

ABSTRAK

Perkembangan media online menyebabkan meningkatnya jumlah berita yang membahas peristiwa serupa dari berbagai sudut pandang. Kondisi ini menimbulkan informasi yang tumpang tindih dan redundan sehingga pembaca harus membaca banyak artikel untuk memahami satu topik secara utuh. Penelitian ini bertujuan mengembangkan sistem otomatis yang mampu mengelompokkan berita berdasarkan kesamaan topik serta menghasilkan ringkasan multi-dokumen yang informatif dan minim redundansi. Proses pengelompokan dilakukan menggunakan *Sentence-BERT* (SBERT) untuk menghasilkan representasi *embedding* dokumen, kemudian tingkat kemiripan dihitung menggunakan *Cosine Similarity* dan dikelompokkan menggunakan *Graph Clustering*. Setelah terbentuk klaster berita dengan topik serupa, dilakukan proses *extractive summarization* menggunakan kombinasi empat fitur, yaitu relevansi berbasis *centroid*, *novelty*, posisi kalimat, dan *keyword* berbasis YAKE. Dataset penelitian diperoleh melalui *web scraping* portal berita CNN Indonesia dan Detik.com. Kualitas ringkasan dievaluasi menggunakan metrik ROUGE dan BERTScore. Hasil penelitian menunjukkan bahwa sistem utama yang mengombinasikan keempat fitur menghasilkan nilai ROUGE-1 sebesar 0,4980, ROUGE-2 sebesar 0,2630, dan ROUGE-L sebesar 0,4710, dengan rata-rata F1 sebesar 0,4107. Nilai ROUGE-1 yang melebihi ambang batas 0,48 menunjukkan bahwa kualitas ringkasan termasuk kategori A+. Selain itu, evaluasi semantik menggunakan BERTScore menunjukkan bahwa penerapan pendekatan *greedy sequential novelty* dengan penurunan *novelty threshold* dari $\tau = 0,95$ menjadi $\tau = 0,70$ meningkatkan kualitas ringkasan, dengan BERTScore F1 tertinggi sebesar 0,8067. Hasil tersebut menunjukkan bahwa kombinasi pengelompokan berbasis SBERT dengan fitur *centroid*, *novelty*, posisi kalimat, dan *keyword* berbasis YAKE mampu menghasilkan ringkasan multi-dokumen berita berbahasa Indonesia yang lebih akurat, informatif, dan minim redundansi.

Kata kunci: Multi-document Summarization, Sentence-BERT, Graph Clustering, Cosine Similarity, YAKE, ROUGE, BERTScore.

ABSTRACT

The development of online media has led to an increasing number of news stories discussing similar events from various perspectives. This condition creates overlapping and redundant information, requiring readers to read multiple articles to fully understand a single topic. This study aims to develop an automated system capable of grouping news based on topical similarities and producing informative multi-document summaries with minimal redundancy. The clustering process was carried out using Sentence-BERT (SBERT) to generate document embedding representations, then the similarity level was calculated using Cosine Similarity and grouped using Graph Clustering. After forming news clusters with similar topics, an extractive summarization process was carried out using a combination of four features: centroid-based relevance, novelty, sentence position, and YAKE-based keywords. The research dataset was obtained through web scraping of the CNN Indonesia and Detik.com news portals. The quality of the summaries was evaluated using the ROUGE and BERTScore metrics. The results showed that the main system combining the four features produced a ROUGE-1 score of 0.4980, ROUGE-2 score of 0.2630, and ROUGE-L score of 0.4710, with an average F1 score of 0.4107. A ROUGE-1 score exceeding the threshold of 0.48 indicates that the summary quality is in the A+ category. Furthermore, semantic evaluation using BERTScore showed that applying the greedy sequential novelty approach by decreasing the novelty threshold from $\tau = 0.95$ to $\tau = 0.70$ improved the summary quality, with the highest BERTScore F1 score of 0.8067. These results indicate that the combination of SBERT-based clustering with YAKE-based centroid, novelty, sentence position, and keyword features is capable of producing multi-document summaries of Indonesian-language news that are more accurate, informative, and minimize redundancy.

Keywords: *Multi-document Summarization, Sentence-BERT, Graph Clustering, Cosine Similarity, YAKE, ROUGE, BERTScore.*