BAB I
PENDAHULUAN

Bab ini menjelaskan kerangka dasar dari penelitian yang terdiri dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, dan sistematika penulisan. Seluruh bagian tersebut disusun untuk memberikan gambaran umum mengenai arah, fokus, serta ruang lingkup penelitian secara menyeluruh.

1.1. Latar Belakang

Era digital telah mengubah cara masyarakat mengonsumsi informasi. Pertumbuhan pengguna internet di Indonesia mencapai 221,5 juta jiwa pada tahun 2024, dengan penetrasi sebesar 79,5% dari total populasi (Prasetiyo dkk., 2024). Kondisi ini memunculkan fenomena *information overload*, di mana pengguna dibanjiri volume berita yang sangat besar dengan konten yang seringkali redundan pada topik yang sama (Guo dkk., 2024). Masalah ini diperparah oleh praktik jurnalisme *online* di Indonesia yang saat ini mengutamakan kecepatan publikasi dibandingkan akurasi, sehingga satu peristiwa dapat diliput oleh puluhan portal berita dengan sudut pandang yang hampir identik dan repetitif (Mahendra & Sukartik, 2024).

Redundansi pada berita digital di Indonesia semakin parah karena banyak media saling mengutip satu sama lain dari sumber yang sama. Kondisi ini tidak lepas dari besarnya jumlah media online di Indonesia. Terdata oleh Dewan Pers mencatat ada 964 media online yang sudah terverifikasi hingga tahun 2023, sementara secara keseluruhan jumlah outlet media di Indonesia diperkirakan mencapai lebih dari 45.000 (Antara News, 2024; Kompas.id, 2023). Dengan persaingan yang sangat ketat, banyak media akhirnya memilih meliput topik yang sama berulang kali dengan sedikit perbedaan, sehingga informasi yang beredar menjadi sangat repetitif. Pengulangan seperti ini pada akhirnya membebani pembaca dan memperparah fenomena *information overload* (Peng dkk., 2021). Fakta ini juga didukung oleh *Reuters Institute Digital News Report* 2024, yang menemukan bahwa 39% pembaca di berbagai negara mengaku kelelahan akibat terlalu banyaknya berita—angka ini naik 11 poin persentase dibanding lima tahun sebelumnya (Newman dkk., 2024). Di Indonesia sendiri kondisinya bahkan lebih parah; laporan *Digital News Report* 2025 mencatat bahwa 75% konsumen berita Indonesia mengaku sering atau sesekali menghindari berita, lebih tinggi

dari rata-rata global yang sebesar 71% (Steele, 2025). Permasalahan diperparah dengan maraknya praktik *clickbait*, di mana media *online* lebih memprioritaskan judul sensasional demi mendongkrak kuantitas klik daripada menyajikan informasi substansial. Akibatnya, pembaca terpaksa menelusuri keseluruhan isi artikel hanya untuk memverifikasi konteks peristiwa yang sebenarnya (Trustyanda dkk., 2021). Tingginya volume teks yang redundan dan terfragmentasi di ruang digital ini berujung pada inefisiensi waktu serta tenaga pembaca dalam mengekstraksi inti informasi yang mereka butuhkan (Widyassari dkk., 2022).

Peringkasan teks otomatis (*automatic text summarization*) muncul sebagai solusi untuk mengatasi *information overload* (Widyassari dkk., 2022). Terdapat dua pendekatan utama yaitu berupa ekstraktif dan abstraktif. Metode abstraktif rentan terhadap anomali halusinasi, yakni menghasilkan informasi yang gramatikal namun faktanya keliru atau tidak pernah ada di dalam dokumen asli (Ji dkk., 2023; Maynez dkk., 2020). Mengingat berita jurnalistik menuntut akurasi faktual yang tinggi, pendekatan ekstraktif menjadi pilihan yang lebih rasional karena menjamin setiap kalimat berasal langsung dari sumber asli. Metode pendekatan ekstraktif dalam peringkasan masih sangat efektif untuk korpus berita berbahasa Indonesia (Koto dkk., 2020).

Arsitektur peringkasan teks secara umum terbagi menjadi dua kategori berdasarkan jumlah sumber dokumen masukan, yaitu peringkasan dokumen tunggal (*single-document summarization*) dan peringkasan multi-dokumen (*multi-document summarization*). Peringkasan dokumen tunggal menitikberatkan pada proses ekstraksi informasi dari satu sumber teks saja. Peringkasan multi-dokumen memiliki tingkat kompleksitas komputasi yang jauh lebih tinggi karena algoritma harus memproses serta mensintesis informasi dari berbagai dokumen berbeda yang membahas topik serupa (Mawaridi dkk., 2024). Tantangan utama dalam peringkasan multi-dokumen mencakup tingginya tingkat redundansi kalimat, variasi gaya penulisan, serta risiko inkonsistensi fakta antar-artikel berita (Koto dkk., 2020; Widyassari dkk., 2022). Mesin peringkas memerlukan mekanisme pemilihan informasi yang cerdas guna menyaring bagian paling representatif dari seluruh sumber dokumen tersebut.

Proses peringkasan multi-dokumen memerlukan kepastian penunggalan tema dari seluruh artikel masukan sebelum pemrosesan semantik di tingkat kalimat dieksekusi. Aliran berita digital yang bersifat heterogen dan acak menuntut adanya penyaringan di tingkat dokumen (*document-level*) agar fakta dari berbagai sub-topik tidak bercampur secara rancu.

Ketiadaan pengelompokan awal pada pemrosesan multi-dokumen akan memicu penurunan drastis pada koherensi ringkasan, karena mesin berisiko menyatukan konteks dari sub-peristiwa yang sebenarnya berbeda (Lennox dkk., 2023). Oleh sebab itu, tahapan pengelompokan dokumen (*document clustering*) menjadi instrumen krusial untuk mengisolasi teks secara spesifik sebelum proses ekstraksi dimulai. Mengingat jumlah peristiwa atau topik berita yang muncul di internet tidak dapat diprediksi secara pasti, metode *Agglomerative Clustering* dipilih sebagai solusi utama. Metode ini telah terbukti sangat tangguh dalam memilah dan mengelompokkan dokumen-dokumen berita secara hierarkis tanpa mengharuskan sistem untuk menentukan jumlah klaster mutlak K secara subjektif di awal pemrosesan (Wicaksana dkk., 2018; Hanley dkk., 2025).

Berbagai penelitian terdahulu telah mengeksplorasi metode ekstraksi dan penilaian relevansi kalimat secara komputasional untuk korpus berita bahasa Indonesia. Pendekatan awal umumnya menggunakan metode statistik konvensional seperti TF-IDF, namun metode ini memiliki keterbatasan fundamental karena hanya mempertimbangkan frekuensi kemunculan kata tanpa memahami konteks semantik. Akibatnya, dua kalimat bermakna sama yang menggunakan kosakata berbeda dinilai tidak memiliki kemiripan, sehingga memicu tingginya redundansi pada hasil ringkasan (Gunawan dkk., 2019). Berupaya memperbaiki kelemahan tersebut, fokus penelitian selanjutnya berkembang pada dua arah solusi utama berupa perbaikan pemetaan topik dan penanganan redundansi teks. Untuk memperbaiki kerangka topik, diterapkan teknik pengelompokan dokumen seperti *Latent Dirichlet Allocation* (LDA) dan *Significance Sentence* (Widjanarko dkk., 2018), hingga kombinasi *K-Means* dan LDA yang mampu mencatatkan skor evaluasi ROUGE-1 sebesar 0,6199 (Twinandilla dkk., 2018). Sementara itu, untuk menanggulangi masalah tingginya redundansi secara spesifik, algoritma seleksi *Maximal Marginal Relevance* (MMR) banyak dieksplorasi. Integrasi algoritma MMR ini terbukti secara empiris sangat andal dalam menyeimbangkan relevansi dan diversitas kalimat, serta secara signifikan mampu mereduksi duplikasi informasi lintas dokumen dibandingkan metode yang murni berbasis probabilitas (Gunawan dkk., 2019; Mawaridi dkk., 2024).

Kinerja model pengelompokan dan seleksi tersebut masih bertumpu pada probabilitas statistik leksikal (pencocokan kata di permukaan teks) meskipun telah menunjukkan beberapa peningkatan tren akurasi. Mengatasi keterbatasan metode yang tidak mengenali makna semantik tersebut, perkembangan model *transformer* BERT hadir merevolusi

pemrosesan bahasa natural melalui representasi kontekstual yang mendalam (Devlin dkk., 2019). Berdasarkan bukti empiris dari literatur terdahulu, representasi berbasis BERT secara konsisten mengungguli metode statistik konvensional dalam menangkap struktur bahasa Indonesia yang kompleks (Subakti dkk., 2022). Mengingat arsitektur dasar BERT tidak efisien untuk komputasi perbandingan antar-kalimat, dikembangkanlah *Sentence*-BERT (SBERT) yang menggunakan jaringan kembar (*Siamese Network*). SBERT terbukti mampu menghasilkan *sentence embeddings* secara seragam, sehingga kalkulasi kemiripan makna kalimat untuk mendeteksi parafrasa dapat dihitung secara cepat menggunakan *Cosine Similarity* (Reimers & Gurevych, 2019). Keputusan untuk menjadikan SBERT sebagai solusi representasi semantik didukung kuat oleh penelitian terbaru yang memvalidasi bahwa integrasi *embeddings* SBERT pada kerangka peringkasan teks ekstraktif terbukti sangat tangguh dalam menghasilkan ringkasan yang koheren secara naratif dan optimal dalam mengontrol redundansi informasi (Annisa dkk., 2026).

Mekanisme evaluasi kuantitatif memiliki peran krusial dalam mengukur kualitas hasil peringkasan secara objektif guna mendampingi pengembangan arsitektur komputasi. Kajian literatur terbaru secara konsisten menggunakan metrik ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) sebagai standar validasi utama. Skenario pengujian ini mencakup ROUGE-1 untuk mengukur perolehan informasi dasar tingkat kata (*unigram*), ROUGE-2 untuk kelancaran struktur frasa (*bigram*), dan ROUGE-L untuk mengevaluasi koherensi struktur kalimat secara utuh (*longest common subsequence*) (Hartawan dkk., 2024). Di samping pemilihan metrik, penentuan rasio kompresi turut menjadi variabel penentu dalam evaluasi model. Mesin peringkas perlu diuji melalui variasi persentase pemotongan untuk mengukur stabilitas kinerjanya. Penelitian terkait peringkasan ekstraktif berita berbahasa Indonesia membuktikan bahwa pengujian pada rentang kompresi 20% hingga 50% dari dokumen asli merupakan ambang batas yang paling komprehensif guna menemukan titik keseimbangan optimal antara retensi fakta esensial dengan tingkat keringkasan teks (Girsang & Amadeus, 2023).

Penelitian ini memposisikan metode usulan untuk mengisi celah riset tersebut menggunakan korpus berita Liputan6 melalui pengembangan alur komputasi yang komprehensif. Tahapan awal dilakukan dengan mengintegrasikan algoritma *Agglomerative Clustering* untuk memilah dan mengelompokkan dokumen berita heterogen ke dalam klaster topik peristiwa yang seragam (Koto dkk., 2020). Langkah berikutnya mengeksekusi

pemahaman semantik menggunakan *Sentence*-BERT (SBERT) dan menerapkan algoritma *Maximal Marginal Relevance* (MMR) sebagai mekanisme seleksi akhir. Peran dari algoritma ini secara matematis dirancang untuk mengekstraksi inti informasi yang paling relevan sekaligus mengeliminasi kemunculan kalimat redundan lintas dokumen melalui kalibrasi parameter penyeimbang λ (*lambda*) (Mawaridi dkk., 2024). Kualitas ringkasan yang dihasilkan kemudian dievaluasi secara komprehensif menggunakan metrik ROUGE pada variasi kompresi tersebut terhadap *human ground truth* sebagai standar rujukan. Untuk menjamin objektivitas standar rujukan ini, uji reliabilitas antar-anotator menggunakan *Cohen's Kappa* diimplementasikan guna mengukur tingkat kesepakatan kognitif peringkas manusia, sehingga data standar emas yang dihasilkan bersifat konsisten dan terbebas dari bias subjektivitas (Badshah & Sajjad, 2025; Röttger dkk., 2022). Secara praktis, metode usulan ini diharapkan bermanfaat bagi jurnalis, peneliti, serta masyarakat umum yang membutuhkan intisari informasi, sekaligus menjadi tolok ukur (*benchmark*) bagi pengembangan teknologi pemrosesan bahasa alami (NLP) di masa depan.

Urgensi penelitian ini secara keseluruhan dilatarbelakangi oleh tingginya kebutuhan akan solusi komputasional yang mampu menyajikan intisari fakta berita secara komprehensif, padat, dan minim redundansi. Sinergi antara pengelompokan dokumen (*Agglomerative Clustering*), ketajaman pemahaman semantik (*Sentence-BERT*), serta ketepatan seleksi kalimat (MMR) diproyeksikan mampu menjadi arsitektur cerdas yang tangguh dalam memitigasi fenomena *information overload*. Melalui rancangan evaluasi kuantitatif yang tervalidasi, penelitian ini dilaksanakan untuk membuktikan sejauh mana arsitektur NLP tersebut dapat memberikan efisiensi kognitif yang optimal bagi pembaca berita digital di Indonesia.

1.2. Rumusan Masalah

Penelitian ini merumuskan beberapa pokok permasalahan sebagai tindak lanjut dari uraian latar belakang tersebut:

1. Bagaimana merancang dan mengintegrasikan metode usulan yang terdiri dari *Agglomerative Clustering*, representasi semantik *Sentence-BERT*, serta algoritma seleksi *Maximal Marginal Relevance* (MMR) untuk melakukan tugas peringkasan multi-dokumen pada teks berita berbahasa Indonesia?

2. Bagaimana tingkat performa metode usulan tersebut saat dievaluasi menggunakan metrik ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) terhadap *human ground truth* sebagai standar rujukan?

1.3. Tujuan

Tujuan dari penelitian ini adalah untuk membangun dan menguji kinerja mesin peringkasan multi-dokumen ekstraktif pada teks berita berbahasa Indonesia dengan mengintegrasikan algoritma *Agglomerative Clustering* sebagai mekanisme pengelompokan topik, *Sentence*-BERT untuk menghasilkan representasi semantik kalimat dalam bentuk *sentence embeddings*, serta algoritma *Maximal Marginal Relevance* (MMR). Selain itu, penelitian ini juga bertujuan untuk menganalisis pengaruh variasi tingkat kompresi dalam menentukan titik keseimbangan informasi yang optimal, serta mengevaluasi efektivitas algoritma tersebut secara komprehensif menggunakan metrik ROUGE dengan membandingkan luaran mesin terhadap *human ground truth* sebagai standar rujukan hasil ringkasan manusia.

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat yang langsung dirasakan oleh masyarakat luas, jurnalis, maupun peneliti di tengah tingginya arus informasi digital. Pengguna mendapatkan akses instan terhadap inti sebuah peristiwa secara utuh, tanpa harus menghabiskan waktu menyaring puluhan artikel yang repetitif. Pendekatan semantik ini memberikan terobosan dalam kajian *Natural Language Processing* (NLP) bahasa Indonesia dengan membuktikan bahwa model mampu mengeliminasi redundansi teks secara cerdas tanpa kehilangan konteks. Selain itu, karena metode yang digunakan murni mengambil kalimat-kalimat asli dari sumber berita (tidak mengarang kalimat baru), ringkasan yang disajikan dijamin akurat dan sesuai fakta. Hal ini sangat membantu pengguna terhindar dari informasi yang menyesatkan atau kalimat yang dikarang secara keliru oleh mesin pemroses.

1.5. Batasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan untuk memperjelas ruang lingkup dan interpretasi hasil, antara lain:

1. Sumber data teks murni dibatasi pada artikel berita berbahasa Indonesia yang diambil dari portal berita Liputan6. Pengujian difokuskan pada enam kategori spesifik, yaitu

Kesehatan, Otomotif, Bisnis, Saham, Bola, dan Tekno. (berita pada tanggal 12 Januari 2026 hingga 18 Januari 2026).

2. Pengukuran kualitas hasil ringkasan dibatasi pada evaluasi komputasional secara kuantitatif menggunakan metrik ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) yang berfokus pada seberapa besar irisan informasi antara keluaran mesin dengan *human ground truth*. Penelitian ini tidak mencakup pengujian kualitatif terkait koherensi paragraf, tata bahasa, maupun tingkat keterbacaan (*readability*) oleh pakar bahasa atau panelis manusia.

1.6. Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran yang terstruktur mengenai isi penelitian. Skripsi terdiri dari lima bab utama, yang masing-masing membahas aspek berbeda dari penelitian ini.

BAB I PENDAHULUAN

Bab ini memberikan gambaran mengenai latar belakang permasalahan, rumusan masalah, tujuan, manfaat, dan batasan penelitian. Sistematika penulisan skripsi terkait pengembangan mesin peringkas multi-dokumen ekstraktif menggunakan metode *Sentence*-BERT dan *Maximal Marginal Relevance* (MMR)

BAB II TINJAUAN PUSTAKA

Bab ini memberikan kajian pustakan yang berhubungan dengan tema skripsi sebagai landasan untuk perumusan dan analisis permasalahan pada skripsi. Kajian pustaka yang digunakan meliputi, metode *Agglomerative Clustering*, teori pra-pemrosesan teks, uji reliabilitas *Cohen's Kappa, Sentence*-BERT, MMR, serta metrik evaluasi ROUGE.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan sistematis dalam pelaksanaan penelitian yang dimulai dari proses pengumpulan data mentah. Rangkaian pemrosesan selanjutnya meliputi pengelompokan topik berita, penyusunan teks rujukan (ground truth), pengujian *Cohen's Kappa*, pra-pemrosesan teks, ekstraksi representasi semantik SBERT, seleksi kalimat menggunakan algoritma MMR, hingga evaluasi kinerja ringkasan dengan metrik ROUGE.

BAB IV HASIL DAN ANALISA

Bab ini menguraikan lingkungan komputasi, hasil eksekusi skenario eksperimen, dan interpretasi kinerja mesin peringkas. Uraian pembahasan mencakup hasil validasi *Cohen's Kappa*, studi kasus peringkasan dokumen, analisis evaluasi skor ROUGE pada enam kategori berita yang berbeda, peninjauan kelemahan metode usulan, serta demonstrasi implementasi prototipe antarmuka program.

BAB V PENUTUP

Bab ini menjabarkan kesimpulan akhir dari seluruh rangkaian pengujian dan analisis yang telah dilakukan pada bab-bab sebelumnya. Saran pengembangan untuk penelitian lebih lanjut di masa mendatang juga dirumuskan pada bagian ini guna menyempurnakan kualitas peringkasan metode usulan.