INTISARI

PERINGKASAN TEKS OTOMATIS DENGAN PARADIGMA *EXTRACT-THEN-ABSTRACT* MENGGUNAKAN DISTILBERT DAN BART

Oleh
Fidzal Adrian
21/480604/PA/20882

Peringkasan teks otomatis merupakan salah satu bidang dalam pemrosesan bahasa alami *(Natural Language Processing)* yang bertujuan menyajikan informasi penting dari sebuah dokumen dalam bentuk ringkas, koheren, dan mudah dipahami. Penelitian ini bertujuan untuk mengembangkan model *text summarization* menggunakan paradigma *extract-then-abstract* yang terdiri dari dua tahap, yaitu tahap ekstraksi menggunakan model DistilBERT dan tahap abstraksi menggunakan model BART. Paradigma ini digunakan untuk mengatasi masalah kurangnya koherensi atau hilangnya makna pada metode ekstraktif dan *computational cost* pada metode abstraktif. Dataset yang digunakan dalam penelitian ini adalah CNN/DM yang merupakan salah satu dataset standar untuk *text summarization*. Metode ini dievaluasi menggunakan metrik ROUGE *(Recall-Oriented Understudy for Gisting Evaluation)* dan BERTScore untuk mengukur kualitas ringkasan yang dihasilkan, serta evaluasi pengukuran waktu komputasi. Hasil penelitian menunjukkan bahwa paradigma *extract-then-abstract* mampu mencapai ROUGE-1 sebesar 41,73, ROUGE-2 sebesar 19,18, ROUGE-L sebesar 38,82, serta BERTScore F1 sebesar 88,60. Selain itu, penggunaan DistilBERT mampu mempercepat waktu inferensi hingga 47-53% dibandingkan model besar seperti BERT. Dengan demikian, paradigma *extract-then-abstract* berbasis DistilBERT dan BART yang dikembangkan tidak hanya kompetitif dari sisi kualitas semantik dan efisiensi komputasi, tetapi juga mampu menghasilkan ringkasan yang relevan, koheren, dan alami.

Kata Kunci: Text Summarization, NLP, *Extract-then-Abstract*, DistilBERT, BART

ABSTRACT

AUTOMATIC TEXT SUMMARIZATION WITH EXTRACT-THEN-ABSTRACT PARADIGM USING DISTILBERT AND BART

By
Fidzal Adrian
21/480604/PA/20882

Automatic text summarization is a field in natural language processing (NLP) that aims to present important information from a document in a concise, coherent, and easy-to-understand form. This study aims to develop a text summarization model using the extract-then-abstract paradigm, which consists of two stages: the extraction stage using the DistilBERT model and the abstraction stage using the BART model. This paradigm is used to overcome the problems of lack of coherence or loss of meaning in extractive methods and computational cost in abstractive methods. The dataset used in this study is CNN/DM, which is one of the standard datasets for text summarization. This method was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BERTScore metrics to measure the quality of the summaries produced, as well as computational time measurements. The results show that the extract-then-abstract paradigm is capable of achieving ROUGE-1 of 41.73, ROUGE-2 of 19.18, ROUGE-L of 38.82, and BERTScore F1 of 88.60. Additionally, the use of DistilBERT was able to accelerate inference time by 47-53% compared to large models such as BERT. Thus, the extract-then-abstract paradigm based on DistilBERT and BART that has been developed is not only competitive in terms of semantic quality and computational efficiency, but also capable of producing relevant, coherent, and natural summaries.

Keywords: Text Summarization, NLP, Extract-then-Abstract, DistilBERT, BART