

ABSTRACT

Audrey Roselian (01082210018)

COMPARATIVE ANALYSIS OF ABSTRACTIVE AND EXTRACTIVE SUMMARIZATION METHODS WITH THE FLAN-T5 MODEL ON NEWS TEXTS

(xvii + 158 pages: 34 figures; 16 tables; 4 appendices)

In the digital era, news has become an integral part of daily life, with the internet serving as the primary source of information. According to data from the Ministry of Communication and Informatics (Kominfo) in January 2024, approximately 79.5% or around 221 million Indonesians are internet users. As the flow of information accelerates, the number of circulating news articles continues to increase, making it difficult for readers to filter and comprehend relevant information within a limited time. Therefore, methods that can efficiently summarize news while preserving essential information are necessary.

To determine the most effective approach for generating accurate and easily understandable news summaries through both quantitative and qualitative analysis, this study compares two summarization methods: abstractive and extractive. The abstractive method generates summaries by reformulating the original text using natural language processing models, whereas the extractive method selects important sentences directly from the original text. The implementation of these methods involves tokenization, modeling with grid search for hyperparameter tuning in FLAN-T5 for abstractive summarization, as well as sentence embedding, segmentation using Stanza, and clustering with KMeans based on cosine similarity for FLAN-T5 extractive summarization.

This study successfully fine-tuned the FLAN-T5 model for both abstractive and extractive summarization and compared their performance. For abstractive summarization, the model was fine-tuned using a learning rate of $2e-5$ for 9 epochs with a dropout of 0.2. Extractive summarization utilized sentence embeddings with FLAN-T5, segmentation using Stanza, and clustering with KMeans. The quality of the summarization was evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore metrics. The results indicate that the extractive method outperformed the abstractive method, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 78.35, 70.64, and 77.98, respectively, along with a BERTScore of 86.71. In contrast, abstractive summarization recorded lower scores of 56.69, 43.88, 52.79, and a BERTScore of 81.13. Although the abstractive approach has the potential for creativity, the extractive method proves to be more effective in terms of quality, time efficiency, and accuracy.

Reference: 76 (2004 – 2024)

ABSTRAK

Audrey Roselian (01082210018)

ANALISIS KOMPARATIF METODE PERINGKASAN ABSTRAKTIF DAN EKSTRAKTIF DENGAN MODEL FLAN-T5 PADA TEKS BERITA

(xvii + 158 halaman: 34 gambar; 16 tabel; 4 lampiran)

Di era digital, berita telah menjadi bagian tak terpisahkan dalam kehidupan sehari-hari, dengan internet sebagai sumber utama informasi. Berdasarkan data Kementerian Komunikasi dan Informatika (Kominfo) pada Januari 2024, sebanyak 79,5% atau sekitar 221 juta penduduk Indonesia merupakan pengguna internet. Seiring dengan pesatnya arus informasi, jumlah berita yang beredar semakin banyak, sehingga pembaca sering mengalami kesulitan dalam menyaring dan memahami informasi yang relevan dalam waktu yang terbatas. Oleh karena itu, diperlukan metode yang dapat membantu merangkum berita secara efisien tanpa menghilangkan esensi informasi utama.

Untuk menentukan pendekatan yang lebih efektif dalam menghasilkan ringkasan berita yang akurat dan mudah dipahami dengan analisis kuantitatif dan kualitatif sehingga penelitian ini membandingkan dua metode peringkasan, yaitu abstraktif dan ekstraktif. Metode abstraktif menghasilkan ringkasan dengan mereformulasi teks asli menggunakan pemodelan berbasis bahasa alami, sedangkan metode ekstraktif memilih kalimat penting langsung dari teks asli. Implementasi metode ini melibatkan tahapan tokenisasi, pemodelan dengan *grid search* untuk *tuning* hyperparameter pada FLAN-T5 abstraktif, serta embedding kalimat, segmentasi menggunakan Stanza, dan klusterisasi dengan KMeans berbasis *cosine similarity* untuk FLAN-T5 ekstraktif.

Penelitian ini berhasil melakukan *fine-tuning* model FLAN-T5 untuk peringkasan abstraktif dan ekstraktif, serta membandingkan kinerjanya. Untuk peringkasan abstraktif, model di-*fine-tune* dengan learning rate $2e-5$ selama 9 epoch dan dropout 0,2. Peringkasan ekstraktif menggunakan embeddings kalimat dengan FLAN-T5, segmentasi menggunakan Stanza, dan klusterisasi dengan KMeans. Kualitas peringkasan dinilai menggunakan metrik ROUGE-1, ROUGE-2, ROUGE-L, dan BERTScore. Hasil menunjukkan bahwa metode ekstraktif unggul dengan nilai ROUGE-1, ROUGE-2, dan ROUGE-L masing-masing 78,35, 70,64, dan 77,98, serta BERTScore 86,71. Sebaliknya, peringkasan abstraktif mencatat nilai yang lebih rendah, yaitu 56,69, 43,88, 52,79, dan BERTScore 81,13. Meskipun abstraktif memiliki potensi kreativitas, metode ekstraktif terbukti lebih efektif dalam hal kualitas, efisiensi waktu, dan akurasi.

Referensi: 76 (2004 – 2024)