

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di zaman sekarang, informasi memegang peranan yang sangat penting dalam kehidupan manusia. Melalui perkembangan teknologi dan internet, siapa saja bisa mendapatkan berita atau artikel dengan cepat, kapan pun dan di mana pun. Setiap hari jumlah berita *online* terus bertambah dari berbagai portal berita nasional maupun internasional. Melimpahnya informasi ini membuat masyarakat semakin mudah memperoleh pengetahuan baru yang bermanfaat untuk berbagai hal, seperti pendidikan, pekerjaan, hingga pengambilan keputusan sehari-hari. Kehadiran berita juga membantu masyarakat untuk selalu mengikuti perkembangan peristiwa terkini, sehingga informasi semakin tidak bisa dipisahkan dari aktivitas modern saat ini.(Alfin et al., n.d.-a)

Salah satu permasalahan yang sering muncul dalam berita *online* adalah banyaknya artikel yang membahas topik yang sama, tetapi dengan sudut pandang, detail, atau penekanan yang berbeda-beda. Hal ini membuat pembaca sering menemukan informasi yang tumpang tindih, sehingga isi berita terasa berulang. Masalahnya, setiap artikel biasanya hanya menampilkan sebagian informasi, tidak semuanya lengkap. Akibatnya, untuk memahami satu peristiwa secara menyeluruh, pembaca harus membuka banyak berita sekaligus dari berbagai sumber. Proses ini tentu memakan waktu lebih lama dan membuat pembaca mudah merasa lelah. Selain itu, banyaknya informasi yang berulang justru bisa menimbulkan kebingungan. Pembaca jadi kesulitan untuk memilah mana informasi yang benar-benar penting dan mana yang hanya pengulangan dari berita lain. Dalam beberapa kasus, inti dari peristiwa tidak tersaji dengan jelas dalam satu bacaan, sehingga pembaca harus terus membandingkan antar artikel. Kondisi ini tidak hanya menghambat pemahaman, tetapi juga membuat pengalaman membaca berita menjadi kurang efisien dan tidak praktis(Fauzi & Abidin, n.d., 2024).

Untuk mengatasi masalah tumpang tindih dan banyaknya informasi dari berbagai berita, diperlukan suatu cara agar pembaca bisa mendapatkan inti informasi dengan lebih cepat dan mudah. Salah satu solusi yang dapat digunakan ialah pengelompokan multi-dokumen dengan topik berita yang sama lalu

peringkasan teks otomatis, dimana keduanya bagian dari bidang *text mining*. Dengan pengelompokan multi-dokumen dan peringkasan ini, sistem dapat mengolah kumpulan berita yang membahas topik serupa lalu menyajikan ringkasan singkat yang berisi informasi penting saja. Dengan begitu, pembaca tidak perlu membuka banyak artikel satu per satu, karena inti peristiwa sudah dirangkum secara efisien dan efektif (Dimas et al., 2022).

Pengelompokan multi-dokumen adalah proses mengelompokkan sejumlah dokumen ke dalam kelompok atau topik yang sama berdasarkan kemiripan makna di dalamnya. Menurut (L. Zhang et al., 2023), pengelompokan dokumen modern tidak lagi hanya mengandalkan *bag-of-words* seperti LDA, tetapi memanfaatkan *contextual embedding* seperti *Sentence-BERT* atau SimCSE untuk menghasilkan representasi dokumen dalam bentuk vektor semantik. Vektor-vektor ini kemudian dipakai untuk menghitung tingkat kemiripan antar dokumen dan membentuk sebuah *semantic graph*, yaitu jaringan yang berisi *node* sebagai dokumen dan *edge* sebagai hubungan kemiripan.

Peringkasan teks otomatis adalah sebuah sistem yang digunakan untuk membuat ringkasan singkat dari sebuah teks, tetapi tetap mempertahankan informasi penting dan makna utamanya. Dengan cara ini, isi utama dari sebuah dokumen atau berita tetap bisa dipahami tanpa harus membaca seluruh teks yang panjang (Bahari & Dewi, 2024a). Peringkasan teks dibagi menjadi dua jenis, yaitu peringkasan dokumen tunggal yang hanya meringkas satu artikel, dan peringkasan multi-dokumen yang merangkum beberapa artikel dengan topik yang sama sehingga hasilnya lebih lengkap (Yulianti et al., 2020). Selain itu, metode peringkasan juga terbagi menjadi dua, yaitu abstraktif dan ekstraktif. Metode ekstraktif lebih sering digunakan karena lebih sederhana, yaitu dengan memilih kalimat-kalimat yang paling penting dari teks asli berdasarkan skor relevansi. Hasil ringkasannya memang tidak terlalu memperhatikan tata bahasa, tetapi tetap bisa mewakili inti informasi dari dokumen (Naufal Ammar, n.d.).

Portal berita CNN Indonesia dan detikcom dipilih sebagai sumber data penelitian karena keduanya merupakan media online nasional yang aktif mempublikasikan berita secara real-time, memiliki jumlah artikel yang besar, serta

sering digunakan dalam penelitian analisis berita, text mining, dan analisis wacana di Indonesia menurut (Pendiidkan & Konsleing, 2023).

Dari penelitian sebelumnya, *extractive summarization* multi-dokumen dapat dilakukan dengan beberapa pendekatan. Metode statistik memilih kalimat berdasarkan frekuensi kata, sedangkan metode graf seperti *TextRank* (Andrew et al., n.d.) melihat hubungan antar kalimat. Pendekatan *centroid* mencari kalimat yang paling dekat dengan topik utama dokumen. Selain itu, ada *machine learning* yang mengklasifikasikan kalimat penting dengan algoritma tertentu, dan *deep learning* yang memanfaatkan representasi kata atau kalimat seperti Word2Vec atau BERT untuk memilih kalimat dengan skor tertinggi. Semua pendekatan tersebut memiliki tujuan yang sama, yaitu memilih kalimat paling penting untuk disusun menjadi ringkasan.

Penelitian sebelumnya tahun 2023 (Arya Ranggianto et al., 2023) juga melakukan kombinasi fitur untuk peringkasan multi-dokumen pada berita Amazon, menggunakan BERT untuk vektorisasi dan kombinasi fitur *centroid* serta *keyword*. Pada penelitian ini, *Sentence-BERT* digunakan untuk menyeleksi dan mengelompokkan berita dengan topik yang sama dari berbagai portal berita. *Sentence-BERT* bekerja dengan mengubah teks berita menjadi *embedding* vektor numerik, kemudian tingkat kemiripan antar berita dihitung menggunakan *Cosine Similarity*. Nilai kemiripan tersebut digunakan untuk membentuk graf kedekatan antar berita dan dilakukan proses *graph clustering* untuk mengelompokkan berita-berita yang membahas topik serupa menjadi satu klaster (*story cluster*)(L. Zhang et al., 2023). Dengan demikian, proses peringkasan hanya dilakukan pada kumpulan berita yang benar-benar memiliki konteks yang sama.

Pada tahap peringkasan, penelitian ini tetap menggunakan kombinasi tiga fitur untuk meningkatkan kualitas ringkasan. Pendekatan *centroid* digunakan untuk menemukan kalimat yang paling dekat dengan topik utama klaster berdasarkan representasi *embedding* kalimat dari *Sentence-BERT*. Fitur posisi digunakan berdasarkan konsep bahwa kalimat pertama dalam dokumen biasanya mengandung informasi paling penting, sehingga kalimat awal diberi nilai lebih tinggi dan nilai tersebut menurun seiring jauhnya posisi kalimat dari awal teks. Fitur *keyword* diambil menggunakan metode YAKE untuk menangkap kata-kata penting yang

mewakili inti informasi dokumen. Penelitian sebelumnya (Arya Ranggiyanto et al., 2023) menunjukkan bahwa kombinasi beberapa fitur mampu mengurangi pengulangan kalimat dan meningkatkan kualitas ringkasan.

Kontribusi penelitian ini adalah penggunaan *Sentence-BERT* untuk mengelompokkan berita berdasarkan kesamaan topik sebelum diringkas, serta penerapan kombinasi empat fitur yaitu, *centroid* berbasis *embedding* SBERT, fitur posisi kalimat, fitur *novelty*/keterbaruan informasi, dan fitur *keyword* berbasis YAKE dalam proses peringkasan multi-dokumen berita berbahasa Indonesia. Pendekatan ini diharapkan dapat menghasilkan ringkasan yang lebih akurat, informatif, dan bebas redundansi dibandingkan metode yang hanya memakai satu atau dua fitur. Penelitian ini diharapkan mampu menghasilkan sistem peringkasan otomatis yang efisien dan akurat, dengan kemampuan mengenali berita yang membahas topik yang sama dan merangkumnya menjadi informasi singkat, padat, dan mudah dipahami. Dengan adanya sistem ini, pembaca dapat memperoleh informasi penting dengan lebih cepat meskipun di tengah banyaknya berita *online* yang terus bertambah setiap hari.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka permasalahan yang dapat dirumuskan dalam penelitian ini adalah sebagai berikut:

1.2.1 Permasalahan

Dengan semakin banyaknya berita *online* yang terbit setiap hari, sering kali satu peristiwa dibahas oleh banyak artikel yang berbeda. Setiap berita menyajikan sudut pandang dan detail yang tidak selalu sama, sehingga pembaca sering menemukan informasi yang berulang. Untuk memahami satu topik secara lengkap, pembaca harus membuka banyak artikel, yang tentu memakan waktu dan membuat proses pencarian informasi menjadi tidak efisien. Sehingga, dibutuhkan sistem otomatis yang mampu mendeteksi berita dengan topik yang sama dan meringkasnya menjadi informasi singkat dan tidak berulang. Hal ini menjadi tantangan karena dibutuhkan metode yang mampu mengenali kemiripan antar berita serta memilih kalimat-kalimat yang paling penting dari kumpulan dokumen tersebut.

1.2.2 Solusi Permasalahan

Penelitian ini menawarkan solusi dengan memanfaatkan *Sentence-BERT* untuk mengubah setiap berita menjadi vektor numerik, kemudian *Cosine Similarity* digunakan untuk menghitung tingkat kemiripan antar berita. Hasil kemiripan ini digunakan untuk mengelompokkan berita yang membahas topik serupa sebelum dilakukan proses peringkasan. Setelah berita yang satu topik terkumpul, proses *extractive summarization* dilakukan menggunakan kombinasi fitur yang pertama relevansi menggunakan *centroid* untuk menentukan kalimat yang paling mendekati inti topik, posisi kalimat yang memberikan bobot lebih tinggi pada kalimat awal dalam dokumen, *novelty* atau keterbaruan informasi yang digunakan apakah berita memuat redundansi, dan *keyword* yang diambil menggunakan metode YAKE untuk menangkap kata-kata penting dalam setiap berita. Dengan menggabungkan teknik pengelompokan berbasis SBERT dan *Cosine Similarity* serta kombinasi fitur tersebut, sistem diharapkan mampu menghasilkan ringkasan multi-dokumen yang lebih akurat, informatif, dan tidak redundan.

1.3 Pertanyaan Penelitian

Berdasarkan latar belakang diatas, maka penelitian ini akan menyelesaikan rumusan masalah dibawah ini:

1. Berapa hasil evaluasi ROUGE dan BERTScore pada peringkasan multi-dokumen berita menggunakan kombinasi fitur yang terdiri dari relevansi menggunakan centroid, novelty, posisi, dan *keyword* berbasis YAKE?

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. *Dataset* berita hanya diambil dari dua portal, yaitu CNN Indonesia pada kategori nasional dan Detik pada kategori berita, melalui proses *web scraping*
2. Pengelompokan topik berita hanya menggunakan *Sentence-BERT* dan *Cosine Similarity*, serta metode *graph clustering* (NetworkX).
3. Metode peringkasan yang digunakan adalah *extractive summarization*.

4. Fitur yang digunakan untuk proses ringkasan terbatas pada 4 fitur, yaitu *centroid*, *novelty*, posisi kalimat, dan *keyword* (YAKE).
5. Hasil ringkasan hanya dalam bentuk teks, tanpa visualisasi atau model bahasa tambahan.
6. Penelitian tidak membahas performa pada bahasa selain bahasa Indonesia.
7. Evaluasi hasil ringkasan dilakukan menggunakan metrik ROUGE dan BERTScore sebagai ukuran kualitas ringkasan.

1.5 Tujuan dan Manfaat

1.5.1 Tujuan

Adapun tujuan penelitian ini dapat dilihat dibawah ini:

1. Mengembangkan sistem yang dapat mendeteksi dan mengelompokkan berita dengan topik yang sama menggunakan *Sentence-BERT*, *Cosine Similarity*, dan *graph clustering*.
2. Menerapkan kombinasi fitur relevansi menggunakan *centroid*, *novelty*, posisi kalimat, dan *keyword* berbasis YAKE untuk menghasilkan ringkasan multi-dokumen yang singkat, akurat, dan tidak berulang.

1.5.2 Manfaat

Penelitian ini diharapkan memberikan manfaat bagi pengguna, terutama pembaca berita *online*. Sistem yang dikembangkan dapat membantu pengguna memperoleh inti informasi dari banyak berita secara cepat tanpa harus membuka dan membaca seluruh artikel satu per satu. Dengan adanya proses pengelompokan berita berdasarkan topiknya, pengguna bisa langsung melihat kumpulan berita yang membahas peristiwa yang sama. Selain itu, ringkasan yang dihasilkan dari kombinasi fitur *centroid*, posisi kalimat, dan *keyword* berbasis YAKE membuat informasi yang ditampilkan menjadi lebih jelas, padat, dan tidak berulang. Hal ini memudahkan pengguna dalam memahami inti peristiwa hanya dalam waktu singkat, sehingga proses pencarian informasi menjadi lebih efisien dan praktis.

1.6 Sistematika Penulisan Skripsi

Penulisan skripsi ini mengacu pada sistematika kepenulisan sebagai berikut:

BAB I PENDAHULUAN

Bab satu berisi latar belakang dilakukannya penelitian tersebut, kemudian dari latar belakang tersebut menghasilkan rumusan masalah, metode usulan, pertanyaan penelitian, tujuan, manfaat, batasan masalah dan sistematika penulisan skripsi.

BAB II KAJIAN PUSTAKA

Bab kedua berisi tinjauan literatur yang digunakan sebagai landasan teori untuk permasalahan dan metode yang akan digunakan dalam penelitian. Tinjauan ini didasarkan pada sumber-sumber yang valid, serta penjelasan terkait dengan penelitian sebelumnya yang sebanding dengan penelitian yang sedang dilakukan.

BAB III METODOLOGI PENELITIAN

Bab tiga berisikan gambaran umum studi kasus, alur proses penelitian, Pengukuran data, penggunaan metode/algoritmanya, proses pengumpulan data dan perhitungannya.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai lingkungan uji coba, tahapan pembuatan program, dan hasil uji coba.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran berdasarkan penelitian yang telah dilakukan.