

ABSTRACT

Proper writing in Indonesian is essential to ensure that messages are clearly understood. However, various texts, such as academic papers, social media content, and news articles, often contain grammatical errors, especially in syntax and morphology. Mistakes like missing subjects or predicates, illogical sentence structures, and incorrect use of prepositions and conjunctions can cause misunderstandings, while automatic grammar error detection tools remain limited. This study developed a deep learning model based on the Transformer architecture using only encoder layers for automatic grammar error detection. The dataset consisted of one million Indonesian sentences. The process involved data pre-processing, including normalization through filtering and cleaning, synthetic sentence generation with incorrect grammar, tokenization using SentencePiece with Unigram Language Model, and token labeling, followed by training, validation, and testing. Transformer models with 1, 2, 3, and 6 encoder layers were compared against a Bi-LSTM model with self-attention. The Transformer with 6 encoder layers achieved the best performance, with 93.46% accuracy and a 75.08% F1-score at the token level, and 78.82% accuracy and a 78.38% F1-score at the sentence level. These results indicate that deeper encoder layers enhance grammar error detection, and token order significantly impacts sentence-level accuracy.

Keywords: sentence, grammar, syntax, morphology, Transformer, encoder layer