

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di Indonesia, pola konsumsi berita digital menunjukkan kemiripan dengan negara lain, namun terdapat tantangan spesifik yang dihadapi. Data Kementerian Komunikasi dan Informatika (Kominfo) mencatat peningkatan aktivitas pencarian informasi daring sebesar 8% dibandingkan tahun 2023 (Budiarto et al., 2024). Tren ini berlanjut pada periode 2024 hingga 2025, sehingga dari peningkatan skor Indeks Masyarakat Digital Indonesia (IMDI) secara nasional serta bertambahnya jumlah pengguna internet yang mencapai 229 juta jiwa atau 80,66% populasi. Peningkatan tersebut menunjukkan bahwa masyarakat semakin aktif memanfaatkan platform digital sebagai sumber utama informasi, termasuk berita daring (Dr. Ir. Bonifasius Wahyu Pudjianto M.T. et al., 2025).

Namun, peningkatan akses informasi ini tidak diimbangi dengan literasi digital yang memadai, sebagaimana tercermin dari skor indeks literasi digital di setiap provinsi yang belum mencapai kategori "baik". Akibatnya, masyarakat mengalami kesulitan dalam menganalisis informasi secara kritis (Biswas et al., 2025). Setiap hari, ribuan artikel berita daring berbahasa Indonesia dipublikasikan secara daring, menghasilkan volume informasi yang sangat besar dan menuntut pengelolaan yang efisien. Seiring meningkatnya volume berita daring, beberapa penelitian menunjukkan bahwa efektivitas sistem klasifikasi berita dapat ditingkatkan apabila dokumen terlebih dahulu diringkas sebelum diproses. Penelitian yang dilakukan (Saadi et al., 2025) membuktikan bahwa RoBERTa yang dilatih menggunakan dokumen teringkas dapat mencapai akurasi tinggi sekaligus mengurangi biaya komputasi secara signifikan, menunjukkan manfaat summarization sebagai praproses sebelum klasifikasi.

Pendekatan SUMMaug juga menunjukkan bahwa ringkasan dapat meningkatkan robustness dan akurasi klasifikasi karena menghasilkan representasi data yang lebih padat dan lebih mudah dipelajari oleh model transformer (Wang & Yoshinaga, 2023). Ringkasan terbukti mampu mengurangi redundansi, menonjolkan informasi inti, dan memperjelas struktur semantik dokumen, sehingga

model deep learning seperti RoBERTa dapat bekerja lebih efisien. Temuan-temuan ini mengindikasikan bahwa perangkuman teks tidak hanya sebatas kompresi, melainkan strategi yang efektif untuk memperkuat pipeline klasifikasi berita.

Klasifikasi teks merupakan salah satu aplikasi utama dalam Natural Language Processing (NLP) dengan berbagai kegunaan praktis, seperti personalisasi konten, sistem rekomendasi, dan analisis sentimen (Kowsari et al., 2019). Sistem klasifikasi otomatis membantu mengelompokkan berita ke dalam kategori tertentu, sehingga pembaca lebih mudah menemukan berita yang relevan (Gormantara & Watomakin, 2020). Berbagai teknik pembelajaran mesin tradisional, termasuk Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Random Forest, telah diterapkan dengan hasil yang bervariasi. Pada penelitian (Sudianto et al., 2022) mengklasifikasikan topik berita menggunakan SVM dan Multi-Layer Perceptron (MLP) berbasis TF-IDF, dengan akurasi SVM sebesar 76% dan MLP 79%.

Perkembangan deep learning telah menghasilkan model berbasis transformer seperti BERT dan RoBERTa yang kini mendominasi berbagai tugas NLP (Devlin et al., 2019). RoBERTa, sebagai pengembangan dari BERT, RoBERTa dilatih lebih lama dengan batch yang lebih besar, dan dataset yang jauh lebih besar dan lebih beragam, serta penyesuaian hiperparameter secara ekstensif melalui penghapusan tugas *Next Sentence Prediction* (NSP) dan penerapan dynamic masking (Liu et al., 2019). Dalam konteks volume artikel berita yang panjang dan variatif, perangkuman teks (*text summarization*) menjadi teknik penting untuk menyederhanakan dokumen tanpa menghilangkan informasi utama. Perangkuman teks dapat dilakukan melalui dua pendekatan utama ekstraktif dan abstraktif. Pada perangkuman ekstraktif, metode memilih kalimat paling penting dari teks asli.

Salah satu metode perangkuman ekstraktif yaitu TextRank, TextRank ialah metode berbasis *graph-based ranking*, salah satu algoritma ekstraktif yang paling banyak digunakan dan telah terbukti efektif dalam berbagai domain, salah satu keunggulan algoritma TextRank adalah bahwa algoritma ini tidak memerlukan pelatihan menggunakan data pelatihan dan telah diterapkan dalam berbagai studi untuk ringkasan teks dalam berbagai bahasa (Manojkumar et al., 2022) (Hidayattullah & Azizi, 2021). Penelitian oleh (Feizi Derakhshi & Abdulreda Kadhim, 2025) menggunakan pendekatan hybrid TextRank yang digabungkan

dengan optimasi seperti *Chi-Square Binary Cuckoo Search* berhasil mencapai ROUGE-L 0,818 pada dataset BBC dan CNN/DailyMail, menunjukkan fleksibilitas TextRank untuk ditingkatkan melalui optimasi lanjutan. Penelitian oleh (Fakhrezi et al., 2021) juga menunjukkan bahwa TextRank mampu menghasilkan ringkasan stabil dan konsisten dengan F-Score 0,6173 pada pengembangan ensiklopedia multi-sumber.

Penelitian yang dilakukan oleh (Sihombing et al., 2024) menunjukkan hasil ringkasan menggunakan TextRank TF-IDF dengan akurasi mencapai 88,533% pada kompresi teks 50%, 85,533% pada kompresi teks 40%, 81,26% pada kompresi teks 30% dalam pengembangan peringkasan untuk artikel ilmiah indonesia. Berbeda dengan perangkuman ekstraktif, perangkuman abstraktif menghasilkan ringkasan baru dengan menyusun ulang informasi menggunakan bahasa alami. Model transformer seperti BART (Lewis et al., 2019), T5 (Raffel et al., 2023) dan Pegasus (Zhang & Yu, n.d.) telah mencapai performa unggulan dalam peringkasan otomatis. Kombinasi metode perangkuman dan model transformer untuk klasifikasi menawarkan pipeline yang saling melengkapi, Pada penelitian ini lebih berfokus kepada pendekatan perangkuman ekstraktif menggunakan TextRank.

Pada pendekatan TextRank classifier telah terbukti efektif dalam berbagai konteks dokumentasi panjang. PageRank-based summarization seperti TextRank terbukti membantu meningkatkan performa model downstream dengan menyaring informasi esensial dan mengurangi noise. Pada penelitian (Medina et al., 2022) menunjukkan bahwa summarization ekstraktif berbasis PageRank mampu meningkatkan skor F1 ketika digunakan sebagai preprocessing sebelum klasifikasi dokumen hukum Portugis, menegaskan manfaat integrasi ringkasan dalam pipeline klasifikasi.

Selain itu, penelitian (Fakhrezi et al., 2021) memperlihatkan bahwa metode berbasis PageRank mampu mengekstraksi informasi inti secara konsisten melalui pembentukan similarity matrix dan ranking yang stabil, sehingga menghasilkan ringkasan yang mempertahankan esensi semantik dokumen panjang. Integrasi peringkasan dan klasifikasi juga didukung oleh penelitian hybrid summarization-classification yang menekankan bahwa ringkasan yang terfokus dan koheren menjadi fondasi kuat bagi tugas NLP skala besar. Studi pada integrasi feature-based

summarization dan classification menunjukkan bahwa perangkuman yang menonjolkan informasi penting (NER, posisi kalimat, cue phrases, dll.) dapat meningkatkan efektivitas model klasifikasi downstream dalam berbagai domain berita dan informasi publik (Feizi Derakhshi & Abdulreda Kadhim, 2025). Secara keseluruhan, bukti-bukti ini menguatkan bahwa ringkasan baik ekstraktif maupun abstraktif dapat berfungsi sebagai reduksi dimensi semantik yang memperbaiki efisiensi dan stabilitas model klasifikasi modern.

Pada penelitian yang telah dilakukan sebelumnya, umumnya menjadikan perangkuman sebagai *hasil akhir* dan berfokus pada kualitas ringkasan (misalnya melalui metrik ROUGE). Namun dalam penelitian ini, perangkuman digunakan terutama sebagai tahap reduksi dimensi dari teks dan penyederhanaan struktur dokumen diharapkan agar proses klasifikasi berita menjadi lebih efisien tanpa kehilangan inti informasi memungkinkan RoBERTa bekerja pada input yang lebih ringkas namun tetap informatif. Dengan demikian, penelitian ini bertujuan mengembangkan klasifikasi berita Indonesia yang mengintegrasikan TextRank sebagai tahap praproses utama dengan RoBERTa sebagai *classifier* sebuah pendekatan yang sejauh ini belum dievaluasi secara sistematis dalam konteks korpus berita berbahasa Indonesia.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Redundansi informasi pada teks dan kompleksitas struktur teks pada dokumen berita yang panjang, pengulangan informasi, dan penyertaan konteks tambahan berpotensi meningkatkan beban komputasi dan waktu komputasi pada proses pembelajaran, sehingga diperlukan pendekatan yang mampu mengelola kompleksitas teks tanpa mengorbankan kinerja klasifikasi.

1.2.2 Metode Usulan

Penelitian ini mengusulkan sebuah pipeline dua tahap yang mengintegrasikan metode peringkasan teks ekstraktif TextRank dengan klasifikasi berita berbasis Indonesian RoBERTa. Pada tahap awal, dokumen berita yang berukuran panjang diringkaskan menggunakan TextRank dengan tingkat kompresi yang ditentukan pada masing-masing skenario pengujian. Hasil

peringkasan tersebut kemudian digunakan sebagai masukan pada tahap klasifikasi berita. Pendekatan ini dirancang untuk mengurangi redundansi dan kompleksitas teks dengan menyajikan representasi dokumen yang lebih ringkas dan terfokus, sehingga memungkinkan pengujian pengaruh reduksi dimensi terhadap efisiensi komputasi serta kinerja klasifikasi berita berbasis Transformer.

1.2.3 Pertanyaan Penelitian

1. Bagaimana kualitas ringkasan yang dihasilkan atau TextRank dalam mempertahankan informasi yang penting untuk klasifikasi berita ?
2. Bagaimana pengaruh penerapan perangkuman teks ekstraktif (TextRank) terhadap performa akurasi dan efisiensi komputasi dalam klasifikasi berita menggunakan IndoRoBERTa ?

1.3 Tujuan & Manfaat

1.3.1 Tujuan

Berdasarkan rumusan masalah di atas, tujuan penelitian ini adalah:

1. Mengevaluasi dampak tingkat kompresi dan kualitas teks ringkasan yang dihasilkan model TextRank dalam mempertahankan informasi diskriminatif yang penting untuk klasifikasi berita Indonesia.
2. Menganalisis pengaruh perangkuman ekstraktif berbasis model TextRank terhadap akurasi klasifikasi berita menggunakan model Indonesian RoBERTa.

1.3.2 Manfaat

1. Memberikan kontribusi ilmiah terhadap pengembangan pendekatan hybrid perangkuman klasifikasi pada bahasa Indonesia, khususnya untuk pemrosesan dokumen panjang.
2. Menyediakan pemahaman empiris mengenai efektivitas perangkuman TextRank sebagai teknik reduksi dimensi semantik untuk meningkatkan efisiensi dan akurasi model transformer RoBERTa.

1.4 Batasan-batasan

Untuk memfokuskan ruang lingkup penelitian dan memastikan hasil yang dapat diukur dengan baik, penelitian ini memiliki beberapa batasan sebagai berikut :

1. Dataset berita berbahasa Indonesia yang dikumpulkan dari portal berita online Indonesia (CNN Indonesia) dengan jumlah 10.000, Kategori berita yang telah ditentukan (yakni: Politik, Ekonomi, Peristiwa, Olahraga, dan Health).
2. Arsitektur pipeline dua tahap secara sekuensial, dimana perangkuman ekstraktif dilakukan terlebih dahulu menggunakan model TextRank, kemudian hasil ringkasan digunakan sebagai input untuk model klasifikasi RoBERTa (IndoRoBERTa).
3. Model perangkuman ekstraktif yang digunakan adalah model TextRank.
4. Penelitian ini membatasi evaluasi pada tiga variasi persentase panjang ringkasan, yaitu 30% (0.3), 40% (0.4), dan 50% (0.5) dari panjang teks asli.
5. Kualitas perangkuman dievaluasi menggunakan metrik standar ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) dan evaluasi manual oleh expert untuk mengukur kelengkapan informasi, koherensi, dan relevansi ringkasan.
6. Menggunakan Quilbot (Summarize) sebagai pseudo-referensi (ringkasan referensi).
7. Performa klasifikasi dievaluasi menggunakan metrik standar seperti akurasi, presisi, *recall*, dan F1-score.

1.5 Sistematika Proposal

Sistematika penulisan laporan proposal skripsi sebagai berikut :

- **BAB I PENDAHULUAN**

Pada bab ini, menyajikan bagian pendahuluan yang berisi latar belakang masalah penelitian yaitu terkait informasi pengaruh akurasi dan efisiensi dalam menerapkan peringkasan dalam klasifikasi teks berita daring. Kemudian dilanjutkan dengan perumusan masalah yang menerapkan algoritma TextRank dalam peringkasan dan klasifikasi menggunakan Deep Learning untuk mengetahui hasilnya, serta tujuan dan manfaat yang diharapkan.

- **BAB II KAJIAN PUSTAKA**

Pada bab ini disajikan teoritis terkait *Dokumen berita indonesia dan tantangan pemrosesan teks panjang, text rank, preprocessing, model*

RoBERTa, model *IndoRoBERTa*, matrik evaluasi kinerja peringkasan dan klasifikasi, serta penelitian terkait dengan menerapkan tunneling (peringkasan lalu klasifikasi) untuk klasifikasi berita daring.

- **BAB III METODE USULAN**

Pada bab ini dibahas mengenai tahapan dari algoritma yang dipilih, arsitektur system yang diterapkan, dataset yang digunakan, scenario uji coba yang hendak diimplementasikan.

- **BAB IV HASIL DAN PEMBAHASAN**

Pada bab IV menyajikan hasil penelitian yang diperoleh dari proses implementasi dan pengujian yang telah dilakukan. Pada bab ini ditampilkan hasil eksperimen, performa sistem atau model berdasarkan metrik evaluasi yang digunakan, serta perbandingan dengan metode pembanding atau penelitian sebelumnya. Selain itu, bab ini juga memuat pembahasan dan analisis terhadap hasil yang diperoleh untuk menjelaskan keterkaitannya dengan tujuan dan rumusan masalah penelitian.

- **BAB V KESIMPULAN DAN SARAN**

Pada bab ini berisi kesimpulan akhir penelitian yang menjawab rumusan masalah dan tujuan penelitian berdasarkan hasil dan pembahasan pada bab sebelumnya. Selain itu, bab ini juga memuat saran untuk pengembangan penelitian selanjutnya, baik dari sisi metode, data, maupun sistem, agar penelitian dapat ditingkatkan atau dikembangkan lebih lanjut.