

ABSTRACT

Health complaints submitted by pregnant women and infants through online health consultation platforms contain valuable information that can be utilized to support early health risk detection. This study aims to develop a text-based health risk classification model for maternal and infant complaints using IndoBERT combined with Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The dataset was collected from the Alodokter health forum and categorized into three risk levels: low, medium, and high, following the risk categories used in the UCI Maternal Health Risk dataset. The research process consisted of data preprocessing, data augmentation, tokenization, hyperparameter optimization, as well as model training and evaluation. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The evaluation results show that the IndoBERT-CNN model achieved an accuracy of 80.15% and a weighted F1-score of 80.03%, while the IndoBERT-BiLSTM model achieved an accuracy of 83.90% and a weighted F1-score of 83.84%. The IndoBERT-BiLSTM model outperformed the IndoBERT-CNN model across all evaluation metrics, indicating that BiLSTM's ability to process text sequences bidirectionally is more effective in capturing the contextual information of health complaints than the local pattern extraction performed by CNN.

Keywords: IndoBERT, CNN, BiLSTM, risk classification, maternal and infant health, natural language processing