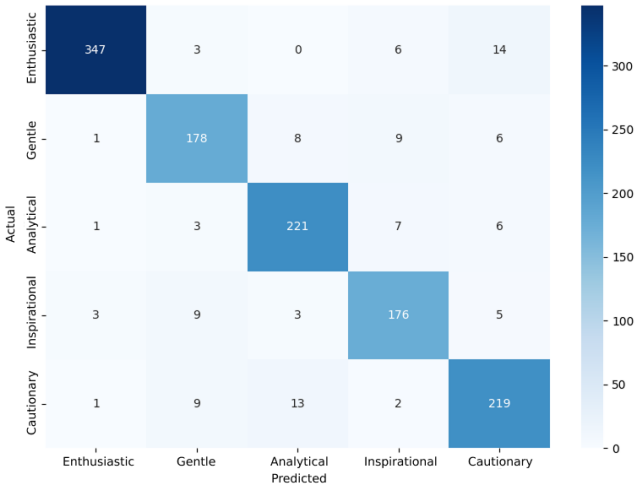

Figure 8. Confusion Matrix on Phase 2.

The confusion matrix presented in Figure 8 shows a noticeable increase in misclassifications compared to what was observed in Phase 1. This rise is likely attributed to label noise that entered the dataset during the auto-labeling process. One of the most common areas of overlap appeared between the ‘Cautionary’ and ‘Inspirational’ categories. Even though these two are conceptually different, they sometimes share similar patterns, such as referencing the future or using directive language, which can blur their boundaries and make interpretation less clear.

There were also moments when responses labeled as ‘Gentle’ were mistaken for ‘Analytical.’ This may be due to the presence of calm, thoughtful language, thoughtful expressions like “mungkin perlu dipikirkan baik-baik” or “It might be worth considering carefully” in English, which can resemble analytical reasoning, even when the actual intent is to provide emotionally supportive responses. These patterns of confusion highlight a broader challenge in emotion classification for counseling contexts, where emotional tone and communicative intent often intertwine in nuanced and context-dependent ways.

The final model was taken from Epoch 3, which yielded the most favorable results in terms of validation accuracy and loss. Although early stopping was enabled, it was not activated during training, indicating that the model had already achieved optimal performance by the end of the scheduled training. This outcome suggests that the chosen number of Epochs was sufficient to maximize model performance without leading to overfitting.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Despite being trained on noisy, auto-labeled data, the model maintained strong performance across all categories. This demonstrates its generalization ability, though the slightly reduced precision and recall compared to Phase 1 indicate the effect of label noise and imbalance. These factors may influence deployment contexts that require high sensitivity for less frequent emotional tones such as 'Gentle' or 'Inspirational'.

3.5. Gesture Mapping

To broaden the scope of the emotion classification model toward building a multimodal virtual counseling agent, this study introduces a gesture mapping component that links emotional labels to expressive physical behaviors. Each of the five functional emotions 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary' was associated with a specific gesture profile designed to reflect its underlying communicative intent. These associations, as previously detailed in Section 2.6, were developed during the system design phase and grounded in principles from counseling psychology and nonverbal communication.

Figure 9. Emotion Visualization Through Distinct Body Gesture Representations.

Figure 9 presents visual examples of these mapped gestures, using full-body character renderings staged in controlled, studio-style poses. The emphasis lies in posture, facial expression, and hand movement each tailored to convey the emotion's tone. For example, Enthusiastic is depicted through open gestures, animated hand motion, and a lively facial expression, while Analytical features a composed posture and attentive gaze.

Unlike the emotion labels, which are predicted automatically by the model, these gesture assignments were defined manually. They serve as a conceptual bridge toward future system upgrades, where emotional understanding could be expressed not only through text but also through embodied, nonverbal cues in animated agents. Although real-time gesture rendering is not yet integrated, the mappings offer a structured foundation for building more emotionally expressive and humanlike chatbot systems in the future.

4. DISCUSSION

The findings from this study confirm that the fine-tuned IndoRoBERTa model is highly effective in identifying functional emotional expressions within early marriage consultation dialogues. In the first training phase, where the model was trained on a balanced dataset of 2,500 manually validated samples were used for fine-tuning, resulting in a validation accuracy of 95.6% and a test accuracy of 92.8%. This result outperformed the previously reported 92.3% accuracy from the IndoBERT model [11]. The high performance in Phase 1 can be credited to the clean, evenly distributed dataset, which allowed the model to clearly differentiate between the five emotion categories.

In contrast, Phase 2 introduced a broader and more complex challenge. The model was retrained on an expanded dataset of 12,500 auto-labeled samples generated through inference from the phase one

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

model, which brought the validation accuracy to 89.9% and the test accuracy to 91.3%. As expected, the validation and test accuracy dropped to 89.9% and 91.3% due to the increased noise and imbalance in the data. Still, this phase demonstrated stronger generalization, with the model maintaining solid classification performance across all emotion categories.

These results suggest that while Phase 1 offered precise learning under controlled conditions, Phase 2 allowed the model to adapt to more diverse and realistic user inputs, mirroring the variability present in actual online counseling scenarios. To provide a more detailed performance comparison, Table 13 presents the overall accuracy of all models.

Table 13. Accuracy Comparison Across Models.

Model	Accuracy	Dataset Size	Emotion Labels
IndoRoBERTa Phase 1	92.8%	2,500	5 Functional Emotions
IndoRoBERTa Phase 2	91.3%	12,500	5 Functional Emotions
IndoBERT	92.3%	7,629	6+1 Basic Emotions (Ekman)

Table 13 offers a direct comparison of results between both training phases and the IndoBERT benchmark, highlighting that the approach taken in this study despite using a different emotional framework delivers comparable, if not superior, outcomes. Additionally, the integration of gesture mapping into the emotion classification pipeline presents a novel step toward multimodal chatbot interaction, an area not previously addressed in related research.

In addition to emotion classification, the system integrates a gesture recommendation module linking each emotional label to corresponding body language, as shown in Table 3. This multimodal enhancement advances emotionally responsive chatbot design, particularly for counseling. For instance, ‘Enthusiastic’ responses are matched with raised eyebrows and bright eyes to convey energy, while ‘Gentle’ replies use soft smiles and relaxed gaze to evoke empathy. Aligning verbal and nonverbal elements aims to build trust and emotional resonance.

However, some limitations remain. Gesture interpretation is subjective and shaped by personal or cultural factors. The five functional labels, while practical, may not fully capture emotional nuance across diverse Indonesian communities. Additionally, the auto-labeling process depends on the Phase 1 model’s accuracy, introducing risks of misclassification over time despite human review. Moreover, the use of LLM-generated labels during the auto-annotation phase may introduce latent biases reflective of the language model's training data, which could affect the objectivity of emotional classification. Despite these challenges, the system holds promise in the Indonesian context, where early marriage remains a concern. It offers a private, empathetic space for youth to reflect and engage. With careful ethical oversight, such tools could complement formal counseling and awareness efforts, not replace them.

The emotion framework is based on functionalist theories emphasizing the communicative roles of emotion rather than universal facial expressions. Recent studies in text-based appraisal modeling and multimodal taxonomy support this perspective, showing that context and purpose improve recognition accuracy and relevance [32]. Additionally, empirical studies show that integrating multimodal features such as gestures, visuals, and audio into chatbot systems leads to greater user engagement and prolonged interaction durations [33].

Future directions may include fine-grained comparative learning strategies to improve the model's capacity for nuanced emotional classification. Real-time interaction studies will also be valuable, helping to fine-tune the system’s responsiveness. Priorities for development include expanding gesture personalization, improving cultural flexibility, and adding support for multiple languages to make the system more widely usable across counseling environments.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

5. CONCLUSION

This study demonstrates that a fine-tuned IndoRoBERTa model performs well in classifying functional emotions within early marriage counseling dialogues and lays the foundation for building more emotionally responsive virtual agents through gesture mapping. Across two training phases, the model achieved 95.6% accuracy in Phase 1 and 89.9% in Phase 2. Despite a drop in accuracy during Phase 2 due to a more diverse, LLM-generated dataset, the increased variability helped improve the model's ability to handle real-world conversations. Strong precision, recall, and F1-scores were observed across all five emotion categories: 'Enthusiastic', 'Gentle', 'Analytical', 'Inspirational', and 'Cautionary'.

Notably, this framework outperformed previous models such as IndoBERT, which had a 92.3% accuracy rate, by balancing clean initial training with broader generalization. Gesture mapping was also introduced, not simply as an added feature, but as a theory-based layer linking emotional text with expressive behavior. Each emotional category was paired with a specific gesture, paving the way for chatbots that can communicate both verbally and physically.This study provides a new benchmark in function-based emotional NLP for low-resource languages, particularly in the context of culturally sensitive counseling dialogues. In addition to providing a labeled dataset and a targeted emotion classification system for Indonesian counselling, it also introduces a gesture design model to facilitate the integration of emotionally expressive virtual agents.

However, the methodology presents certain limitations that warrant attention. Notably, the reliance on synthetic data generated by large language models during the second phase of training introduces potential discrepancies. These artificial data points may lack the depth, ambiguity, and emotional variability inherent in actual human interactions. Consequently, the ecological validity of the system could be compromised if not further validated. Furthermore, future directions may include real-time gesture generation, multimodal input support (e.g., voice or facial cues), and mobile or web platform deployment, each offering new possibilities for enhancing adolescent mental health tools and early marriage awareness efforts.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest related to the conduct of this study. All research activities, including data generation, annotation, model training, and evaluation, were carried out independently and without any affiliation, sponsorship, or influence from external institutions, government bodies, or organizations involved in early marriage or public policy in the region. The chatbot and datasets used in this study were created solely for academic and research purposes, and no personal, political, or commercial interests were associated with the object or outcome of this research.

REFERENCES

[1] R. Nabila, R. Roswiyani, and H. Satyadi, "A Literature Review of Factors Influencing Early Marriage Decisions in Indonesia," in *Proceedings of the 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021)*, Atlantic Press, 2022, pp. 1392–1402. doi: 10.2991/assehr.k.220404.223.

[2] D. Fadilah, "Tinjauan Dampak Pernikahan Dini dari Berbagai Aspek," *Pamator Journal*, vol. 14, no. 2, pp. 88–94, Nov. 2021, doi: 10.21107/pamator.v14i2.10590.

[3] Z. Ayudiputri, A. Nur, S. Amanda, and F. Hanifa, "Determinants of Child Marriage in Indonesia : A Systematic Review," *Journal of Community Medicine and Public Health Research*, vol. 5, no. 2, pp. 216–227, Nov. 2024, doi: 10.20473/jcmphr.v5i2.45777.

[4] L. Wang, D. Wang, F. Tian, Z. Peng, X. Fan, Z. Zhang *et al.*, "CASS: Towards Building a Social-Support Chatbot for Online Health Community," in *Conference on Computer-Supported*

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Cooperative Work & Social Computing (CSCW), Feb. 2021, pp. 1–31. doi: https://doi.org/10.48550/arXiv.2101.01583.

[5] S. Khandelwal, "SOCIAL COMPANION CHATBOT FOR HUMAN COMMUNICATION USING ML AND NLP," *International Journal of Engineering Applied Sciences and Technology*, vol. 8, pp. 321–324, 2023, doi: https://doi.org/10.33564/IJEAST.2023.v08i01.048.

[6] R. E. Guingrich and M. S. A. Graziano, "Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines," in *Oxford Intersections: AI in Society*, Oxford University Press, 2025. doi: https://doi.org/10.1093/9780198945215.001.0001.

[7] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 728–735. doi: 10.1016/j.procs.2021.01.061.

[8] N. Hilmiaji, K. M. Lhaksmana, and M. D. Purbalaksono, "Identifying Emotion on Indonesian Tweets using Convolutional Neural Networks," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 584–593, Jun. 2021, doi: 10.29207/resti.v5i3.3137.

[9] N. G. Ramadhan, "Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1083–1089, Dec. 2021, doi: 10.29207/resti.v5i6.3547.

[10] M. I. K. Sinapoy, Y. Sibaroni, and S. S. Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, pp. 657–662, Jun. 2023, doi: 10.29207/resti.v7i3.4830.

[11] 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.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, vol. abs/1907.11692, Jul. 2019, doi: https://doi.org/10.48550/arXiv.1907.11692.

[13] Y. O. Sihombing, R. F. Rachmadi, S. Sumpeno, and Moh. J. Mubarok, "Optimizing IndoRoBERTa Model for Multi-Class Classification of Sentiment & Emotion on Indonesian Twitter," in *Proceeding - IEEE 10th Information Technology International Seminar, ITIS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 12–17. doi: 10.1109/ITIS64716.2024.10845566.

[14] T. Widarmanti, M. P. Widodo, D. P. Ramadhani, and M. Danlami, "Text Emotion Detection: Discover the Meaning Behind YouTube Comments Using Indo RoBERTa," in *ICACNIS 2022 - 2022 International Conference on Advanced Creative Networks and Intelligent Systems: Blockchain Technology, Intelligent Systems, and the Applications for Human Life, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022, p. 1. doi: 10.1109/ICACNIS57039.2022.10055265.

[15] F. M. Plaza-Del-Arco, A. Curry, A. C. Curry, and D. Hovy, "Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 5696–5710. doi: 10.48550/arXiv.2403.01222.

[16] A. Koufakou and E. Nieves, "Review of recent emotion-annotated text corpora and resources," *Lang Resour Eval*, pp. 1–35, Jun. 2025, doi: 10.1007/s10579-025-09828-1.

[17] E. (Grace) Park, "I Trust You, but Let Me Talk to AI: The Role of the Chat Agents, Empathy, and Health Issues in Misinformation Guidance," *International Journal of Strategic Communication*, vol. 19, no. 2, pp. 231–260, Mar. 2025, doi: 10.1080/1553118X.2025.2462087.

[18] Y. Li, K. Li, H. Ning, X. Xia, Y. Guo, C. Wei *et al.*, "Towards an Online Empathetic Chatbot with Emotion Causes," in *Proceedings of the 44th International ACM SIGIR Conference on*

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

Research and Development in Information Retrieval, New York, NY, USA: ACM, Jul. 2021, pp. 2041–2045. doi: 10.1145/3404835.3463042.

[19] H. Li, G. K. Rajbahadur, D. Lin, C.-P. Bezemer, Z. Ming, and Jiang, "Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting," vol. 12, pp. 70676–70689, Jan. 2024, doi: 10.1109/ACCESS.2024.3402543.

[20] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[21] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, and Y. F. Zhao, "A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management," *J Intell Manuf*, vol. 34, no. 8, pp. 3305–3340, Dec. 2023, doi: 10.1007/s10845-022-02017-9.

[22] Y. Li, X. Ren, F. Zhao, and S. Yang, "A Zeroth-Order Adaptive Learning Rate Method to Reduce Cost of Hyperparameter Tuning for Deep Learning," *Applied Sciences*, vol. 11, no. 21, p. 10184, Oct. 2021, doi: 10.3390/app112110184.

[23] J. S. Hwang, S. S. Lee, J. W. Gil, and C. K. Lee, "Determination of Optimal Batch Size of Deep Learning Models with Time Series Data," *Sustainability (Switzerland)*, vol. 16, no. 14, Jul. 2024, doi: 10.3390/su16145936.

[24] S. Ahn, S. Kim, J. Ko, and S.-Y. Yun, "Fine tuning Pre trained Models for Robustness Under Noisy Labels," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, (IJCAI-24)*, International Joint Conferences on Artificial Intelligence Organization, Oct. 2023, pp. 3643–3651. doi: https://doi.org/10.48550/arXiv.2310.17668.

[25] 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*, International Committee on Computational Linguistics, Nov. 2020, pp. 757–770. doi: https://doi.org/10.48550/arXiv.2011.00677.

[26] W. Stigall, M. A. Al Hafiz Khan, D. Attota, F. Nweke, and Y. Pei, "Large Language Models Performance Comparison of Emotion and Sentiment Classification," in *Proceedings of the 2024 ACM Southeast Conference, ACMSE 2024*, Association for Computing Machinery, Inc, Apr. 2024, pp. 60–68. doi: 10.1145/3603287.3651183.

[27] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Transactions of the Association for Computational Linguistics 2024*, vol. 12, pp. 820–836, Apr. 2024, doi: 10.1162/tacl_a_00675.

[28] T. Schlosser, M. Friedrich, T. Meyer, D. Kowerko, and J. Professorship, *A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision*. 2024. doi: 10.13140/RG.2.2.14331.69928.

[29] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[30] S. Jiang, J. Li, Y. Wang, B. Huang, Z. Zhang, and T. Xu, "Delving into Sample Loss Curve to Embrace Noisy and Imbalanced Data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, Dec. 2021, pp. 7024–7032. doi: https://doi.org/10.48550/arXiv.2201.00849.

[31] A. Diwan, R. Sunil, P. Mer, R. Mahadeva, and S. P. Patole, "Advancements in Emotion Classification via Facial and Body Gesture Analysis: A Survey," *Expert Syst*, vol. 42, no. 2, p. e13759, 2025, doi: https://doi.org/10.1111/exsy.13759.

[32] J. Hofmann, E. Troiano, K. Sassenberg, and R. Klinger, "Appraisal Theories for Emotion Classification in Text," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain: International Committee on Computational Linguistics, Nov. 2020, pp. 125–138. doi: doi.org/10.48550/arXiv.2003.14155.

Vol. 6, No. 5, October 2025, Page. 3729-3749
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5038

[33] L. Zhang, J. Yu, S. Zhang, L. Li, Y. Zhong, G. Liang *et al.*, "Unveiling the Impact of Multi-Modal Interactions on User Engagement: A Comprehensive Evaluation in AI-driven Conversations," *CoRR*, vol. abs/2406.15000, Jun. 2024, doi: 10.48550/arXiv.2406.15000.