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

Ind. Journal on Computing
Vol. 10, Issue. 2, February 2026. pp. 1-19
doi:10.21108/indojc.v10i2.9709

Implementation of IndoRoBERTa to Improve the Clarity of the Context of Homograph Words in the Text-to-Speech System for Education Chatbot Early Marriage in Lombok

Fikri Rahmanda Noor[1], Rifki Wijaya[2], Ade Romadhony[3]

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

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

Abstract

This study presents the implementation of IndoRoBERTa, a pre-trained Indonesian language model, to improve the contextual clarity of homograph words in Text-to-Speech (TTS) systems, particularly for virtual chatbot applications addressing early marriage education in Lombok. The proposed system integrates IndoRoBERTa into the TTS pipeline to classify the context of homographs prior to grapheme-to-phoneme (G2P) conversion, ensuring accurate pronunciation based on meaning. The research was conducted in two fine-tuning phases: the first utilized 500 manually labeled conversational samples, achieving 96% test accuracy, while the second expanded the dataset with 2,000 auto-labeled samples and yielded 88% accuracy. Evaluation metrics including precision, recall, and F1-score demonstrated the model's effectiveness across 20 homograph categories. Despite strong results, the study acknowledges limitations in data authenticity and challenges in underrepresented classes. Future work is recommended to incorporate real-world dialogue data and enhance the system's generalization in more complex linguistic settings. This research contributes to the advancement of Indonesian NLP in TTS systems, particularly in socially impactful educational contexts.

Keywords: IndoRoBERTa, Text-to-Speech, Homograph, Early Marriage, Indonesian NLP

Abstrak

Penelitian ini mengkaji pemanfaatan IndoRoBERTa, sebuah model bahasa pralatih untuk Bahasa Indonesia, dalam meningkatkan keakuratan pelafalan kata homograf pada sistem Text-to-Speech (TTS). Fokus utamanya adalah pada chatbot virtual untuk edukasi pencegahan pernikahan dini di wilayah Lombok. Dalam sistem ini, IndoRoBERTa diintegrasikan ke dalam proses TTS guna mengklasifikasikan konteks kata homograf sebelum tahapan grapheme-to-phoneme (G2P), agar pelafalan sesuai dengan makna dalam kalimat. Metode penelitian melibatkan dua tahap pelatihan model. Tahap pertama menggunakan 500 data percakapan berlabel manual dan mencapai akurasi 96%. Tahap kedua memperluas data dengan 2.000 entri yang diberi label secara otomatis oleh model, menghasilkan akurasi 88%. Evaluasi menggunakan metrik presisi, recall, dan F1-score memperlihatkan performa model yang cukup solid dalam mengidentifikasi 20 konteks homograf

Received on September, 2025. Accepted on February, 2026

berbeda. Namun, penelitian ini juga mengakui beberapa keterbatasan, seperti ketidakotentikan data yang digunakan serta kurangnya representasi pada beberapa kelas kata. Untuk itu, riset lanjutan disarankan memakai dialog nyata dan memperkaya cakupan data agar model lebih adaptif terhadap variasi bahasa alami. Studi ini menjadi kontribusi penting dalam pengembangan NLP berbahasa Indonesia, khususnya pada aplikasi TTS yang menyasar isu sosial edukatif.

Kata Kunci: IndoRoBERTa, Text-to-Speech, Homograf, Pernikahan Dini, Indonesia NLP

I. Introduction

THE rapid pace of digital transformation has reshaped various aspects of life, including how we learn and access information. Despite these advancements, early marriage remains a deeply rooted and complex social issue. It continues to impact the well-being of younger generations affecting their health, education, and long-term quality of life. Child marriage involving individuals under the age of 18 is both a breach of human rights and a clear indicator of persistent gender inequality [1]. Before the Covid pandemic drew worldwide attention, UNICEF reported in 2021 that around 100 million girls had already been married off before turning 18 [2]. West Nusa Tenggara (WNT) is among the Indonesian provinces with a high prevalence of child marriage. Records from the Ministry of Religion's Regional Office in WNT show that in 2020, West Lombok Regency (Praya) received 135 applications for marriage dispensation. Meanwhile, the Praya Religious Court reported 136 child marriage cases between February and June 2021 [3]. A comparable pattern is seen in Lombok, particularly in the Lingsar region, where early marriage is still common within the Sasak community. Data from West Lombok's DP3AK2B shows that in 2017 alone, 1,038 underage girls and 357 boys were married before reaching 20 years old [4]. Early or child marriage brings serious consequences for women, affecting their health, dignity, and independence. Each year, an estimated 12 million girls are married off before turning 18 [5]. Within the Sasak community's religious and cultural framework, there are no explicit rules that set a clear age for marriage, which makes it challenging to curb the practice of child marriage [6].

Despite ongoing educational efforts, many remote communities still lack sufficient access to reliable information. As a solution, virtual chatbots have emerged as effective, interactive tools for educational outreach. Their growing use in language learning is largely due to their ability to engage users through conversational, natural-sounding language [7]. However, voice-based chatbot systems continue to face technical hurdles one of the most prominent being mispronunciation of homographs. These are words spelled identically but spoken and interpreted differently depending on context. If mispronounced, they can create confusion and alter the intended meaning of a message.

In this light, context-aware Text-to-Speech (TTS) systems play a crucial role. One of the core challenges in AI-driven conversational systems is producing speech that mirrors human voice quality [8]. To address this, Natural Language Processing (NLP) technologies offer a promising avenue. Applying NLP effectively, however, demands a careful balance between human linguistic insight and computational power [9].

Among the most advanced NLP models tailored for Indonesian is IndoRoBERTa, a variant of RoBERTa (Robustly Optimized BERT Pretraining Approach). Unlike the original BERT model that includes a Next Sentence Prediction objective, RoBERTa discards that component and benefits from more extensive training data and longer training time. IndoRoBERTa is a variant of BERT, specifically adapted from the RoBERTa architecture and trained using Indonesian language data. In comparative evaluations, IndoRoBERTa consistently outperformed other BERT-based models designed for Indonesian, demonstrating superior performance [10]. IndoRoBERTa, developed by the IndoNLU team, is trained on a large-scale Indonesian corpus, allowing it to more accurately capture the syntactic and semantic intricacies of the language. These strengths make it particularly suitable for tasks requiring nuanced contextual understanding, such as homograph classification in TTS systems.

Yet, despite its potential, IndoRoBERTa has rarely been applied in the domain of homograph articulation refinement in Indonesian TTS. Correct pronunciation is essential especially in educational settings where the clarity of spoken language influences comprehension, more so when dealing with sensitive topics like early marriage. In response to this gap, the current study proposes a solution: a TTS framework that leverages IndoRoBERTa to detect and interpret homographs based on context, prior to the grapheme-to-phoneme (G2P) conversion. This ensures that pronunciation accurately reflects the intended meaning [11].

By incorporating IndoRoBERTa into the workflow, this system is designed to improve the accuracy of homograph pronunciation and deliver clearer, contextually precise communication in voice-based educational chatbots. The central goal of the study is to enhance the intelligibility of homograph articulation in Indonesian TTS for early marriage awareness chatbots through contextual classification. The underlying hypothesis is that using IndoRoBERTa will significantly boost the system's ability to correctly identify and pronounce homographs, producing speech output that is both more understandable and context appropriate.

Nonetheless, several limitations must be recognized when considering real-world application. Spoken inputs may still be misinterpreted due to regional accents or dialectal differences, which could lower the accuracy of homograph pronunciation. In addition, the correction module may not always achieve flawless results, and automated educational tools inevitably bring ethical challenges particularly for sensitive topics like early marriage. These concerns involve ensuring that responses remain culturally appropriate and that the system avoids unintentionally misleading users in crucial counseling scenarios.

The study evaluates success not only through standard metrics such as accuracy, precision, recall, and F1-score but also through its practical benefits. These include enhancing the clarity of chatbot output, preserving cultural sensitivity in the Lombok context, and creating a foundation for future assessments of user experience. By addressing both technical and socio-cultural dimensions, this research contributes to developing TTS systems that are not only effective but also socially responsible.

II. Literature Review

A few studies have explored the use of machine learning and transformer-based models in tackling various text classification tasks involving the Indonesian language. One such study applied the BERT classification algorithm to detect hoax news, achieving a validation accuracy of 76%. This result reflects a solid performance in identifying false or misleading information within online news content [12].

In another research effort, Random Forest paired with optimized grid search parameters was employed to classify mobile phone pricing data. By adjusting hyperparameters such as max_depth and criteria like gain_ratio, gini_index, and accuracy, the study attained a peak classification accuracy of 88.50%, showcasing the strength of traditional ensemble models when applied to structured categorical datasets [13].

A more recent investigation focused on sentiment and emotion classification using Indonesian Twitter data and leveraged IndoRoBERTa an enhanced version of RoBERTa pre-trained on a large Indonesian corpus. The findings demonstrated that IndoRoBERTa outshined other BERT-based Indonesian models, achieving 98% accuracy and 97.4% F1-score in sentiment classification. For emotion detection, the model reached an F1-score of 83% and an accuracy of 82.7%. Further analysis using confusion matrices indicated that the model consistently provided accurate predictions across all sentiment categories [14].

Overall, these findings underline the growing effectiveness of both classical machine learning techniques and modern pre-trained language models in Indonesian NLP. In particular, the outstanding performance of IndoRoBERTa highlights its potential for addressing more intricate classification problems such as homograph disambiguation within context-sensitive applications like Text-to-Speech (TTS) systems for educational chatbots.

A. Early Marriage

Broadly defined, early marriage refers to the legally recognized union between two individuals of opposite genders who are still in their teenage years, bound under the institution of family [15]. In Indonesia, Law Number 16 of 2019 Article 7 Paragraph (1) stipulates that marriage is only allowed if both parties male and female are at least 19 years old. In Lombok, this practice is known in the Sasak language as merariq kodeq, a long-standing custom that continues to be widely practiced. It's viewed as part of the Sasak community's cultural heritage, handed down across generations [16].

Recent data reveals a noticeable rise in early marriage rates across four Indonesian provinces. South Kalimantan saw an increase to 21.2%, followed by Central Kalimantan at 20.2%, Central Sulawesi at 16.3%, and West Nusa Tenggara (NTB) at 16.1% [17].

B. Text-to-speech

Text-to-Speech (TTS) refers to a technology that automatically converts written text in natural language into spoken audio that mimics the pronunciation of a native speaker [18]. This technology brings together several fields of study, including Natural Language Processing (NLP), linguistics, and speech synthesis. In Indonesian TTS systems, one of the core processes involved is Grapheme-to-Phoneme (G2P) conversion.

Grapheme-to-Phoneme (G2P) plays a central role by translating written characters (graphemes) into their corresponding sound units (phonemes). This process is essential for speech technologies like TTS and Automatic Speech Recognition (ASR), as it enables systems to generate accurate pronunciations, even for words not found in the existing vocabulary. By doing so, G2P helps speech systems operate more flexibly, reducing dependence on manually curated pronunciation dictionaries [11].

C. RoBERTa and IndoRoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is an improved adaptation of the original BERT model, developed to overcome its known limitations. It refines the pretraining strategy by significantly increasing the size of training data and batch capacity. RoBERTa also introduces dynamic masking throughout the training process and discards the Next Sentence Prediction (NSP) task entirely [19]. These enhancements allow RoBERTa to learn more nuanced and comprehensive representations of language, leading to significant improvements in performance across a variety of Natural Language Processing (NLP) tasks. At present, the field of Natural Language Processing (NLP) is largely driven by general-purpose pretrained language models such as RoBERTa, which deliver impressive results on natural language understanding (NLU) tasks by being trained on massive datasets containing billions of words [20].

Following the improvements brought by RoBERTa to BERT pretraining framework, IndoRoBERTa emerged as a localized adaptation tailored to the Indonesian language. IndoRoBERTa, or Indonesian RoBERTa, is a variant of the RoBERTa model that has been pretrained on a corpus specifically composed of Indonesian language data [21].

In the case of homograph disambiguation, IndoRoBERTa strength lies in its ability to process words within their complete sentence-level context. This contextual awareness allows the model to distinguish between different meanings and pronunciations of identically spelled words, depending on how they're used in surrounding text. Because of this, IndoRoBERTa serves as a strong foundation for building more precise and context-sensitive Text-to-Speech (TTS) systems in Indonesian particularly for educational tools that aim to address sensitive topics like early marriage prevention.

D. Homograph

Homographs are words that are spelled the same but can have different meanings and sometimes different pronunciations, depending on how they are used in a sentence. In Indonesian language, as in the example context of homographs in this study, for example, kecap may refer to soy sauce or the movement of the mouth, and apel

can mean either the fruit or a formal gathering. Such words often introduce ambiguity in both written and spoken language because their correct meaning cannot be determined from the spelling alone. In the study of semantics and pragmatics, homographs are a key focus for understanding how meaning shifts in spoken communication [22]. Their dual meanings or inherent ambiguity reflect not only the expressive richness of a language but also the collective mindset of its speakers.

In Natural Language Processing (NLP), managing homographs poses a significant challenge across tasks like text classification, machine translation, and speech processing. This issue is especially critical in Text-to-Speech (TTS) systems, where mispronouncing a homograph can distort the intended message or create confusion in the spoken output. For example, if the word apel in the sentence "Besok ada apel di kantor" is pronounced as the fruit rather than an assembly, the meaning conveyed would shift entirely.

III. RESEARCH METHOD

This research introduces a systematic framework that leverages the IndoRoBERTa model to classify the contextual meanings of homograph words in Indonesian Text-to-Speech (TTS) applications. The methodology is divided into several stages, including data preprocessing, a two-phase fine-tuning of IndoRoBERTa, integration with the TTS pipeline, and performance evaluation of the model. An overview of the entire process is depicted in the block diagram in Fig. 1, which highlights the essential stages of model construction, training, and implementation.

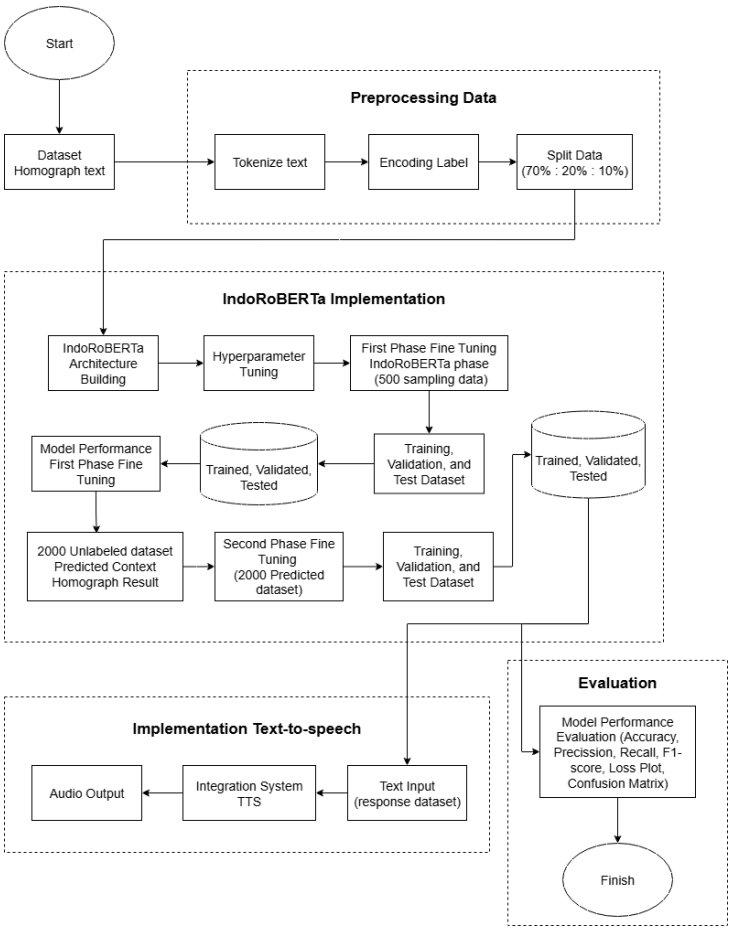

Fig. 1. Flow diagram

A. Data Collection

This research employs a dataset of 2,500 Indonesian-language entries created specifically for contextual homograph classification in a Text-to-Speech (TTS) system for an educational chatbot addressing early marriage. The dataset was synthetically produced using a Large Language Model (LLM) to mimic natural dialogues between adolescents and psychological counselors discussing early marriage. The use of LLM-generated data was motivated by the lack of publicly available counseling datasets in Indonesia, particularly in Lombok, as well as the limited availability of homograph-rich text for TTS tasks. Gathering real counseling dialogues would also involve significant ethical and privacy challenges, making synthetic generation the most feasible approach. Each entry includes the adolescent's utterance, the chatbot's response, a homograph word, and a context label that defines its intended meaning. Of the total entries, 500 were manually annotated into 20 context categories two meanings for each of 10 homograph words while the remaining 2,000 were automatically labeled using the IndoRoBERTa model after its initial fine-tuning. A manual spot check was conducted to validate the synthetic labels, confirming that the data quality and model outcomes were comparable to those from fully curated datasets.

The dataset is entirely text-based, as the study centers on context detection before the grapheme-to-phoneme (G2P) stage in the TTS pipeline. Minor variations, such as typographical changes, were introduced to simulate realistic conversation. Finally, the dataset was split using a stratified ratio of 70% for training, 20% for validation, and 10% for testing, ensuring balanced label representation across all 20 categories while supporting strong generalization and robust evaluation.

TABLE I
EXAMPLE DATASET OF THE HOMOGRAPH WORD "APEL"

Input	**Response**	**Homograph**	**Context**	**G2P**
Dulu waktu sekolah, aku jualan apel di pasar. Kalau nikah muda, aku bisa usaha bareng pasangan lagi, kan? " When I was in school, I sold apples at the market. If I get married young, I can start another business with my partner, right?"	Usaha jualan apel bisa jalan terus meski belum nikah. Bahkan, kamu bisa lebih bebas belajar dan berkembang kalau fokus dulu ke usaha dan diri sendiri. Nikah nanti aja kalau semuanya udah siap. " You can continue your apple business even if you're not married. In fact, you'll have more freedom to learn and grow if you focus on your business and yourself first. Marry later when everything is ready."	Apel	Buah "fruit"	uˈsaha dʒuˈalan ˈapɛl ˈbisa ˈdʒalan təˈrus məsˈki bəˈlum ˈnikah. ˈbahkan, ˈkamu ˈbisa ləˈbih ˈbebas bəˈladʒar dan bərkəmˈbaŋ kaˈlau ˈfokʊs ˈdulu kə uˈsaha dan ˈdiri sənˈdiri. ˈnikah ˈnanti ˈadʒa kaˈlau səmuˈaŋa ˈudah ˈsiap.
Waktu apel sekolah tadi, temen-temen ngebahas soal nikah muda. Katanya seru dan romantis. Aku jadi kepikiran juga. " During school assembly earlier, my friends were discussing early marriage. They said it was exciting and romantic. It got me thinking too."	Seru pas dengerin di apel, tapi kenyataan nikah muda gak sesimpel itu. Banyak tanggung jawab yang belum tentu bisa kamu pegang sekarang. Lebih baik fokus dulu sama diri sendiri dan cita-cita. "It's exciting to hear about at the roll call, but the reality of getting married young isn't that simple. There are many responsibilities you might not be able to handle right now. It's better to focus on yourself and your dreams first."	Apel	Kumpul "gather"	səˈru pas deŋəˈrin di ˈapɛl, ˈtapi kəŋaˈtaan ˈnikah ˈmuda gaʔ səˈsimpəl ˈitu. ˈbaŋaʔ ˈtaŋgʊŋ ˈdʒawap jaŋ bəˈlum tənˈtu ˈbisa ˈkamu pəˈgaŋ səˈkaraŋ. ləˈbih ˈbaɪʔ ˈfokʊs ˈdulu ˈsama ˈdiri sənˈdiri dan ˈʧita ˈʧita.

To enhance articulation within the Text-to-Speech (TTS) system, every text sample was subjected to a series of preprocessing steps such as text normalization, noise filtering, and phonetic correction. The questions submitted by adolescents were thoroughly examined to guarantee precise labeling of homograph contexts, ensuring the dataset-maintained consistency and reliability. Table 2 illustrates the balanced distribution of labels in the manually annotated initial dataset.

TABLE II
MANUAL LABELING OF 500 DATASETS

Homograph	**Context**	**Sample Total**
Apel	Buah “Fruit”	25
	Kumpul “Gather”	25
Tahu	Makanan “food”	25
	Situasi “Situation”	25
Serang	Melawan “Oppose”	25
	Nama Tempat “Place Name”	25
Memerah	Berubah Warna “Change Color”	25
	Memerah Susu “Express Milk”	25
Kecap	Gerakan Mulut “Mouth Movements”	25
	Bumbu “Seasoning”	25
Bulan	Kalendar “Calendar”	25
	Satelit “Satelite”	25
Bunga	Perbankan “Banking”	25
	Tanaman “Plant”	25
Hak	Kewenangan “Authority”	25
	Sepatu “Shoe”	25
Bebek	Binatang “Animal”	25
	Jenis Rujak “Type of Salad”	25
Mental	Rohani “Spiritual”	25
	Gerakan Terlempar “Thrown Movement”	25

The initial dataset was further expanded with the addition of 2,000 new entries featuring homograph words. To streamline the process, these entries originally unlabeled, were automatically labeled using a context classification model that had already been fine-tuned on manually labeled data. The model leveraged its prior training to predict the correct contextual meaning for each homograph. Table 3 displays the distribution of context labels generated through this automated annotation process.

TABLE III
AUTO LABELING OF 2000 DATASETS

Homograph	**Context**	**Sample Total**
Apel	Buah “Fruit”	153
	Kumpul “Gather”	89
Tahu	Makanan “Food”	163
	Situasi “Situation”	126
Serang	Melawan “Oppose”	65
	Nama Tempat “Place Name”	105
Memerah	Berubah Warna “Change Color”	123
	Memerah Susu “Express Milk”	38
Kecap	Gerakan Mulut “Mouth Movements”	24
	Bumbu “Seasoning”	81
Bulan	Kalendar “Calendar”	97
	Satelit “Satelite”	100

Bunga	Perbankan “Banking”	71
	Tanaman “Plant”	166
Hak	Kewenangan “Authority”	146
	Sepatu “Shoe”	71
Bebek	Binatang “Animal”	87
	Jenis Rujak “Type of Salad”	107
Mental	Rohani “Spiritual”	129
	Gerakan Terlempar “Thrown Movement”	59

Table 3 presents the distribution of homograph context classifications generated by the model. Prior to training, both the manually and automatically labeled datasets were processed through a series of preprocessing steps, including tokenization and label encoding. The complete dataset was then split into three key subsets 70% for training, 20% for validation, and 10% for testing using a stratified sampling method to ensure that the distribution of context classes remained balanced throughout.

B. Preprocessing Data

In the preprocessing stage, raw text entries are tokenized using the IndoRoBERTa tokenizer, and context labels are encoded into numerical format. The dataset is split into training, validation, and test sets with a stratified ratio of 70%: 20%: 10% to ensure balanced label distribution. The 70% training split was chosen because allocating most data to training typically enhances a deep learning model’s ability to generalize, particularly for classification tasks, while still leaving enough samples for dependable validation and testing [23]. These steps are critical to prepare the data for efficient and accurate learning.

C. IndoRoBERTa Implementation

This study utilizes the IndoRoBERTa model, a monolingual adaptation of RoBERTa [24], specifically pretrained on Indonesian text. IndoRoBERTa removes the Next Sentence Prediction (NSP) component found in BERT and adopts a training approach that uses larger batch sizes, longer sequences, and dynamic masking to improve contextual language representation.

As depicted in Figure 2, IndoRoBERTa is employed in this research to classify context homographs words with identical spelling but different pronunciations and meanings depending on the sentence context.

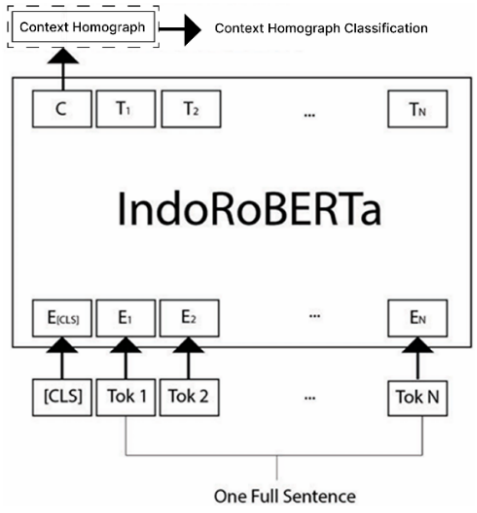

Fig. 2. IndoRoBERTa architecture [25]

Figure 2 illustrates how the IndoRoBERTa model processes and classifies homograph contexts. Each input consists of a full sentence taken from a simulated conversation between a teenager and a counselor, where the model must accurately interpret an ambiguous homograph based on its surrounding context. The sentence is first broken down into subword tokens (Tok 1, Tok 2, ..., Tok N) using the IndoRoBERTa tokenizer, which was pretrained on Indonesian-language text.

A special token, CLS, is added to the beginning of the sequence to represent the overall meaning of the sentence. Every token, including CLS, is then converted into a vector form (E[CLS], E1, E2 ..., EN) and passed through twelve layers of transformer encoders. As the sentence is processed, the model produces contextual embeddings (C, T1, T2, ..., TN), with the final hidden state of the CLS token marked as C in the diagram acting as a summary representation of the entire sentence. This vector is passed to a classification layer, which predicts one of twenty possible context categories. These include: buah, kumpul, makanan, situasi, melawan, nama tempat, berubah warna, memerah susu, gerakan mulut, bumbu, kalendar, satelit, perbankan, tanaman, kewenangan, sepatu, binatang, jenis rujak, rohani, and gerakan terlempar.

The implementation consists of two phases:

1) First Phase Fine-Tuning: The IndoRoBERTa model is fine-tuned using the manually labeled 500-entry dataset. The model is trained, validated, and tested to assess its initial ability in classifying homograph contexts.
2) Second Phase Fine-Tuning: Using the trained model from Phase 1, context labels are predicted for the remaining 2,000 entries. These predictions are then used to retrain the model. This second fine-tuning phase aims to improve generalization and performance on more varied data.

D. Integration With Text-to-Speech System

After completing both training phases, the final model is integrated into a prototype Indonesian TTS system. The output from chatbot responses is passed as input to the classifier, which assigns the correct context for the homograph. The result is then processed by the Grapheme-to-Phoneme (G2P) component to produce audio output with accurate pronunciation.

E. Evaluation

In this study, the IndoRoBERTa model was assessed to determine how well it improves the articulation of homographs in Indonesian texts. The evaluation relied on a set of standard performance metrics accuracy, precision, recall, and F1-score to give a well-rounded view of the model overall effectiveness.

As defined in Equation (1), accuracy measures the ratio of correctly predicted instances both positive and negative against the total number of test samples. It's often used as a broad measure of how well a model performs, especially in situations where the dataset's class distribution is fairly even [26].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In Equation (2), precision reflects how accurately the model predicts the positive class that is, the proportion of correct positive predictions out of all predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

In Equation (3), recall measures the model's ability to identify all actual positive instances in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In Equation (4), the F1-score calculates the harmonic mean between precision and recall, making it especially useful when dealing with imbalanced data distributions.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

True Positive (TP): Positive samples that the model successfully predicts as positive.

True Negative (TN): Negative samples accurately classified as negative by the model.

False Positive (FP): Negative samples that the model incorrectly labels as positive.

False Negative (FN): Positive samples that the model fails to recognize, predicting them as negative instead.

IV. Results And Discussion

This chapter outlines the training results of the IndoRoBERTa based model designed to interpret homograph meanings by analyzing contextual cues within counseling-style conversations, particularly those centered around early marriage education. The training was carried out in two separate stages. The first phase involved 500 manually labeled samples to establish the model's baseline accuracy. The second phase expanded the training set with 2,000 additional entries labeled automatically, aimed at improving the model adaptability across varied contexts. The model's effectiveness was assessed using standard classification metrics accuracy, precision, recall, and F1-score. To enhance clarity and insight into the training dynamics, visual representations are included, such as training loss and accuracy graphs, along with confusion matrices, showcasing the model's performance and its reliability across the 20 homograph context categories.

A. First Phase Fine tuning on 500 Manually Labeled Data

During the first phase of training, IndoRoBERTa was fine-tuned on a set of 500 data samples that had been manually annotated across 20 distinct homograph context categories. These samples were selected from a larger pool of 2,500 entries to assess the model's ability to interpret homograph meanings within counseling-themed dialogue. Each entry included a response sentence paired with a context label identifying the intended meaning of the homograph used in the sentence.

This fine-tuning phase followed a supervised learning strategy. To maintain balanced class representation, the dataset was divided using stratified sampling 70% for training, 20% for validation, and 10% for testing. The model employed in this phase was flax-community/indonesian-roberta-base, and its hyperparameters were carefully tuned to accommodate the smaller dataset and reduce the risk of overfitting. Table 4 outlines the full configuration used in this setup.

TABLE IV
CONFIGURATION HYPERPARAMETER

Hyperparameter	Configuration
Base Model	flax-community/indonesian-roberta-base
Learning rate	5e-5
Optimizer	AdamW
Batch size	16
Total epoch	5

The fine-tuning in this research utilized the IndoRoBERTa base model "flax-community/indonesian-roberta-base", with the hyperparameter settings summarized in Table 7. A learning rate of 5e-5 was applied to maintain stable convergence, paired with the AdamW optimizer to manage weight decay effectively. Training was carried out with a batch size of 16 over 5 epochs, providing a practical balance between accuracy and computational cost [27].

All experiments run in a GPU supported environment using an NVIDIA GeForce GTX 1650, enabling efficient batch processing for the transformer architecture. These settings are essential for ensuring reproducibility and allow future researchers to replicate the training under comparable conditions.

After completing the training phase, the model was tested on a separate dataset to determine how well it could generalize its understanding of homograph contexts to new, unseen data. The evaluation utilized precision, recall, and F1-score metrics for each of the 20 distinct context categories. The outcomes of this classification performance assessment are summarized in Table 5.

TABLE V
TEST MATRIX DATA EVALUATION RESULTS PHASE 1

Context	Precision	Recall	F1-Score	Support
Berubah warna "Change Color"	1.00	1.00	1.00	2
Binatang "Animal"	1.00	1.00	1.00	2
Buah "Fruit"	0.60	1.00	0.75	3
Bumbu "Seasoning"	1.00	1.00	1.00	2
Gerakan mulut "Mouth Movements"	1.00	1.00	1.00	3
Gerakan terlempar "Thrown Movement"	1.00	1.00	1.00	2
Jenis rujak "Type of Salad"	1.00	1.00	1.00	3
Kalendar "Calendar"	1.00	1.00	1.00	3
Kewenangan "Authority"	1.00	1.00	1.00	2
Kumpul "Gather"	1.00	0.33	0.50	3
Makanan "Food"	1.00	1.00	1.00	3
Melawan "Oppose"	1.00	1.00	1.00	2
Memeras susu "Express Milk"	1.00	1.00	1.00	3
Nama tempat "Place Name"	1.00	1.00	1.00	3

Perbankan “Banking”	1.00	1.00	1.00	3
Rohani “Spiritual”	1.00	1.00	1.00	2
Satelit “Satelite”	1.00	1.00	1.00	2
Sepatu “Shoe”	1.00	1.00	1.00	2
Situasi “Situation”	1.00	1.00	1.00	2
Tanaman “Plant”	1.00	1.00	1.00	3

The evaluation findings indicate that the model excelled at identifying homograph context labels in the test set. Out of 50 samples spread across 20 categories, the majority of classes reached perfect scores for precision, recall, and F1-score. Only two "buah" and "kumpul" experienced a slight decline, recording F1-scores of 0.75 and 0.50, respectively. The particularly low recall for “kumpul” (0.33) points to some difficulty in capturing all instances of that category. Still, performance across the remaining 18 labels stayed consistently strong.

In total, the model scored 96.00% accuracy on the test set. The macro-average F1-score came in at 0.9625, while the weighted-average F1-score was slightly lower at 0.9550. These figures highlight the model’s solid generalization capability when dealing with unfamiliar data especially in conversations centered around early marriage topics.

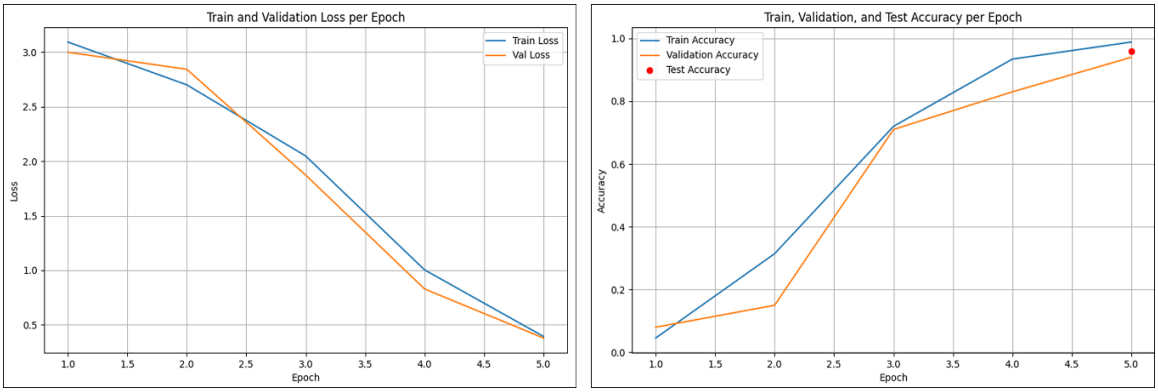

Fig. 3. Plot loss and accuracy phase 1

Figure 3 illustrates the performance progression of the model throughout five epochs of training. The left graph displays a consistent decrease in loss values for both training and validation datasets. Initially, the training loss begins above 3.0 and steadily declines to below 0.4 by the fifth epoch. Similarly, the validation loss starts around 2.9 and decreases significantly to approximately 0.3. This downward trend in both losses suggests an effective and stable learning process without major fluctuation.

The graph on the right illustrates how the model’s accuracy steadily improved over the course of training. During the first epoch, training accuracy started at just 7% but surged past 97% by the fifth epoch. Validation accuracy followed a similar upward trend, starting around 6% and rising above 90% by the end. A red marker at the fifth epoch represents the final test accuracy, which closely mirrors the validation curve.

Taken together, these graphs suggest that the model not only absorbed patterns effectively from the training set but also maintained strong generalization on validation and test sets showing no signs of overfitting during the fine-tuning stage.

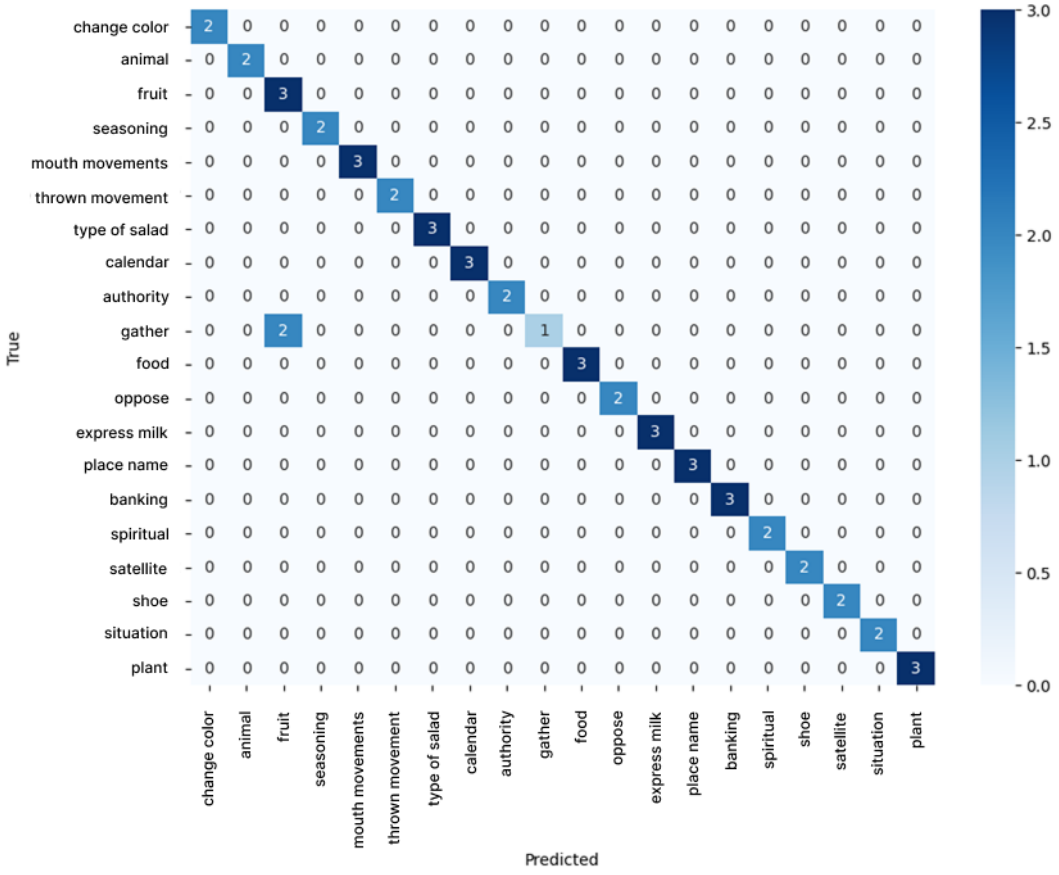

Fig. 4. Confusion matrix phase 1

Figure 4 shows the confusion matrix derived from how the model handled 50 test samples across 20 homograph context categories. The vertical axis reflects the actual labels, while the horizontal axis shows the labels predicted by the model. Each value along the diagonal line represents a correct match between the actual and predicted labels.

Overall, the model demonstrated excellent classification capability, successfully identifying most context categories with minimal error. Many labels such as “mouth movement”, “authority”, “place name”, “banking”, and “plant” were classified perfectly without any mispredictions. Only a few minor misclassifications occurred, such as one instance of “gather” being misclassified as “fruit”, and another “fruit” instance being predicted as “gather”.

This confusion matrix reinforces earlier evaluation results, confirming that the model not only performs with high accuracy but also maintains consistent reliability in distinguishing between homograph meanings based on their contextual usage. The errors are sparse and evenly distributed, further validating the model robustness in classification tasks.

B. Second Phase Fine tuning on 2000 Labeled Data

In the second phase of training, the model underwent additional fine-tuning using 2,000 contextually labeled data points, each categorized into one of 20 homograph context classes. This phase built upon the initial phase, which had utilized only 500 manually labeled examples, by targeting improved interpretation of more varied and syntactically complex sentences. The training inputs included chatbot responses paired with homograph

context labels that were automatically annotated, leveraging the predictions made by the earlier model, as described previously.

This phase aimed to enhance the model's ability to semantically understand and generalize homograph usage across diverse contexts. Fine-tuning was performed using the top-performing model checkpoint from the first phase, following a continual learning approach. Key training modifications included adopting the AdamW optimizer with a learning rate of 1e-5 and extending the training to 5 epochs. Model performance was assessed using a balanced test dataset of 200 samples evenly distributed across the 20 context classes.

TABLE VI
Test Matrix Data Evaluation Results Phase 2

Context	**Precision**	**Recall**	**F1-Score**	**Support**
Berubah Warna "Change Color"	0.86	1.00	0.92	12
Binatang "Animal"	1.00	1.00	1.00	9
Buah "Fruit"	0.94	1.00	0.97	15
Bumbu "Seasoning"	1.00	0.88	0.93	8
Gerakan Mulut "Mouth Movements"	1.00	0.50	0.67	2
Gerakan terlempar "Thrown Movement"	0.80	0.67	0.73	6
Jenis Rujak "Type of Salad"	1.00	0.91	0.95	11
Kalendar "Calendar"	0.67	0.40	0.50	10
Kewenangan "Authority"	0.94	1.00	0.97	15
Kumpul "Gather"	1.00	0.89	0.94	9
Makanan "Food"	0.88	0.94	0.91	16
Melawan "Oppose"	1.00	0.67	0.80	6
Memeras susu "Express Milk"	1.00	0.75	0.86	4
Nama tempat "Place Name"	0.47	0.90	0.62	10
Perbankan "Banking"	0.88	1.00	0.93	7
Rohani "Spiritual"	0.91	0.77	0.83	13
Satelit "Satelite"	1.00	1.00	1.00	10
Sepatu "Shoe"	1.00	1.00	1.00	7
Situasi "Situation"	0.79	0.85	0.81	13
Tanaman "Plant"	1.00	0.88	0.94	17

From the classification report, the model performs robustly across most categories, particularly on classes such as "animal", "fruit", "authority", "shoe", and "satellite", all of which achieved perfect scores. However, lower performance was observed on "mouth movement", "calendar", and "place name", indicating that these contexts may still pose a challenge due to lexical ambiguity or limited representative examples. In total, the second phase scored 88.00% accuracy on the test set.

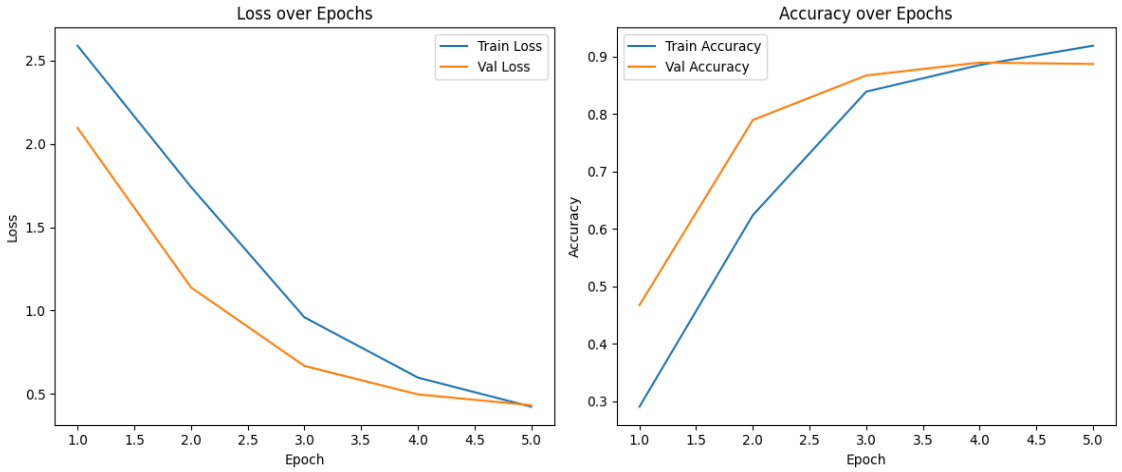

Fig. 5. Plot loss and accuracy phase 2

Figure 5 illustrates the model's training dynamics across five epochs. The left subplot displays the training and validation loss, showing a steady decline from epoch 1 to epoch 5. The training loss dropped significantly from around 2.6 to below 0.5, while the validation loss decreased from approximately 2.1 to 0.45, indicating effective learning.

The right subplot depicts the accuracy per epoch. Training accuracy increased consistently, from 30% in the first epoch to over 90% by the final epoch. Similarly, validation accuracy rose from 47% to 89%, demonstrating stable performance and minimal signs of overfitting.

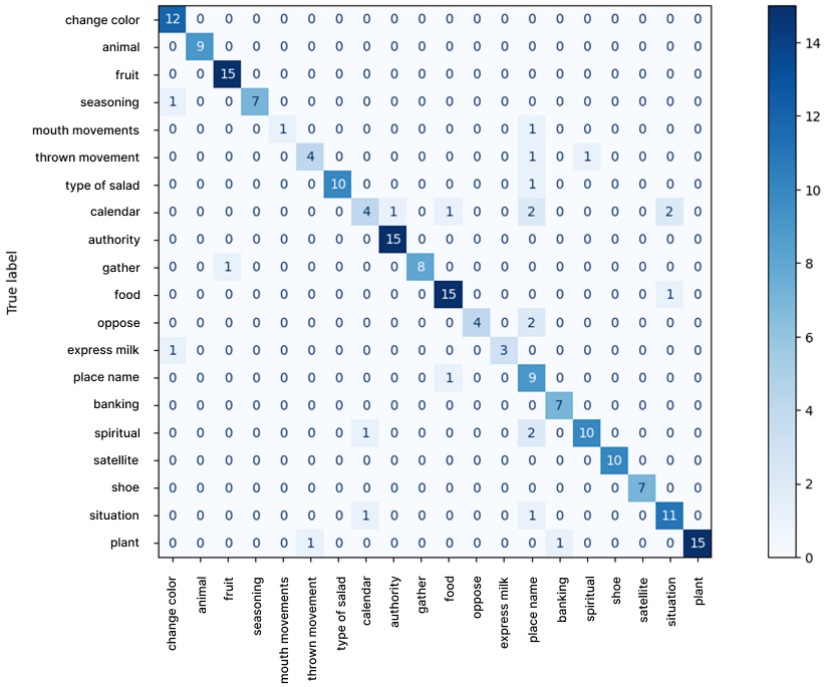

Fig. 6. Confusion matrix phase 2

Figure 6 presents the confusion matrix generated from the model's predictions on a test set containing 200 samples. The diagonal entries represent accurate classifications, where the predicted context aligns with the

actual label. This matrix highlights the model's overall capability to correctly identify most homograph contexts, with several categories such as "fruit", "animal", "shoe", and "satelite" achieving perfect accuracy across all test examples.

Nevertheless, some classification errors remain. Notably, two instances from the "thrown movement" class were misclassified into different categories. The kalendar context exhibited the highest rate of confusion, with only 4 out of 10 samples correctly predicted, suggesting that the model had difficulty confidently recognizing temporal expressions. Likewise, while "place name" category achieved relatively high recall, its predictions were scattered across other classes. Despite these challenges, the model performed robustly overall, confirming the effectiveness of the fine-tuning phase using 2,000 auto annotated entries in improving contextual homograph classification.

C. Comparison Model Performance

To evaluate the effectiveness of IndoRoBERTa in classifying homograph contexts within a Text-to-Speech (TTS) educational chatbot, this study conducts a comparative analysis with other commonly used classification models. This comparison helps place IndoRoBERTa's accuracy in context and highlights its advantages in dealing with complex linguistic subtleties. The benchmark models chosen for comparison include RoBERTa, Naïve Bayes, and Support Vector Machine (SVM), all of which are frequently employed in text classification tasks. Although these models have been tested across various domains and task settings, the comparison sheds light on IndoRoBERTa capabilities in scenarios requiring nuanced, context-aware classification like disambiguating homographs.

TABLE VII
COMPARISON OF MODEL ACCURACY

Model	Accuracy	Study Case
IndoRoBERTa	88%	Context Homograph Classification in TTS for Early Marriage Chatbot
RoBERTa	96.2%	Sentiment Prediction on Corporate University Social Media Texts
Transformer (SoundChoice)	74%	Disambiguating Italian Homographic Heterophones using Text-Based Dataset
Naïve Bayes	71%	A comparison of classification algorithms for hate speech detection
IndoBERT	83.5%	Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models

Based on Table 7, IndoRoBERTa achieved an 88% accuracy, demonstrating its strength in classifying homograph contexts within conversational texts for Text-to-Speech applications. RoBERTa recorded a slightly higher accuracy of 96.2% in sentiment analysis tasks [28], which, although impressive, involves a less complex semantic disambiguation task compared to homograph classification.

The SoundChoice Transformer model reported an accuracy of 74% when applied to Italian homograph heterophone disambiguation, which also relies on a text-based dataset for TTS [29]. This highlights that your

study and the SoundChoice project target a similar objective disambiguating homographs in text to improve speech synthesis though they operate in different.

By comparison, conventional classification approaches such as Naïve Bayes and SVM achieved accuracies of 71% [30] and 83.5% [31] in their respective applications but are less effective for tasks requiring deep semantic interpretation like homograph disambiguation. These findings emphasize IndoRoBERTa strength in capturing subtle contextual meaning within Indonesian TTS, positioning it as an excellent option for systems demanding high semantic accuracy.

D. Comparison Model Performance

Table 7 shows that RoBERTa delivered the highest accuracy at 96.2% in sentiment analysis, a task considered less demanding than homograph classification. IndoRoBERTa, by contrast, reached 88% accuracy in identifying homograph contexts for TTS, proving its strength in grasping fine semantic distinctions in Indonesian language use. On the other hand, the Transformer (SoundChoice) model achieved only 74% when disambiguating Italian homographic heterophones, underlining how difficult homograph handling remains in TTS even with transformer-based designs.

Traditional methods like Naïve Bayes (71%) and IndoBERT (83.5%) in sentiment analysis performed less effectively in areas requiring deeper semantic comprehension. Overall, the results suggest IndoRoBERTa is a robust option for context-sensitive TTS systems, as it offers strong accuracy while more effectively capturing intricate contextual meanings compared to conventional approaches.

V. Conclusion

This research introduces a context-based homograph classification system aimed at improving pronunciation clarity in Indonesian Text-to-Speech (TTS) applications, particularly for chatbots that focus on educating users about early marriage. The system relies on IndoRoBERTa a transformer model trained specifically on Indonesian language data to interpret homograph meanings based entirely on chatbot-generated replies. Training was divided into two phases. The first involved fine-tuning IndoRoBERTa with 500 manually labeled samples, covering 20 context categories from ten homograph words. The second phase extended training using 2,000 additional entries, which were automatically annotated based on the model's earlier predictions. This approach enhanced the model's adaptability to varied conversational styles.

Performance evaluations showed that IndoRoBERTa delivered solid results, achieving an F1-score of 0.86 and 88% accuracy on the second test set. Visualizations like the confusion matrix and training curves indicated consistent learning without signs of overfitting. When compared with other classification models RoBERTa, Naïve Bayes, and IndoBERT, IndoRoBERTa stood out in its ability to grasp the contextual richness of the Indonesian language, despite RoBERTa slightly higher accuracy in other domains. By embedding contextual understanding into the text pipeline, this study pushes forward the quality of speech synthesis in Indonesian. Future work could focus on integrating prosodic features or testing the system in real-time within TTS platforms for educational or counseling use.

Acknowledgment

The author sincerely extends appreciation to Telkom University for the valuable support and opportunities offered throughout the course of this research. Special thanks are also directed to the researchers whose prior works have laid the foundation and inspired the conceptual and methodological direction of this study. Gratitude is likewise expressed to all individuals and institutions who contributed either materially or morally to the successful completion of this research.

References

[1] S. Fan and A. Koski, "The health consequences of child marriage: a systematic review of the evidence," *BMC Public Health*, vol. 22, no. 1, p. 309, Feb. 2022, doi: 10.1186/s12889-022-12707-x.

[2] S. O. Gunawan and S. Bahri, "Impacts of Early Childhood Marriage in Indonesia Viewed from Child Protection Laws Perspectives," *El-Usrah: Jurnal Hukum Keluarga*, vol. 6, no. 2, p. 362, Dec. 2023, doi: 10.22373/ujhk.v6i2.20262.

[3] P. Hariyanti, I. Darmawan, and D. P. Mayangsari, "Child Marriage: An Exploratory Study in Aik Mual, West Lombok, West Nusa Tenggara," *Proceedings of International Conference on Communication Science*, vol. 3, no. 1, pp. 196–201, Jan. 2024, doi: 10.29303/iccsproceeding.v3i1.453.

[4] Supi Yanti, "PENCEGAHAN PERNIKAHAN DINI DAN EDUKASI DIRI," *ALAINA: Jurnal Pengabdian Masyarakat*, vol. 1, no. 1, Jan. 2024, doi: 10.61798/alaina.v1i1.54.

[5] 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.

[6] S. Aminah, "RELIGIOUS AND CULTURAL CONSTRUCTS OF THE SASAK COMMUNITY AGAINST CHILD MARRIAGE PRACTICES," *SANGKéP: Jurnal Kajian Sosial Keagamaan*, vol. 6, no. 2, pp. 167–178, Dec. 2023, doi: 10.20414/sangkep.v6i2.8496.

[7] W. Huang, K. F. Hew, and L. K. Fryer, "Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning," *J. Comput. Assist. Learn.*, vol. 38, no. 1, pp. 237–257, Feb. 2022, doi: 10.1111/jcal.12610.

[8] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12644–12652, Jun. 2023, doi: 10.1609/aaai.v37i11.26488.

[9] D. Hussen Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. Mohammed Sadeeq, and K. Hussein Sharif, "State of Art for Semantic Analysis of Natural Language Processing," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 21–28, Mar. 2021, doi: 10.48161/qaj.v1n2a44.

[10] 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.

[11] M. S. Ribeiro, G. Comini, and J. Lorenzo-Trueba, "Improving grapheme-to-phoneme conversion by learning pronunciations from speech recordings," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.16643

[12] A. R. Hanum *et al.*, "Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, pp. 537–546, Jul. 2024, doi: 10.25126/jtiik.938093.

[13] A. Arisusanto, N. Suarna, and G. Dwilestari, "Analisa Klasifikasi Data Harga Handphone Menggunakan Algoritma Random Forest Dengan Optimize Parameter Grid," *Jurnal Teknologi Ilmu Komputer*, vol. 1, no. 2, pp. 43–47, 2023, doi: 10.56854/jtik.v1i2.51.

[14] 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.

[15] R. Seks Bebas dan Pernikahan Dini Bagi Kesehatan Reproduksi Pada Remaja Shanty Natalia, I. Sekarsari, F. Rahmayanti, and N. Febriani, "Journal of Community Engagement in Health," vol. 4, no. 1, 2021, doi: 10.30994/jceh.v4i1.113.

[16] R. Susilawati, "Upaya Pencegahan Pernikahan Dini Meningkatkan Generasi Berkualitas di Lombok Timur (Studi Kasus UPTD PPA Lombok Timur)," *attaujih*, vol. 1, no. 1, pp. 40–48, Dec. 2022, doi: 10.37216/taujih.v1i1.755.

[17] N. Fitria Aprianti *et al.*, "Nomor 1 Januari," *Indonesian Journal of Community Dedication*, vol. 5, 2023.

[18] T. D. Chala, A. C. Guta, and M. H. Asebel, "Design and Development of a Text-to-Speech Synthesizer for Afan Oromo," *SN Comput. Sci.*, vol. 3, no. 5, Sep. 2022, doi: 10.1007/s42979-022-01306-7.

[19] W. Suwarningsih, R. A. Pratama, F. Y. Rahadika, and M. H. A. Purnomo, "RoBERTa: language modelling in building Indonesian question-answering systems," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 6, pp. 1248–1255, Dec. 2022, doi: 10.12928/TELKOMNIKA.v20i6.24248.

[20] Y. Zhang, A. Warstadt, H.-S. Li, and S. R. Bowman, "When Do You Need Billions of Words of Pretraining Data?," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.04946

[21] 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*, vol. 14, no. 5, pp. 5489–5501, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5489-5501.

[22] M. Saeful, D. Ayu andhirah, P. Pendidikan Guru Sekolah Dasar, and F. Keguruan dan Ilmu Pendidikan, "Representasi Makna Ganda (Homograf)dalam Bahasa Makassar: Studi Linguistik pada Masyarakat di Kelurahan Pattenne Kecamatan Polong Bangkeng Selatan Kabupaten Takalar," 2024, doi: 10.62383/dilan.v1i1.2120.

[23] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," 2024. [Online]. Available: www.ijacsa.thesai.org

[24] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[25] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[26] 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.

[27] 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.

[28] Y. O. Sihombing, N. V. Situmorang, B. K. Negara, and J. M. Sutoyo, "Prediksi Sentimen Pada Teks Media Sosial Corporate University Menggunakan RoBERTa," 2024.

[29] M. Nanni, J. Sjons, and F. Von Kartaschew, "Disambiguating Italian homographic heterophones with SoundChoice and testing ChatGPT as a data-generating tool," 2023.

[30] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 830, no. 3, p. 032006, Apr. 2020, doi: 10.1088/1757-899X/830/3/032006.

[31] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 3, pp. 746–757, Aug. 2023, doi: 10.26555/jiteki.v9i3.26490.