ABSTRACT

OPTIMIZATION OF BIOBART MODEL FINE-TUNING USING ADAMW OPTIMIZER IN AUTOMATIC TEXT SUMMARIZATION OF BIOMEDICAL SCIENTIFIC ARTICLES WITH ROUGE AND BERTSCORE

By

Meilina Risdianti

In general, health and biomedical articles have technical and specialized content that makes it difficult for the general public or non-experts to understand and filter information in articles that tend to be long and complex. The increasing number and complexity of biomedical literature, especially in long scientific documents, makes NLP (Natural Language Processing) an important approach in decision making. NLP is capable of making long texts more concise, relevant, and informative through automatic text summarization. In this study, the BioBART model was optimized using the AdamW Optimizer with a fine-tuning process to improve the quality of summaries of long biomedical articles by utilizing the PMC (PubMed Central) dataset, which is known to have a high level of heterogeneity in terms of document length, article structure variation, and terminology complexity. A dataset of 10,000 articles, with 90% for the model and 10% for testing. Of the 9,000 model data, 75% is training data and 25% is validation data. The research process consisted of a preprocessing stage involving text cleaning, normalization, and the application of a sliding window to handle long documents. Then, the BioBART model was fine-tuned by exploring hyperparameters to obtain the best model configuration. Model evaluation was performed using the ROUGE and BERTScore metrics. The results of the study show that the fine-tuned and hyperparameter-optimized BioBART model is capable of producing better and more competitive summaries than the pretrained BioBART model, both lexically and semantically, despite the complexity of long documents and heterogeneity in the PMC dataset.

Keywords: PMC, Biomedical Articles, NLP, BioBART Model, AdamW Optimizer, Fine-Tuning, ROUGE, BERTScore, Automatic Text Summarization.

ABSTRAK

OPTIMALISASI *FINE-TUNING* MODEL BIOBART MENGGUNAKAN ADAMW *OPTIMIZER* DALAM PERINGKASAN TEKS OTOMATIS PADA ARTIKEL ILMIAH BIOMEDIS DENGAN EVALUASI ROUGE DAN BERTSCORE

Oleh

Meilina Risdianti

Secara umum, artikel kesehatan dan biomedis memiliki konten teknis dan spesialis yang menyebabkan pembaca non-ahli kesulitan memahami dan menyaring informasi dalam artikel yang cenderung panjang dan kompleks. Meningkatnya jumlah dan kompleksitas literatur biomedis, khususnya pada dokumen ilmiah yang panjang (*long documents*), menjadikan NLP (*Natural Language Processing*) sebagai pendekatan dalam pengambilan keputusan. NLP mampu menjadikan teks panjang menjadi lebih ringkas, relevan, dan informatif melalui *automatic text summarization*. Dalam penelitian ini, model BioBART dioptimalkan menggunakan AdamW *Optimizer* dengan *fine-tuning* untuk meningkatkan kualitas ringkasan teks artikel biomedis panjang dengan memanfaatkan dataset PMC (*PubMed Central*) yang dikenal memiliki tingkat heterogenitas tinggi dari sisi panjang dokumen, variasi struktur artikel, dan kompleksitas terminologi. Data sebanyak 10000 artikel, dengan 90% untuk model dan 10% pengujian. Dari 90% data model dibagi menjadi 75% data *training* dan 25% data *validation*. Proses penelitian terdiri dari tahap *preprocessing* yaitu pembersihan teks, normalisasi, serta penerapan *sliding window* untuk menangani dokumen panjang. Kemudian proses *fine-tuning* model BioBART dengan eksplorasi *hyperparameter* untuk memperoleh konfigurasi model terbaik. Evaluasi model dilakukan menggunakan metrik ROUGE dan BERTScore. Hasil penelitian menunjukkan bahwa model BioBART hasil *fine-tuning* dan optimalisasi hyperparameter mampu menghasilkan ringkasan yang lebih baik dan kompetitif dibandingkan model BioBART *pretrained*, baik secara leksikal maupun semantik, meskipun dihadapkan pada kompleksitas dokumen panjang dan heterogenitas pada dataset PMC.

Kata-kata kunci: PMC, Artikel Biomedis, NLP, Model BioBART, AdamW *Optimizer*, *Fine-Tuning*, ROUGE, BERTScore, *Automatic Text Summarization*.