

ABSTRAK

Pertumbuhan jumlah berita digital yang masif setiap harinya menyebabkan informasi berlebih (*information overload*) yang menyulitkan pembaca dalam menangkap inti informasi secara efisien. Untuk menjawab tantangan tersebut, teknologi peringkasan teks otomatis diperlukan guna menyajikan informasi penting dalam bentuk yang lebih ringkas. Salah satu metode yang sering digunakan adalah Latent Semantic Analysis (LSA), terutama dalam kombinasi dengan pembobotan TF-IDF. Namun, metode ini memiliki keterbatasan dalam memahami makna kata secara kontekstual. Sementara itu, *embedding* kontekstual seperti BERT-Base Indonesian memiliki potensi besar untuk menghasilkan representasi kalimat yang lebih bermakna, namun belum banyak diteliti dalam konteks integrasi dengan LSA untuk bahasa Indonesia. Oleh karena itu, penelitian ini bertujuan untuk membandingkan kualitas ringkasan antara pendekatan LSA berbasis TF-IDF dan LSA berbasis *embedding* BERT-Base Indonesian, khususnya dalam kaitannya dengan variasi *compression rate*.

Penelitian ini menggunakan data dari dataset IndoSum yang terdiri dari ribuan dokumen berita berbahasa Indonesia. Setiap kalimat pada dokumen diubah menjadi representasi vektor menggunakan dua pendekatan berbeda, yakni TF-IDF dan BERT-Base Indonesian 522M. Representasi tersebut kemudian diproses menggunakan metode LSA untuk memilih kalimat-kalimat penting dalam dokumen dan membentuk ringkasan berdasarkan rasio kompresi yang bervariasi antara 0,1 hingga 0,8. Evaluasi kualitas ringkasan dilakukan dengan membandingkan hasil ringkasan sistem terhadap ringkasan referensi manusia menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L.

Hasil penelitian menunjukkan bahwa pendekatan LSA berbasis *embedding* BERT-Base Indonesian memberikan performa lebih baik dibandingkan LSA berbasis TF-IDF. Nilai ROUGE-1 tertinggi sebesar 0,4101 dicapai pada *compression rate* 0,4, ROUGE-2 tertinggi sebesar 0,3248 pada *compression rate* 0,7, dan ROUGE-L mencapai 0,3521 juga pada *compression rate* 0,7. Sementara itu, pendekatan LSA berbasis TF-IDF memperoleh nilai ROUGE-1 tertinggi sebesar 0,4012 pada *compression rate* 0,4, ROUGE-2 sebesar 0,3260 pada *compression rate* 0,8, dan ROUGE-L sebesar 0,3515 pada *compression rate* 0,6. Temuan ini memperkuat pentingnya pemilihan parameter kompresi yang tepat, dan menunjukkan bahwa integrasi representasi kontekstual berbasis BERT mampu meningkatkan kualitas ringkasan dalam aspek kebermaknaan dan struktur kalimat. Penelitian ini juga memberikan dasar empiris untuk mengarahkan pengembangan sistem peringkasan otomatis berbahasa Indonesia yang lebih adaptif dan akurat.

Kata kunci: *compression rate*, Latent Semantic Analysis, TF-IDF, BERT-base Indonesian, ROUGE

ABSTRACT

The massive growth of digital news in recent years has caused an information overload, making it difficult for readers to efficiently grasp key points in the text. To address this issue, automatic text summarization technologies are needed to present essential information in a more concise form. One widely used method is Latent Semantic Analysis (LSA), especially in combination with TF-IDF weighting. However, this method is limited in its ability to understand contextual word meaning. On the other hand, contextual embeddings such as BERT-Base Indonesian offer significant potential for generating more meaningful sentence representations but have rarely been investigated in combination with LSA for the Indonesian language. Therefore, this study aims to compare summary quality between LSA based on TF-IDF and LSA based on BERT-Base Indonesian embeddings, particularly in relation to varying compression rates.

The research uses the IndoSum dataset, which contains thousands of Indonesian news articles. Each sentence in a document is converted into a vector representation using two approaches: TF-IDF and the pretrained BERT-Base Indonesian 522M model. These representations are then processed using LSA to select key sentences and generate summaries based on different compression rates, ranging from 0.1 to 0.8. Summary quality is evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics by comparing system-generated summaries with human-written references.

The results show that LSA based on BERT-Base Indonesian embeddings outperforms the TF-IDF-based LSA approach. The highest ROUGE-1 score (0.4101) was achieved at a compression rate of 0.4, ROUGE-2 reached 0.3248 at 0.7, and ROUGE-L peaked at 0.3521 at the same rate. In comparison, the TF-IDF-based LSA model achieved its best ROUGE-1 score of 0.4012 at a compression rate of 0.4, ROUGE-2 of 0.3260 at 0.8, and ROUGE-L of 0.3515 at 0.6. These findings underscore the importance of selecting an appropriate compression rate and show that integrating BERT-based contextual embeddings enhances both the semantic and structural quality of summaries. This research contributes empirical insights for the development of more effective and adaptive Indonesian summarization systems.

Keywords: *compression rate, extractive summarization, Latent Semantic Analysis, TF-IDF, BERT-base Indonesian, ROUGE*