OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
journals.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 10, Issue. 1, August 2025. pp. 72-87
doi:10.21108/indojc.v10i1.9708

Speech to Text Correction for Indonesian Early Marriage Counseling Chatbots Using IndoRoBERTa and Mistral-7B

Firdhaus Dwi Sukma[1], Rifki Wijaya[2], Ade Romadhony[3]

School of Computing, Telkom University
Telecommunication Street No. 1, Bandung 40257, Indonesia

[1] usdhauss@student.telkomuniversity.ac.id
[2] rifkiwijaya@telkomuniversity.ac.id
[3] aderomadhony@telkomuniversity.ac.id

Abstract

Early marriage among individuals of immature age continues to draw significant attention in Lombok. As of 2021, the prevalence rate stands at 16.59%, indicating that this social issue remains unresolved within the region's community dynamics. Limited access to counseling services particularly in rural areas poses a significant barrier to prevention efforts. This study introduces a virtual counseling chatbot designed to detect and correct Indonesian language text errors during user interactions. The system integrates IndoRoBERTa for error detection and Mistral-7B-Instruct to refine speech to text transcriptions. IndoRoBERTa was trained on synthetic datasets to classify user input as accurate or incorrect, while Mistral-7B-Instruct generates context aware corrections. Achieving an accuracy rate of 98.90%, IndoRoBERTa outperformed benchmark models such as BERT and RNN. The proposed chatbot offers an adaptive and accessible digital solution, especially for communities with limited access to conventional counseling services. This approach highlights the potential of AI-driven tools to support early intervention strategies and reduce the incidence of child marriage in underserved regions.

Keywords: Early Marriage, Virtual Counseling, IndoRoBERTa, Mistral-7B-Instruct, Speech to Text

Abstrak

Pernikahan dini pada usia yang belum matang terus menjadi sorotan di Lombok, dengan catatan prevalensi sebesar 16,59% pada tahun 2021. Angka ini mencerminkan bahwa isu tersebut belum sepenuhnya teratasi dalam dinamika sosial masyarakat setempat. Terbatasnya layanan konseling, khususnya di daerah pedesaan, menjadi hambatan serius dalam upaya pencegahan. Studi ini merancang chatbot konseling virtual yang mampu mendeteksi dan memperbaiki kesalahan teks bahasa Indonesia selama percakapan. Sistem ini mengintegrasikan IndoRoBERTa untuk deteksi kesalahan dan Mistral-7B-Instruct untuk koreksi transkripsi dari speech to text. IndoRoBERTa dilatih menggunakan data sintetis untuk mengidentifikasi masukan yang tidak tepat, sementara Mistral-7B-Instruct menghasilkan koreksi kontekstual. Dengan tingkat akurasi untuk IndoRoBERTa mencapai 98,90%, model ini melampaui kinerja BERT dan RNN. Chatbot ini diharapkan menjadi solusi digital yang adaptif, terutama bagi masyarakat dengan keterbatasan akses terhadap layanan konseling konvensional. Pendekatan ini membuka peluang baru dalam pemanfaatan AI untuk mendukung intervensi dini dan menekan angka pernikahan anak di daerah terpencil.

Kata Kunci: Pernikahan Dini, Virtual Konseling, IndoRoBERTa, Mistral-7B-Instruct, Speech to Text

Received on January, 2025. Accepted on August, 2025

I. Introduction

EARLY marriage is a persistent issue in Indonesia, and it's especially widespread in the Lombok region, where it's a deeply rooted part of the local culture. The practice is still happening, even with many new rules and policies, because it's deeply rooted in old traditions and how the community sees things. Even with the continuous progress in digital technology, early marriage continues to be a major social concern, especially in rural regions like Lombok. With a staggering prevalence rate of 16.59% in 2021, Lombok was positioned as one of the regions in Indonesia with a concerningly high incidence of early marriage [1]. Looking at 2024 data from Statistics Indonesia (BPS), West Nusa Tenggara stands out for its high rate of early marriage. Specifically, nearly 15% (14.96%) of women there who are now between the ages of 20 and 24 were already married or living with a partner before they were 18. Early marriage is a significant risk for adolescent girls in rural communities, largely because conservative social norms, stigma, and discrimination prevent them from accessing vital reproductive health information [2].

Through counseling, young people learn about the risks of early marriage, which helps them make more thoughtful decisions about their lives[3], [4]. The practical value of conventional counseling diminishes significantly in rural settings due to three main constraints: underdeveloped infrastructure, financial barriers posed by high costs, and geographical disparities in service access [5], [6]. The culmination of these difficulties has meant that many adolescents lack the critical support and educational grounding they otherwise would have obtained.

To meet the growing need for safer communication channels, we've developed an AI powered virtual assistant. By leveraging Natural Language Processing (NLP), it can grasp user intent and respond in a way that feels natural and relevant. With features like round-the-clock availability and strong privacy safeguards, this tool fosters a more inclusive and open space for dialogue. There's a recurring issue with many transcription tools, they tend to miss the little things subtle mishearings, awkward phrasing, those quiet mistakes that slip by unnoticed. This study steps into that gap. By cleaning up those oversights automatically, the transcriptions don't just look neater they feel more natural to read. And that makes all the difference when people are trying to follow along or engage with spoken content.

Transformer based architectures have become the cornerstone of natural language processing. A notable leap was made with RoBERTa, which refined the pre-training strategy and led to significant performance gains. IndoRoBERTa, a language model designed for Indonesian, performs exceptionally well when classifying sentiment and emotion in social media data [7]. However, not much research has focused on applying this to a narrower domain, specifically identifying linguistic inaccuracies in Indonesian texts. In this study, a special prompt was used to guide the Mistral-7B-Instruct language model in correcting errors. The goal of this prompt was to get the model to revise questions from teenagers about early marriage, adapting a style that a psychological counselor would use.

This research uses the IndoRoBERTa model to find and correct language errors in Indonesian transcripts. It specifically focuses on questions that come up during counseling sessions for early marriage. The classification centers on two categories texts containing language inaccuracies and those that adhere to proper linguistic conventions. The types of errors analyzed include misspellings, inconsistent capitalization, and ambiguous sentence structures. Complementing this, the Mistral-7B-Instruct model is employed to automatically correct erroneous texts using contextually tailored prompts. The findings are expected to serve as a foundation for developing more natural and effective Indonesian text correction systems, while also supporting the creation of empathetic and linguistically intelligent virtual counseling agents.

Within this framework, the study is anchored by several guiding questions that shape the direction of the analysis. First, what kinds of linguistic errors are most found in Indonesian speech-to-text transcriptions,

particularly within the context of early marriage counseling? Second, how accurately can the IndoRoBERTa model identify these errors? And third, to what degree is the Mistral-7B-Instruct model capable of correcting such errors in a way that reflects not only linguistic precision but also empathetic and context-sensitive communication? Overall, the study highlights the potential of IndoRoBERTa and Mistral in specialized linguistic domains and lays the groundwork for future multimodal counseling systems in Indonesia.

II. Literature Review

Research on error detection in Indonesian texts has been extensively conducted, including a study by Nihalani and colleagues, which employed a BERT-based model. The study achieved an accuracy rate of 90.53%. Despite this promising result, the model still encounters key challenges, such as overfitting, a strong reliance on data quality, and inefficiencies in consistently identifying grammatical errors [8]. Another study conducted by Zhenhui He and his team explored the application of Recurrent Neural Networks (RNN) for detecting errors in English text. However, the method demonstrated a relatively modest accuracy rate of approximately 77% and was tested only on a limited-scale dataset. Moreover, the approach struggled to handle the complexity of grammatical structures and remained heavily reliant on rule-based detection techniques [9].

Another study by Evi Yulianti focuses on the application of Named Entity Recognition (NER) in Indonesian legal documents, assessing the performance of various transformer-based models, with IndoRoBERTa being one of the key benchmarks. Interestingly, the findings reveal that RoBERTa models tailored for the Indonesian language such as IndoRoBERTa outperformed the BiLSTM-CRF model in terms of F1-score, with a performance margin ranging from 2.1% to 2.6%, and achieving a peak F1-Score of 92.46% [10]. These results indicate that the transformer-based architecture offers a more adaptive and effective approach to processing Indonesian text.

This study adds to the growing research on how we can better spot text errors, especially with new methods for both Indonesian and English. Historically, researchers have often used reliable language models like BERT and Recurrent Neural Networks (RNNs) because they've consistently proven to be effective for various language-related tasks. Moreover, the application of the IndoRoBERTa model to Named Entity Recognition (NER) tasks in Indonesian legal texts has shown promising outcomes. Not only does it exhibit a strong ability to capture critical entities embedded in native linguistic structures, but it also adapts effectively to the nuanced nature of legal language, which often poses challenges for conventional models. This study turns its attention to the application of the IndoRoBERTa model, exploring its viability as either a substitute for or a supplement to existing approaches. This study seeks to explore how effectively IndoRoBERTa identifies linguistic errors in Indonesian texts compared to other existing models. Rather than merely benchmarking raw performance, the investigation emphasizes practical efficiency and accuracy, aiming to determine which model demonstrates the most reliable results in real-world language processing scenarios.

A. Early Marriage

Early marriage, where a boy and a girl marry as minors, is a tradition often dictated by family custom and the prevailing cultural and social norms of a community [11], [12], [13].

B. Indonesian Linguistic Norms and Common Errors

The Indonesian language operates under a standardized orthographic framework known as Ejaan Yang Disempurnakan (EYD), officially referred to today as the Pedoman Umum Ejaan Bahasa Indonesia (PUEBI). This system plays a crucial role in regulating the use of capitalization, punctuation, word formation, and overall sentence construction, thereby ensuring consistency and precision in written communication [14], [15]. Standard Indonesian typically adheres to a Subject-Predicate-Object (SPO) syntactic structure. The consistent application of Enhanced Spelling System (EYD) conventions not only fosters clarity of meaning but also strengthens the overall cohesion and coherence of the text [16].

The inherent divergence of natural conversation from standard pronunciation and grammar, marked by clipped words and mixed phrasing, often compromises the accuracy of automated transcription. Phonetic errors, such as the incorrect transcription of 'bantu' to 'bantuk', exemplify this problem. Consequently, there is a clear demand for text correction technologies that surpass the capabilities of conventional spell-checking. An effective system must be engineered to not only interpret phonetic discrepancies but also to process the dynamic and informal slang characteristic of youth language.

C. Psycholinguistic Characteristics of Adolescent Speech

Adolescent speech is marked by unique psycholinguistic characteristics that often diverge from conventional linguistic norms. Typically, their expressions are informal, emotionally charged, fragmented, and heavily reliant on contextual cues [18], [19]. Departing from formal grammar, teenage communication is characterized by its brevity and informality. The frequent omission of subjects or verbs, coupled with the use of syntactically dense slang, results in significant deviations in sentence structure and discourse. These linguistic anomalies present a substantial obstacle for automated language processing systems attempting to parse such fluid speech patterns.

Teens are great at adapting their language. In casual chats with friends, they'll use informal expressions that fit their social scene. But when they're talking to an adult, like a parent or teacher, they can easily switch to more formal Indonesian. From a psychological perspective, this isn't a mistake it’s a smart communication strategy that changes with the situation. This means that any language corrections we make need to be handled carefully. We shouldn't just focus on grammar, but also on preserving the emotion and sincerity in what they say, especially in sensitive situations like counseling on early marriage, where empathy is just as important as being grammatically correct.

D. IndoRoBERTa

As an extension of RoBERTa [21], IndoRoBERTa was specifically trained on Indonesian-language textual data to enhance the model’s accuracy in capturing the unique semantic and contextual nuances of the Indonesian language. IndoRoBERTa was developed by pretraining the RoBERTa architecture on a large-scale Indonesian-language dataset. The training corpus was sourced from OSCAR [22], comprising approximately 17.05 GB of text. This model retains the original RoBERTa search architecture and consists of around 124 million parameters, offering robust performance in capturing the linguistic nuances of the Indonesian language.

E. Mistral-7B-Instruct

Mistral 7B stands out for its impressive efficiency, delivering powerful performance from a lean 7-billion-parameter framework. By employing advanced techniques like Grouped-Query Attention (GQA) and Sliding Window Attention (SWA), it accurately solves complex challenges without requiring extensive computational resources. The fine-tuned Mistral-7B-Instruct version builds on this success, surpassing even larger models such as LLaMA 2-13B in specialized areas. It particularly excels in mathematical and coding tasks, a direct result of its focused training on public instructional data. Notably, the developers made a conscious decision to exclude any proprietary or sensitive data from the training process. Evaluation results demonstrate its impressive capabilities surpassing all other 7B models and performing on par with 13B models in MT-Bench tests and human preference assessments [23]. Equipped with prompt engineering support and built-in moderation mechanisms, this model is well-suited for real-world applications that demand both safety and operational control.

Therefore, the Mistral-7B-Instruct model can be utilized directly without the need for fine-tuning. Because it was fine-tuned on publicly available instruction datasets, Mistral-7B–Instruct emerges as a highly capable model surpassing all other 7B instruction-following counterparts and even rivaling the performance of larger 13B-scale chat models such as LLaMA 2–13B Chat [23]. Nevertheless, to ensure its effectiveness in identifying and correcting errors in Indonesian texts particularly those related to the issue of early marriage it is essential to construct prompts carefully so that the model can interpret the context of the questions accurately.

III. Research Method

This study developed a systematic approach focused on leveraging the IndoRoBERTa and Mistral-7B-Instruct models to facilitate the detection and correction of errors in Indonesian-language texts. This phase emphasizes the development of a comprehensive strategy, outlining a series of deliberate steps designed not only to construct the system but also to establish a structured framework for its gradual evaluation. The entire process is methodically structured into several stages, as illustrated in the block diagram in Fig. 1, to ensure a coherent and well-organized development workflow.

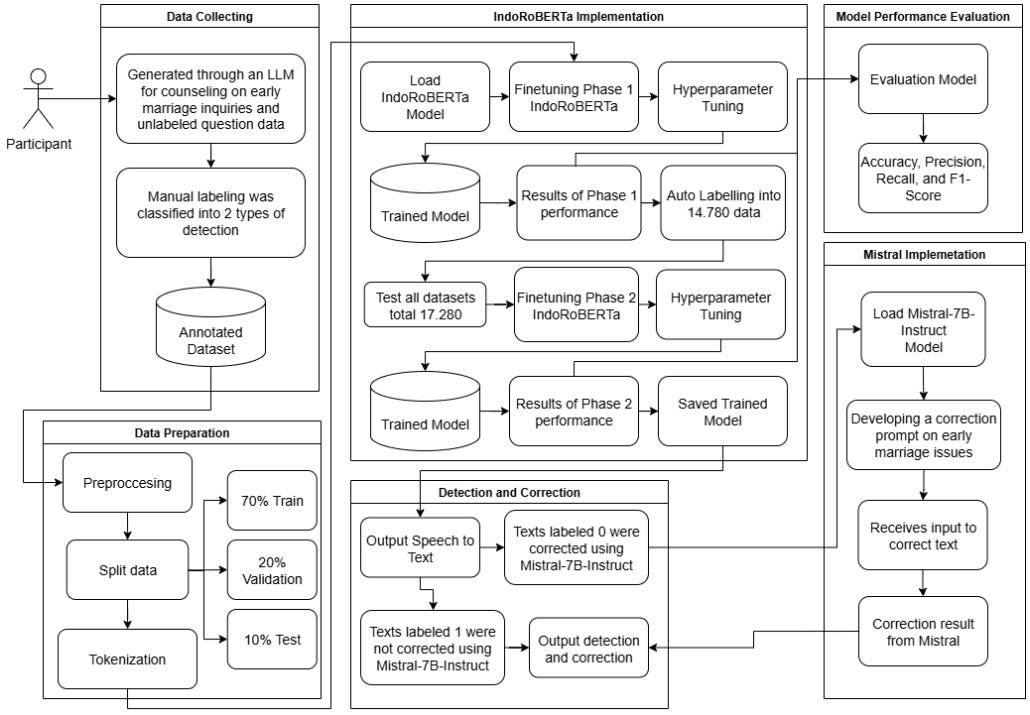

Fig. 1. System flow block diagram

A. Data Collection

In this study, data collection was conducted using Large Language Models (LLMs), which were employed to generate a series of questions articulated from the perspectives of children and adolescents who are either curious about or contemplating early marriage. The workflow for this process is illustrated in Fig 1. These questions were deliberately crafted to reflect a diverse range of backgrounds, encompassing economic, social, religious, cultural, and educational contexts that may shape the respondents' viewpoints. Given the cultural sensitivity surrounding early marriage in Lombok, all outputs generated by the LLM underwent manual evaluation by the research team. This review aimed to ensure that the questions produced were both contextually appropriate and reflective of realistic counseling scenarios, while remaining respectful of local socio-cultural values.

In total, the data collection process yielded 17.280 entries. From this dataset, the researchers manually annotated a subset of 2.500 entries. For each question in this sample, they assigned labels based on its accuracy and language use classifying whether it conveyed correct information, was purposefully misleading, or exhibited linguistic flaws when interpreted in the Indonesian context. The remaining 14.780 entries were not manually labeled instead, they were automatically annotated using a model previously trained on the manually labeled subset.

In this research, we introduce a detection system aimed at sorting textual inputs into two primary classifications. The initial class, marked as 0 (Incorrect), encompasses Indonesian text segments that exhibit linguistic irregularities these may include typographical slips, inconsistent use of capital letters, or sentence constructions that appear vague or challenging to decode. On the other hand, label 1 (Correct) is attributed to sentences that align with established linguistic norms and display no notable issues in terms of grammar or sentence construction. This classification approach facilitates a more systematic and nuanced evaluation of language quality, particularly within the domain of natural language processing. An example of the dataset can be seen in TABLE I.

TABLE I
EXAMPLE OF QUESTION DATA WITH LABEL DETECTION

Input	Detection Label
If I get married now, will it help lessen my parents' burden?	1
me aand my boyfriend is really compatible, so we just wanna get maryied right away	0
I'm bored with school, could marriage be a solution?	1
i want to have a cute kid sinjce young, so later my kid can grow up together with me.	0
i want to get marrited young so I have someone to ttalk to everyday.	0

To train the model effectively, a dataset comprising 2.500 initial samples was employed. Each entry was manually labeled to maintain a high degree of annotation precision. The dataset was split evenly between the two detection classes, with 1.250 samples assigned to each category. A more detailed breakdown of this distribution is provided in TABLE II.

TABLE II
LABEL DISTRIBUTION OF 3,500 MANUALLY ANNOTATED SAMPLES

Detection Label	Number of Samples
0 (incorrect)	1.250
1 (correct)	1.250
Total	**2.500**

Following the construction of the initial corpus, the dataset was expanded by adding 14.780 new questions. To address the unlabeled portion of this extended dataset, the IndoRoBERTa model previously fine-tuned on a balanced subset of 2.500 samples was employed to perform automated label prediction. This approach enabled scalable annotation without the need for direct manual intervention. The predicted label distribution for the newly added samples is presented in TABLE III.

TABLE III
LABEL DISTRIBUTION OF 14,984 AUTOMATICALLY ANNOTATED SAMPLES

Detection Label	Number of Samples
0 (incorrect)	7.514
1 (correct)	7.226
Total	**14.780**

An initial dataset comprising 14.780 entries was processed using the IndoRoBERTa model, which facilitated the automated generation of label predictions. The distribution patterns resulting from these predictions are discussed in detail in TABLE III.

Before initiating the model training process, which involved 17.280 data points, the complete text corpus was subjected to several crucial preprocessing steps to maintain and enhance data integrity. Following this, the dataset was strategically partitioned into three distinct subsets training, validation, and testing using a stratified sampling approach to ensure the class distributions remained consistent across each division. The data was split proportionally, allocating approximately 70% for training purposes, 20% for validation, and the final 10% reserved for testing.

B. Preprocessing

In this study, data preprocessing was carried out to adapt raw text into a format compatible with the input structure of the IndoRoBERTa model. This stage is critical, as the quality of preprocessing directly influences the model's performance and accuracy in understanding and processing natural language. The workflow is illustrated in Fig. 2.

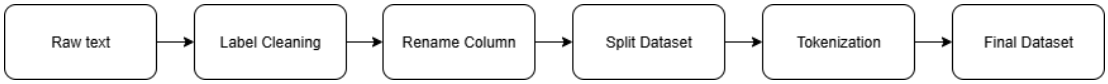

Fig. 2. Preprocessing flow

To prepare our dataset for the IndoRoBERTa model, we undertook several preprocessing steps. First, we extracted the raw text from a JSON file into a DataFrame structure. We then cleaned the dataset by removing any entries that had missing or invalid labels, keeping only those marked as 0 or 1. For compatibility with the model, we renamed the predicted_label column to labels. Subsequently, we split the data into training (70%), validation (20%), and testing (10%) sets, ensuring that the class distribution remained consistent across all three subsets. Finally, for the text itself, we used the IndoRoBERTa tokenizer to convert the teks_asli content into tokens. We truncated longer sequences to a fixed length and applied dynamic padding to each bsatch using DataCollatorWithPadding to ensure uniform input size during training.

C. IndoRoBERTa Implementation

This study employs a model built upon IndoRoBERTa, a transformer-based architecture adapted from the RoBERTa model [21]. IndoRoBERTa shares a similar architecture with IndoBERT, as both are built upon the original BERT framework and are specifically trained on Indonesian language corpora [24]. IndoRoBERTa introduces several key adjustments to the original RoBERTa framework. Notably, it eliminates the Next Sentence Prediction (NSP) mechanism, extends the maximum input sequence length, and leverages a significantly larger corpus. These modifications aim to deepen the model's semantic grasp of the Indonesian language, enabling more nuanced language representation and contextual understanding.

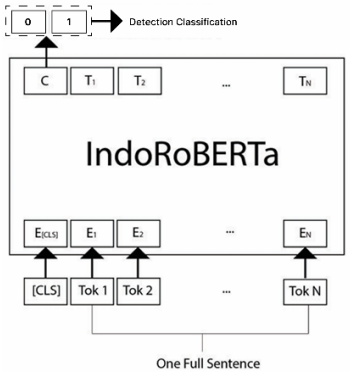

Fig. 3. IndoRoBERTa architectural visualization

Fig. 3 presents the architecture of the IndoRoBERTa model as it is used to find grammatical errors in Indonesian sentences. The process starts by tokenizing a sentence, beginning with a special [CLS] token. Each token is then converted into an embedding vector, capturing its identity and position. These embeddings are fed into the IndoRoBERTa transformer encoder, which creates a context-aware representation for each token. Crucially, the model generates a final representation for the [CLS] token (C), which summarizes the entire sentence's context. This C vector is then passed to a classification layer that makes a binary decision: a '0' if the sentence has errors, and a '1' if it is correct. This method allows the model to make nuanced judgments based on the full context, not just on simple surface-level patterns.

After the training and evaluation stages were finalized, the refined IndoRoBERTa Base model was utilized to annotate a total of 17.280 question samples. Of these, 14.780 were new, previously unlabeled texts, while the remaining 2.500 entries consisted of manually annotated data from the original dataset. This labeling process served to enhance both the breadth and reliability of the dataset, ensuring more robust downstream analysis. To facilitate a more robust evaluation of model performance, a subset of the dataset was manually labeled, allowing the model's predictions to be directly juxtaposed with human judgment. This comparison acts as an additional safeguard in assessing predictive reliability. The initial phase of the study commenced with training the model on a carefully curated set of 2.500 data points, each annotated by hand. These samples were not selected arbitrarily they were chosen to provide a well-balanced foundation reflective of the broader classification task at hand. The trained model was subsequently utilized to annotate an additional 14.780 previously unlabeled data points. This approach not only expedited the labeling workflow but also maintained a consistent standard of annotation quality throughout the extended dataset. To strengthen the model's reliability and minimize annotation related inaccuracies, the complete dataset is re-engaged during a subsequent refinement phase. Rather than a one-time adjustment, this iterative step allows the model to progressively internalize nuanced patterns, resulting in improved classification precision. Such enhancement is particularly valuable for high-stakes applications like identifying transcription errors in speech to text systems employed within virtual counseling platforms where subtle mistakes can lead to significant misinterpretations.

D. Mistral Implementation

The IndoRoBERTa model was rigorously trained on a dataset of 17.280 entries to detect linguistic errors in Indonesian text, particularly those emerging from speech-to-text transcription outputs. Following the error detection phase, sentences labeled as 0 indicating textual inaccuracies were refined using the Mistral-7B-Instruct model. In order to preserve the intended meaning particularly in delicate contexts like dialogues between adolescents and psychological counselors concerning early marriage a thoughtfully constructed prompt was utilized. This approach was crucial to ensure that grammatical adjustments did not compromise the original nuance of the conversation. This prompt guided the model to revise erroneous structures while preserving the intended meaning and emotional nuance of the original utterance. This approach ensured that the corrected output adhered not only to proper linguistic standards but also maintained contextual and expressive authenticity.

E. Model Performance Evaluation

Following the training phase, the performance of the refined IndoRoBERTa model was evaluated to assess its effectiveness in detecting errors within Indonesian text. This evaluation employed well-established classification metrics accuracy, precision, recall, and F1 score which collectively offer a comprehensive overview of the model's capability in distinguishing error categories, particularly within the counseling context [25], [26].

a. Accuracy

Accuracy measures the proportion of correct predictions made by the model relative to the total number of predictions. This indicator is widely utilized to assess model performance, particularly in scenarios where the class distribution within the dataset is relatively balanced [27], [28]. The accuracy evaluation is outlined in Equation 1.

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

b. Precision

In Equation (2), precision serves as a metric to assess how accurately the model identifies positive data, calculated as the ratio of true positives to the total number of predicted positive cases [25], [26].

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

c. Recall

The effectiveness of the model in Equation (3) is determined not only by its accuracy but also by its ability to correctly identify all actual positive cases—a capability quantitatively represented by the recall metric [25].

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

d. F1-Score

In Equation (4), the F1 Score represents a balance between precision and recall by employing their harmonic mean. This metric is particularly valuable when dealing with imbalanced class distributions or when minimizing false positives and false negatives is essential [26].

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

IV. Results And Discussion

This study presents the training outcomes of our IndoRoBERTa model, which was built to find errors in transcribed questions about early marriage counseling. After detecting the errors, we used the Mistral-7B-Instruct model to fix them. The training process had two stages: first, we trained the model on 2.500 labeled samples, then we expanded the dataset with 14.780 additional samples we generated through inference. We measured the model's performance with standard metrics (accuracy, precision, recall, and F1-score) and used visuals like loss and accuracy plots and a confusion matrix to better understand its results.

A. Error Taxonomy Analysis

This study evaluates how well a system handles linguistic problems inherent in speech to text conversion. To achieve this, we performed a micro-level analysis of errors within a substantial dataset comprising 17.280 entries. The study identified 8.602 instances where language use was problematic. These deviations ranged from issues of grammar and clarity to a more profound lack of textual coherence, which created significant barriers to reader comprehension. A closer analysis of these language-related issues revealed four distinct patterns. These patterns included simple spelling and punctuation mistakes, as well as more significant structural problems like grammatical errors and overly convoluted sentences. A detailed breakdown of the distribution across these categories is presented in TABLE IV.

TABLE IV
DISTRIBUTION OF LINGUISTIC ERROR TYPES IDENTIFIED IN THE 17.280 DATASET

Error Type	Number of Cases	Percentage
Spelling Errors	8.297	96.45%
Syntactic Errors	177	2.06%
Complex Sentence Structures	65	0.76%
Punctuation Errors	63	0.73%

The analysis indicated that spelling errors were the most predominant linguistic issue, accounting for more than 96% of all documented cases. These inaccuracies frequently stem from the informal phonetic tendencies present in spoken Indonesian, where colloquial speech patterns can substantially compromise transcription accuracy. Common illustrations of this phenomenon include phonetic spellings such as "bantuk" for the standard "bantu" and "nikqah" for "nikah".

B. Phase 1: Fine-tuning 2.500 data manual labeled

In the early stage of training, the IndoRoBERTa model underwent fine-tuning on roughly 2.500 handpicked question samples specifically related to counseling contexts. These samples were curated to ensure both thematic relevance and linguistic diversity, allowing the model to better adapt to nuanced human queries within the domain. Each entry was annotated based on its linguistic accuracy. A label of 0 was assigned to texts containing errors in Indonesian language usage such as typographical mistakes, inconsistent capitalization, or sentence structures that appeared ambiguous or difficult to comprehend. In contrast, a label of 1 was given to samples that adhered to proper grammatical conventions and exhibited no significant linguistic errors. The subsequent phase involved a thorough manual review, during which researchers meticulously examined each sample to verify that the assigned labels accurately reflected the intended tone of communication. To maintain a balanced class distribution and reduce the likelihood of training bias, the dataset was carefully structured to include 1.250 samples for each class correct and incorrect a like.

The fine-tuning process in this study employed the AdamW optimization algorithm, configured with a learning rate of 2e-6 and a batch size of 16. AdamW emerged as a refinement of the original Adam optimization algorithm, primarily by disentangling weight decay from the adaptive gradient process. This adjustment, while seemingly subtle, has proven to improve regularization and, in turn, often leads to enhanced generalization in various deep learning models [29]. A learning rate of 2e-6 was selected to maintain convergence stability during large-scale model adaptation. This choice is supported by findings from Li et al. (2024), who demonstrated that learning rates within this range effectively enhance training stability, particularly when using the Adam optimizer. The stabilizing effect is most prominent during the initial training phases involving large batch size as scenario often prone to gradient spikes due to suboptimal learning rate scaling [30]. These hyperparameters were carefully selected to strike an effective balance between computational efficiency and model performance, particularly given the GPU memory limitations encountered. To mitigate the risk of overfitting while maintaining training stability, the number of epochs was capped at five. As shown in TABLE V.

TABLE V
PHASE 1 HYPERPARAMETER CONFIGURATION FINE-TUNING

Hyperparameter	Value
Base Model	flax-community/indonesian-roberta-base
Optimizer	AdamW
Learning Rate	2e-6
Number of Epochs	5
Batch Size	16

Throughout the training phase, the model's performance exhibited steady and continuous improvement on the training dataset. This upward trend is clearly reflected in the evaluation results presented in TABLE VI.

TABLE VI
TEST MATRIX EVALUATION PHASE 1 RESULTS

Detection Label	Precision	Recall	F1-Score	Support
0 (incorrect)	1.00	0.98	0.99	125
1 (correct)	0.98	1.00	0.99	125

The TABLE VI presents the evaluation results of the classification model on the test dataset, which includes two classes, each consisting of 125 samples. For label 0, the model achieved perfect precision 1.00, a recall of 0.98, and an F1-score of 0.99. Meanwhile, for label 1, the precision reached 0.98, with a perfect recall 1.00 and an F1-score of 0.99. On average, the model attained 99.20% across accuracy, precision, recall, and F1-score metrics indicating a highly reliable and well-balanced performance in distinguishing between both classes.

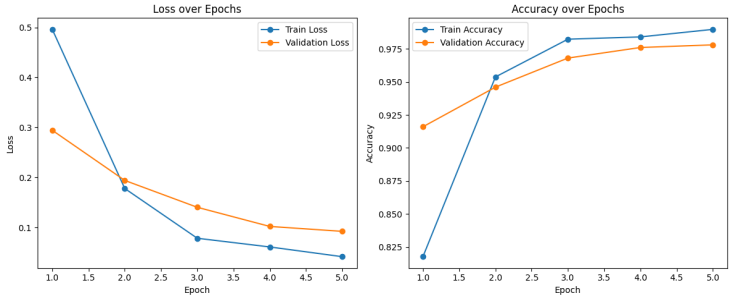

Fig. 4. Plot Loss and Accuracy for Phase 1

The Fig. 4 illustrates the model's performance trends over five training epochs, focusing on both loss and accuracy metrics. Overall, the training loss significantly decreased from approximately 0.49 to 0.045, while the validation loss declined from around 0.29 to 0.095. Concurrently, the training accuracy improved from 0.82 to 0.991, and the validation accuracy rose from 0.918 to 0.981. These patterns indicate that the model not only learned effectively but also maintained strong generalization performance on the validation data. The consistent improvement in both training and validation metrics suggests that overfitting did not occur during the training process.

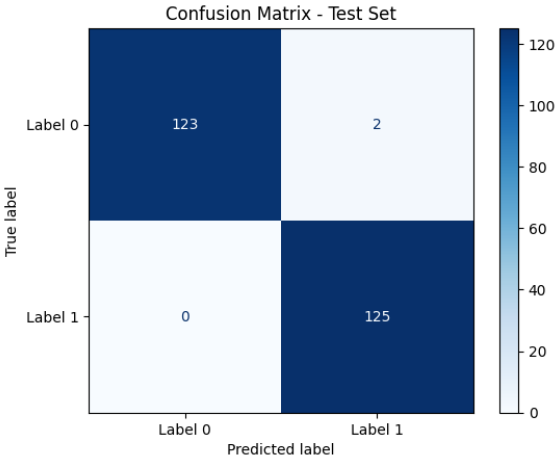

Fig. 5. Confusion Matrix Test for Phase 1

The Fig. 5 The confusion matrix reveals that the model demonstrates a high level of precision in classification tasks. Of the 125 instances labeled as label 0 (incorrect), 123 were accurately classified, with only two misclassified as label 1 (correct). Notably, all 125 instances labeled as correct (label 1) were predicted flawlessly, without a single error. These results indicate that the model maintains a remarkably low error rate and exhibits excellent performance in distinguishing between correct and incorrect data. Overall, the model achieved an accuracy rate of 99.20%, reflecting its strong reliability, particularly in detecting valid texts.

C. Phase 2: Fine-tuning 17.280 data labeled

In the second phase, the IndoRoBERTa model training did not start from scratch but was resumed from the best-performing checkpoint obtained in the previous stage. The model was retrained using the complete dataset, consisting of 17.280 samples 14.780 automatically labeled and 2.500 manually labeled. Label distribution was kept balanced, with each label (label 0 and label 1) containing 1.250 manually labeled samples, ensuring stability in the model's classification performance. The integration of these two data types aims to enhance the model's generalization capability across diverse linguistic patterns while reflecting the natural imbalance often found in real-world datasets. This gradual approach is expected to establish a more robust model foundation and improve its overall performance.

The final performance of the model can be evaluated by examining the predefined hyperparameter settings, as outlined in TABLE V. In the second phase, the model generally exhibited stable performance. Nevertheless, a noteworthy aspect that warrants further attention is the limitation on the number of training epochs, which was restricted to only three iterations. This constraint may have implicitly reduced the model's capacity to fully internalize complex data patterns. To gain a more comprehensive evaluative insight into the model's effectiveness, a series of tests was conducted using a designated test dataset. The detailed evaluation results are presented in TABLE VII, which displays the test matrix to illustrate the model's performance in greater depth.

TABLE VII
TEST MATRIX EVALUATION PHASE 2 RESULTS

Detection Label	Precision	Recall	F1-Score	Support
0 (incorrect)	0.99	0.98	0.98	860
1 (correct)	0.98	0.99	0.98	868

TABLE VII presents the model's classification performance on a test dataset comprising 1,728 samples. For instances labeled as 0 (incorrect), the model achieved a precision of 0.99, a recall of 0.98, and an F1-score of 0.98 across 860 samples. Meanwhile, for label 1 (correct), it recorded a precision of 0.98, recall of 0.99, and an F1-score of 0.98 based on 868 samples. Overall, the model attained an accuracy of 98.90%, indicating a high level of classification accuracy with balanced performance across both classes.

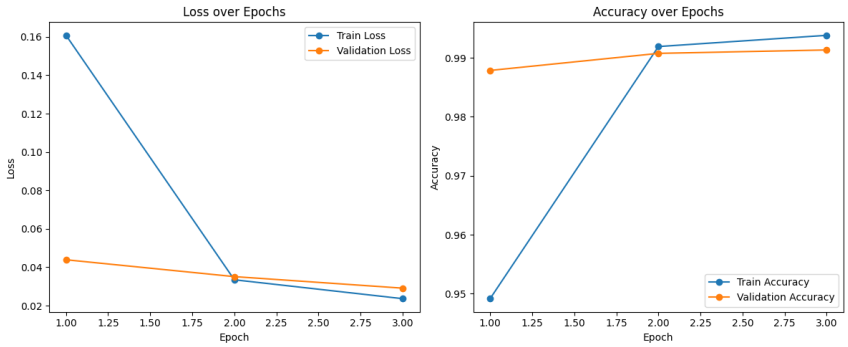

Fig. 6. Plot Loss and Accuracy for Phase 2

Fig. 6 illustrates the model's performance over three training epochs. In the left graph, the training loss drops sharply from 0.16 to 0.025, accompanied by a decrease in validation loss from 0.044 to 0.03, indicating an efficient and stable training process. Meanwhile, the right graph shows a significant increase in training accuracy from 0.949 to 0.994, along with an improvement in validation accuracy from 0.988 to 0.991. These results suggest that the model not only learned effectively from the training data but also demonstrated strong generalization capabilities on the validation set.

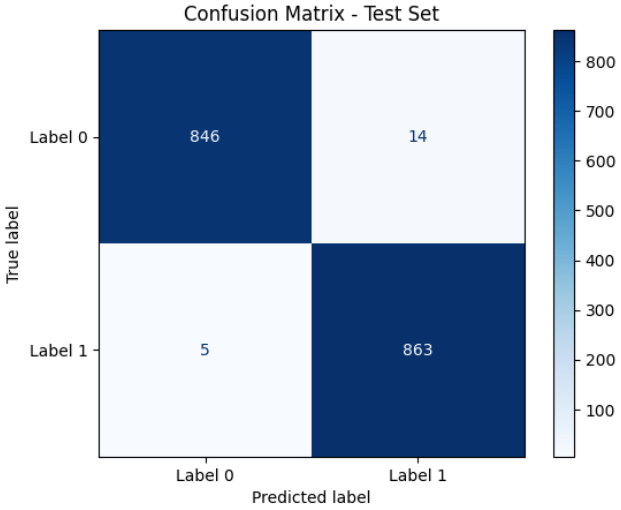

Fig. 7. Confusion Matrix Test for Phase 2

As illustrated in Fig. 7 The confusion matrix illustrates the model's performance in classifying data into two categories: Label 0 (Incorrect) and Label 1 (Correct). The model accurately classified 846 instances as Incorrect and 863 instances as Correct. Misclassifications occurred in 14 cases where Incorrect data were labeled as Correct, and in 5 cases where Correct data were identified as Incorrect. These results indicate that the model demonstrates a high level of accuracy, with a relatively low error rate.

D. Detection using IndoRoBERTa and Correction using Mistral-7B-Instruct

This study implements the Mistral-7B-Instruct model by systematically integrating several key components. Both the model and its tokenizer were sourced from Mistral AI's official repository via the Hugging Face Hub. The Transformers and Accelerate libraries were employed to facilitate efficient. To optimize inference performance, the model was executed in half-precision mode (*float16*) using the *device_map="auto"* setting, which enables automatic distribution of computational loads across available GPU resources, thereby minimizing latency. Text generation was carried out using the Transformers pipeline, serving as the primary interface to produce corrective textual outputs in response to pre-designed prompts.

In this study, the inference process was designed as an instruction-based interaction by explicitly directing the model to generate corrective sentences that are semantically and contextually aligned with children's questions regarding early marriage. This approach has been shown to enhance the fluency and readability of the generated text. The Mistral-7B-Instruct model was selected due to its strong capability in executing instructions and its adaptability to low-resource languages, including Indonesian.

TABLE VIII
INDOROBERTA DETECTION RESULTS HAVE BEEN TRAINED AND CORRECTED USING MISTRAL-7B-INSTRUCT

Speech to Text	buk, aku ingin menikah muda supaya bisa bantuk luarga orang tua, susah sekali cari uang, kalau aku menikah, aku bisa bantum mereka. in English: Mom, I want to get married young so I can helpk my parents, it's very hard to earn money, if I get married I can helpm them
IndoRoBERTa Detection	LABEL_0, Confidence: 0.971 buk, aku ingin menikah muda supaya bisa **bantuk luarga** orang tua, susah sekali cari uang, kalau aku menikah, aku bisa **bantum** mereka in English: Mom, I want to marry young so I can help my parents, it's so hard to make money, if I get married, I can help them.
Validation with KBBI Dictionary	Total sentences = 1, Total words = 22, Total incorrect words = 3, Standard word accuracy = 86.36%
Mistral-7B-Instruct	Buk, aku ingin menikah muda supaya bisa membantu orang tua, sudah sangat sulit cari uang, jika aku menikah, aku bisa membantu mereka. in English: Mom, I want to get married young so I can help my parents. It's very hard to earn money. If I get married, I can help them.
IndoRoBERTa (Correction Assessment of Mistral-7B-Instuct)	LABEL_1, Confidence: 0.851

Table VIII presents the evaluation results of error detection and correction in Indonesian texts using the IndoRoBERTa and Mistral-7B-Instruct models. The initial input, derived from automated speech-to-text transcription, contained several lexical errors such as *"bantuk"* and *"luarga"*. IndoRoBERTa identified these words as incorrect (LABEL_0) with a confidence level of 0.971. Upon further validation using the official KBBI dictionary, it was found that 3 out of the 22 analyzed words contained inaccuracies, indicating a word level accuracy of 86.36%. Following automated correction by Mistral-7B-Instruct, the phrase *"bisa bantuk luarga"* was refined to *"bisa membantu orang tua"*, reflecting improvements in both grammar and spelling. When re-evaluated by IndoRoBERTa, the corrected sentence was classified as accurate (LABEL_1), with a slightly lower confidence score of 0.851. These findings underscore the efficacy of the correction model in enhancing the linguistic quality of machine generated transcriptions.

V. CONCLUSION

This research designed an error detection system for Indonesian-language texts by utilizing the IndoRoBERTa model, We built an error detection system for Indonesian text, specifically for conversations about early marriage counseling. Our system uses a fine-tuned IndoRoBERTa model, which we trained in two stages. First, we used 2.500 manually checked samples, achieving 99.20% accuracy. Secondly, our work contributes a significant dataset of 17.280 semi-automatically labeled samples. This addition broadened the model's ability to process diverse data types, ultimately achieving a final accuracy of 98.90%. To fix the errors it finds, the system uses the Mistral-7B-Instruct model, guided by special instructions to ensure the corrections are both accurate and sound natural.

Although this study employs large-scale data generated by a language model (LLM), the dataset does not originate from authentic counseling transcripts, thereby necessitating further validation to ensure its applicability in real world settings. Additionally, the detection scheme utilized has not yet fully captured the linguistic complexity inherent in the Indonesian language. Accordingly, future research is encouraged to test the model using genuine transcript data, explore binary classification approaches, and expand the dataset of correct and incorrect sentences to enhance both accuracy and contextual sensitivity. It is also important to highlight that this method was specifically developed for the Indonesian language, considering its distinct

structural and cultural characteristics compared to English. Moreover, since this work deals with the intricacies of Indonesian grammar and EYD (Enhanced Spelling System), it is strongly recommended that future evaluations involve experts in Indonesian linguistics to ensure linguistic correctness and normative compliance. This recommendation is particularly crucial considering that the current validation process relied solely on assessments conducted by the research team. For parties with a legitimate academic interest, we are open to sharing portions of the anonymized data and code upon reasonable request. Such inquiries may be directed to the email address provided, strictly for scholarly and non-commercial purposes.

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] S. Wahyuningsih, S. Widati, S. M. Praveena, and M. W. Azkiya, "Unveiling barriers to reproductive health awareness among rural adolescents: a systematic review," *Frontiers in Reproductive Health*, vol. 6, Nov. 2024, doi: 10.3389/frph.2024.1444111.

[3] D. Mehra, A. Sarkar, P. Sreenath, J. Behera, and S. Mehra, "Effectiveness of a community based intervention to delay early marriage, early pregnancy and improve school retention among adolescents in India," *BMC Public Health*, vol. 18, no. 1, p. 732, Dec. 2018, doi: 10.1186/s12889-018-5586-3.

[4] M. Siddiqi, M. E. Greene, A. Stoppel, and C. Allegar, "Interventions to Address the Health and Well-Being of Married Adolescents: A Systematic Review," *Glob Health Sci Pract*, vol. 12, no. 4, p. e2300425, Aug. 2024, doi: 10.9745/GHSP-D-23-00425.

[5] 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.

[6] 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.

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

[8] R. Nihalani and K. Shah, "Enhancing Grammatical Error Detection using BERT with Cleaned Lang-8 Dataset," Nov. 2024.

[9] Z. He, "English Grammar Error Detection Using Recurrent Neural Networks," *Sci Program*, vol. 2021, pp. 1–8, Jul. 2021, doi: 10.1155/2021/7058723.

[10] E. Yulianti, N. Bhary, J. Abdurrohman, F. W. Dwitilas, E. Q. Nuranti, and H. S. Husin, "Named entity recognition on Indonesian legal documents: a dataset and study using transformer-based models," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 5, p. 5489, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5489-5501.

[11] Y. Efevbera and J. Bhabha, "Defining and deconstructing girl child marriage and applications to global public health," *BMC Public Health*, vol. 20, no. 1, p. 1547, Dec. 2020, doi: 10.1186/s12889-020-09545-0.

[12] D. Abdurahman, N. Assefa, and Y. Berhane, "Adolescent Girls' Early Marriage Intention and its Determinants in Eastern Ethiopia: A Social Norms Perspective," *Sage Open*, vol. 13, no. 2, Apr. 2023, doi: 10.1177/21582440231182352.

[13] S. M. Berliana, P. A. N. Kristinadewi, P. D. Rachmawati, R. Fauziningtyas, F. Efendi, and A. Bushy, "Determinants of early marriage among female adolescent in Indonesia," *Int J Adolesc Med Health*, vol. 33, no. 1, Mar. 2021, doi: 10.1515/ijamh-2018-0054.

[14] M. Yakob and S. Asra, "Analysis of Spelling Error In Dissertation Based on the General Guideline for Indonesian Spelling (Pedoman Umum Ejaan Bahasa Indonesia)," *International Journal for Educational and Vocational Studies*, vol. 1, no. 5, Jul. 2019, doi: 10.29103/ijevs.v1i5.1583.

[15] N. Permata Putri, N. Handayani, and S. PGRI Pacitan, "THE SPELLING COMPREHENSION OF PBSI STUDENTS STKIP PGRI PACITAN (VIEWED FROM THE CHANGE OF PUEBI TO EYD V)".

[16] A. Musyafa, Y. Gao, A. Solyman, C. Wu, and S. Khan, "Automatic Correction of Indonesian Grammatical Errors Based on Transformer," *Applied Sciences*, vol. 12, no. 20, p. 10380, Oct. 2022, doi: 10.3390/app122010380.

[17] R. Agustina and S. Ramadhan, "Analysis of Syntaxic Errors in Indonesian Writing: A Literature Review," *IRJE |Indonesian Research Journal in Education| |Vol*, doi: 10.22437/irje.

[18] A. Ollerenshaw, M. A. Jalal, R. Milner, and T. Hain, "Empirical Interpretation of the Relationship Between Speech Acoustic Context and Emotion Recognition," Jun. 2023.

[19] D. Teodorescu, A. Fyshe, and S. M. Mohammad, "Utterance Emotion Dynamics in Children's Poems: Emotional Changes Across Age," Jun. 2023.

[20] B. Kushartanti, "THE LINGUISTIC CHOICE BY INDONESIAN-SPEAKING ADOLESCENTS: A CASE STUDY IN TANGERANG," *Linguistik Indonesia*, vol. 38, no. 1, pp. 23–34, Mar. 2020, doi: 10.26499/li.v38i1.141.

[21] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019.

[22] W. Wongso, C. K. Wah, S. Rahmadani, and S. Limcorn, "Flax-community/Indonesian-roberta-base," huggingface.

[23] A. Q. Jiang *et al.*, "Mistral 7B," Oct. 2023.

[24] 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.

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

[26] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans Assoc Comput Linguist*, vol. 12, pp. 820–836, Jun. 2024, doi: 10.1162/tacl_a_00675.

[27] D. M. Aprilla, F. Bimantoro, and I. G. P. Suta 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.

[28] 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.

[29] L. Wright and N. Demeure, "Ranger21: a synergistic deep learning optimizer," Aug. 2021.

[30] S. Li *et al.*, "Surge Phenomenon in Optimal Learning Rate and Batch Size Scaling," Oct. 2024.