ABSTRAK

Konten tekstual yang ada di internet menyebabkan orang-orang mengonsumsi lebih banyak waktu untuk menemukan informasi yang mereka inginkan. Untuk mengatasi hal tersebut perlu adanya sistem peringkasan terhadap sebuah teks agar pembaca dapat secara efisien menemukan informasi utama dalam teks yang dibaca. Sistem tersebut mengadopsi ilmu dari *Natural Language Processing* untuk dapat memproses data tekstual dalam sebuah komputer.

Penelitian ini menggunakan TextRank sebagai metode peringkasannya dengan BERT sebagai metode *sentences embedding* atau vektorisasi terhadap kalimat. Dataset yang digunakan adalah Indosum yang berisi sebanyak 100 teks dengan pasangan ringkasannya yang digunakan untuk evaluasi hasil peringkasan. Pada penerapannya, TextRank menggunakan *cosine similarity* untuk menghitung bobot, sehingga perlu adanya metode vektorisasi yang dapat secara khusus mentransformasikan kalimat kedalam sebuah vektor yang memiliki makna informasi yang kuat, berdasarkan hal tersebut diusulkanlah BERT. BERT dapat mentransformasikan kalimat kedalam sebuah vektor dengan memperhatikan konteks kalimat tersebut. Metodologi penelitian ini dimulai dari studi literatur, analisis masalah, pengumpulan data, pengolahan data, pre-processing, membuat model TextRank, evaluasi model, hasil, dan laporan.

Adapun hasil perbandingan dari penggunaan metode TextRank yang diimplementasikan BERT dengan metode vektorisasi lain seperti TFIDF dan Word2Vec dimana BERT mendapatkan nilai tertinggi dengan perolehan ROUGE-1 sebesar 0.48, ROUGE-2 sebesar 0.38, dan ROUGE-L sebesar 0.47. Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwa penggunaan BERT mampu meningkatkan performa dari TextRank dari perolehan nilai ROUGE.

Kata kunci: TextRank, BERT, peringkasan teks

ABSTRACT

Textual content that exists on the internet causes people to consume more time to find the information they want. To overcome this, it is necessary to have a summary system for a text so that readers can efficiently find the main information in the text they are reading. The system adopts knowledge from Natural Language Processing to be able to process textual data in a computer.

This study uses TextRank as a summary method with BERT as a sentence embedding method or vectorization of sentences. The dataset used is Indosum which contains as many as 100 texts with their summary pairs used to evaluate the results of the summary. In its application, TextRank uses cosine similarity to calculate weights, so it is necessary to have a vectorization method that can specifically transform sentences into a vector that has strong informational meaning, based on this, BERT is proposed. BERT can transform sentences into a vector by paying attention to the context of the sentence. The research methodology starts from literature study, problem analysis, data collection, data processing, pre-processing, creating TextRank models, model evaluation, results, and reports.

The comparison results from using the TextRank method implemented by BERT with other vectorization methods such as TFIDF and Word2Vec where BERT gets the highest score with the acquisition of ROUGE-1 of 0.48, ROUGE-2 of 0.38, and ROUGE-L of 0.47. Based on the research that has been done, it can be concluded that the use of BERT can improve the performance of TextRank from the acquisition of ROUGE values.

***Keywords**: TextRank, BERT, text summarization*