ABSTRAK

Perkembangan model *Transformer* telah mendorong kemajuan pada tugas *abstractive text summarization*, yaitu menghasilkan ringkasan dengan menyusun kembali informasi penting menggunakan kalimat yang lebih ringkas dan alami. Meskipun berbagai arsitektur *encoder-decoder* telah dikembangkan, penelitian yang membandingkan penggunaan decoder non-generatif dan decoder generatif pada tugas *abstractive summarization* Bahasa Indonesia masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis dan membandingkan performa model BERT2BERT dan BERT2GPT dari aspek kualitas konten, keterbacaan, dan konsistensi faktual (*faithfulness*). Penelitian ini mengadaptasi kerangka kerja Cross-Industry Standard Process for Data Mining (CRISP-DM) yang meliputi tahapan *business understanding*, *data understanding*, *data preparation*, *modelling*, dan *evaluation*. Penelitian menggunakan Liputan6 Indonesian Summarization Dataset sebagai data pelatihan dan pengujian model. Evaluasi dilakukan menggunakan metrik ROUGE, BERTScore, *Flesch-Kincaid Grade Level* (FKGL), *Gunning Fog Index* (GFI), Formula Dwiyanto Djoko Pranowo, serta pendekatan *LLM-as-a-Judge* yang diadaptasi dari HaluEval untuk mengevaluasi konsistensi faktual. Hasil penelitian menunjukkan bahwa model BERT2BERT memperoleh performa yang lebih baik dibandingkan BERT2GPT, dengan nilai ROUGE-1 sebesar 0,3964, ROUGE-2 sebesar 0,2206, dan ROUGE-L sebesar 0,3302, serta menghasilkan tingkat konsistensi faktual yang lebih baik berdasarkan pendekatan evaluasi yang digunakan. Penelitian ini memberikan kontribusi berupa analisis komparatif terhadap dua arsitektur *encoder-decoder* pada tugas *abstractive summarization* Bahasa Indonesia serta memberikan gambaran mengenai karakteristik masing-masing arsitektur.

Kata Kunci: *Abstractive Text Summarization*, BERT2BERT, BERT2GPT, *Transformer*, *LLM-as-a-Judge*.

ABSTRACT

The advancement of Transformer models has significantly improved the performance of abstractive text summarization, which aims to generate concise and natural summaries by rewriting the essential information from source documents. Although various encoder-decoder architectures have been proposed, studies comparing non-generative and generative decoders for Indonesian abstractive summarization remain limited. Therefore, this study aims to analyze and compare the performance of BERT2BERT and BERT2GPT in terms of content quality, readability, and factual consistency (faithfulness). This research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, covering the stages of business understanding, data understanding, data preparation, modelling, and evaluation. The Liputan6 Indonesian Summarization Dataset was used for model training and evaluation. Performance was assessed using ROUGE, BERTScore, Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Dwiyanto Djoko Pranowo Formula (DDP), and an LLM-as-a-Judge approach adapted from HaluEval to evaluate factual consistency. The results indicate that BERT2BERT outperformed BERT2GPT, achieving ROUGE-1 of 0.3964, ROUGE-2 of 0.2206, and ROUGE-L of 0.3302, while also demonstrating better factual consistency based on the adopted evaluation approach. This study contributes a comparative analysis of two encoder-decoder architectures for Indonesian abstractive text summarization and provides insights into the characteristics of each architecture.

Keywords: *Abstractive Text Summarization, BERT2BERT, BERT2GPT, Transformer, LLM-as-a-Judge.*