Abstract

This study focuses on enhancing the accuracy of homograph pronunciation in Indonesian Text-to-Speech (TTS) systems through the use of the IndoBERT model. Homographs words with the same spelling but different meanings and, in some cases, pronunciations pose a unique challenge in TTS, especially in voice-based educational chatbots addressing sensitive topics like early marriage. The research was carried out in two training phases. In the first phase, IndoBERT was fine-tuned on 500 contextually annotated samples, achieving 98% accuracy with near-perfect F1-scores across all categories. The second phase incorporated an additional 2,000 samples labeled automatically, after which the model was tested on 200 samples. At this phase, the model achieved 97% accuracy, demonstrating strong capability in identifying homograph contexts even with more complex input. Performance assessments, including confusion matrix analysis and training curves, indicated steady accuracy gains and a consistent reduction in loss throughout the process. This work delivers a multi-class homograph classification framework that enhances the contextual accuracy of TTS systems, enabling more precise and meaningful speech generation in educational chatbots. Nonetheless, the findings also highlight limitations—automatically labeled datasets may introduce errors that impact both model performance and the naturalness of TTS output. Despite this, the approach offers significant potential for expanding voice-based services, particularly in regions with limited access to reliable information.

Keywords: homograph, child marriage, indoBERT, text-to-speech, chatbot.
