

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi digital yang pesat saat ini memicu fenomena ledakan informasi yang mengubah pola konsumsi berita di masyarakat Indonesia. Kemudahan akses terhadap berbagai platform media daring menyebabkan informasi tidak hanya mudah diakses, tetapi juga mudah diproduksi dan didistribusikan secara masif dalam waktu singkat (Pratiwi & Nugraha, 2022). Kondisi tersebut berdampak pada meningkatnya volume informasi yang beredar, termasuk informasi yang tidak valid sehingga menimbulkan tantangan bagi pengguna dalam memperoleh informasi yang sesuai dengan kebutuhan secara cepat. Permasalahan ini berkaitan dengan kesulitan dalam menyaring dan mengidentifikasi informasi yang relevan di tengah tingginya arus data.

Fokus utama dalam ekosistem berita saat ini tidak lagi terletak pada keterbatasan akses data, tetapi pada kompleksitas dalam menyaring informasi yang akurat. Tantangan ini semakin meningkat terutama pada konteks bahasa dengan keterbatasan sumber daya (*low-resource languages*), termasuk bahasa Indonesia yang meskipun digunakan oleh jutaan penutur masih menghadapi keterbatasan dalam ketersediaan dataset, model, dan *benchmark* yang memadai (Cahyawijaya dkk., 2021). Dalam konteks tersebut, sistem temu kembali informasi (*information retrieval*) memegang peranan penting sebagai solusi untuk menemukan dan menyajikan informasi yang relevan berdasarkan kebutuhan pengguna (Sharma & Panda, 2023). Hal ini juga didukung oleh analisis tren penelitian yang menunjukkan bahwa topik *information retrieval* masih menjadi salah satu area yang aktif dikembangkan dalam literatur ilmiah yang ditunjukkan melalui kepadatan istilah terkait dalam berbagai penelitian terkini (Hambarde & Proenca, 2023).

Evolusi *information retrieval* terus berkembang seiring dengan upaya peningkatan relevansi serta kualitas hasil pencarian. Salah satu metode konvensional yang masih menjadi standar utama adalah *Best Matching 25* (BM25), yang bekerja dengan mengukur tingkat kecocokan istilah antara kueri dan dokumen. Dalam penelitian yang dilakukan oleh Lasminingrum dkk., (2026),

metode BM25 mampu mencapai performa yang tinggi dengan nilai *Normalized Discounted Cumulative Gain* (NDCG@5) sebesar 0.8866 dan *Precision@5* (P@5) sebesar 0.8067. Namun demikian, sebagai pendekatan leksikal, BM25 masih memiliki keterbatasan dalam memahami hubungan makna antar kata yang berbeda secara terminologi, sehingga belum mampu menangkap kesamaan semantik dan konteks secara mendalam, terutama ketika istilah yang digunakan berbeda meskipun memiliki makna yang sama (Sharma & Panda, 2023).

Perkembangan bidang *Natural Language Processing* (NLP) membawa perubahan signifikan dalam meningkatkan pemahaman bahasa oleh mesin melalui pendekatan berbasis *neural*. Salah satu kemajuan penting ditandai dengan diperkenalkannya arsitektur *Transformer* yang mampu meningkatkan performa berbagai tugas bahasa alami secara signifikan (Vaswani dkk., 2017). Inovasi ini kemudian melahirkan model *Bidirectional Encoder Representations from Transformers* (BERT) yang mampu memahami konteks secara dua arah dan mencapai kinerja tinggi pada berbagai *benchmark* NLP (Devlin dkk., 2019). Pengembangan lebih lanjut melalui *Sentence-BERT* (SBERT) memungkinkan representasi kalimat yang lebih efisien, di mana proses pencarian kemiripan yang sebelumnya membutuhkan waktu hingga 65 jam dengan BERT dapat diselesaikan hanya dalam 5 detik (Reimers & Gurevych, 2019).

Pengembangan model berbasis semantik terus mengalami adaptasi melalui penerapan pada bahasa tertentu, termasuk pada bahasa Indonesia melalui model *Indo-Sentence-BERT* (IndoSBERT). Model ini dirancang untuk merepresentasikan makna teks secara kontekstual sehingga mampu meningkatkan kualitas pencarian informasi berbasis kesamaan semantik (Ubaidillah dkk., 2025). Meskipun demikian, pendekatan *semantic retrieval* tidak selalu menunjukkan performa terbaik pada semua kondisi pencarian. Model berbasis *embedding* cenderung mengalami kesulitan dalam menangkap istilah spesifik atau frasa penting yang bersifat eksplisit dalam kueri. Hal ini menyebabkan beberapa dokumen yang relevan secara leksikal tidak teridentifikasi dengan baik terutama pada kueri yang bersifat pendek dan berbasis kata kunci (Lasminingrum dkk., 2026).

Perbedaan karakteristik antara pendekatan leksikal dan semantik menghasilkan sinyal relevansi yang berbeda pada proses *information retrieval*

(Kodri dkk., 2025). Pendekatan semantik unggul dalam memahami konteks makna, namun tidak selalu mampu menangkap sinyal relevansi yang bersifat lemah atau implisit dalam dokumen. Sebaliknya, pendekatan leksikal seperti BM25 tetap efektif dalam mengidentifikasi kecocokan istilah secara eksplisit yang sering kali diabaikan oleh model berbasis *embedding*. Temuan ini diperkuat oleh penelitian Wang dkk., (2021) yang menunjukkan bahwa integrasi antara *dense retrieval* dan BM25 melalui mekanisme interpolasi mampu meningkatkan efektivitas *information retrieval* secara signifikan. Dengan demikian, pengembangan sistem *hybrid retrieval* menjadi pendekatan yang potensial untuk menggabungkan keunggulan kedua metode tersebut. Salah satu pendekatan yang umum digunakan dalam *hybrid retrieval* adalah *Convex Combination*, yaitu penggabungan skor BM25 dan model berbasis BERT melalui parameter pembobotan tertentu. Pendekatan ini terbukti efektif dan stabil dalam menggabungkan skor leksikal dan semantik, di mana performanya sangat dipengaruhi oleh pengaturan bobot yang digunakan (Bruch dkk., 2024). Selain itu, metode ini diketahui sangat efisien dalam penggunaan sampel kueri untuk mencapai nilai optimal pada domain target.

Meskipun pendekatan *hybrid retrieval* yang menggabungkan metode leksikal dan semantik telah banyak diteliti dan terbukti meningkatkan performa *information retrieval*, penelitian yang ada umumnya memanfaatkan model berbasis BERT atau model *multilingual* (Hambarde & Proenca, 2023), sementara eksplorasi yang spesifik berfokus pada model bahasa Indonesia seperti IndoSBERT masih relatif terbatas. Selain itu, meskipun metode penggabungan seperti *Convex Combination* telah umum digunakan, analisis mendalam mengenai pengaruh variasi pembobotan terhadap performa sistem khususnya pada dataset berita berbahasa Indonesia juga masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis pengaruh variasi pembobotan *Convex Combination* dalam sistem *hybrid* BM25 dan IndoSBERT terhadap performa *information retrieval* pada dataset berita berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana efektivitas pemeringkatan (*ranking effectiveness*) metode BM25, IndoSBERT, dan pendekatan *hybrid* dalam *Information Retrieval* pada dataset berita berbahasa Indonesia yang diukur menggunakan metrik *Precision@5*, *Precision@10*, *NDCG@5* dan *NDCG@10*, serta bagaimana efisiensi waktu respons (*latency*) dari masing-masing metode tersebut?
2. Bagaimana pengaruh variasi pembobotan metode *convex combination* dalam menggabungkan skor BM25 dan IndoSBERT terhadap efektivitas pemeringkatan sistem berdasarkan metrik evaluasi yang sama?

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis efektivitas pemeringkatan (*ranking effectiveness*) metode BM25, IndoSBERT, dan pendekatan *hybrid* dalam *information retrieval* pada dataset berita berbahasa Indonesia yang diukur menggunakan metrik *Precision@5*, *Precision@10*, *NDCG@5* dan *NDCG@10*, serta mengevaluasi efisiensi waktu respons (*latency*) dari masing-masing metode tersebut.
2. Menganalisis pengaruh variasi pembobotan pada metode *Convex Combination* dalam proses penggabungan skor BM25 dan IndoSBERT terhadap efektivitas pemeringkatan sistem berdasarkan metrik evaluasi yang sama.

1.4 Manfaat

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut:

1.4.1 Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan kajian di bidang *Information Retrieval*, khususnya terkait analisis pengaruh variasi pembobotan *Convex Combination* pada pendekatan *hybrid retrieval* yang menggabungkan model leksikal (BM25) dan model semantik berbasis *transformer* (IndoSBERT). Selain itu, penelitian ini diharapkan dapat memperkaya literatur mengenai pemanfaatan kombinasi leksikal dan semantik dalam meningkatkan efektivitas pemeringkatan (*ranking effectiveness*) pada data berita berbahasa Indonesia.

1.4.2 Manfaat Praktis

Penelitian ini diharapkan dapat menjadi referensi langsung dalam pengembangan sistem pencarian portal berita yang lebih efektif dan akurat melalui penerapan metode *Convex Combination*. Selain itu, hasil penelitian ini dapat menjadi acuan bagi pengembang sistem atau peneliti selanjutnya dalam menentukan dan mengoptimalkan parameter pembobotan pada sistem *hybrid retrieval*, khususnya untuk mempercepat waktu respons (*latency*) pada konteks bahasa Indonesia.

1.5 Batasan Masalah

Mengingat luasnya cakupan pembahasan, maka ditetapkan batasan masalah sebagai berikut:

1. Dataset yang digunakan merupakan data sekunder yang diperoleh dari repositori terbuka *Kaggle* dengan nama *Indonesian News Title Dataset* yang bersumber dari portal berita *detik.com*.
2. Dataset awal hanya memuat atribut *title* dan *category*, sehingga dilakukan proses *web scraping* untuk memperoleh isi berita (*content*).
3. Representasi dokumen dibangun menggunakan kombinasi *title* dan *content* yang telah melalui tahapan *preprocessing*. Proses *stemming* pada model leksikal (BM25) dilakukan menggunakan pustaka *Sastrawi*. Potensi kesalahan *stemming* yang mungkin terjadi akibat karakteristik algoritma *stemming* tidak menjadi fokus penelitian dan diposisikan sebagai keterbatasan sistem.
4. Pendekatan *semantic retrieval* menggunakan model *pretrained* IndoSBERT tanpa dilakukan proses *fine-tuning* tambahan pada domain spesifik.
5. Korpus yang digunakan dalam proses *retrieval* berjumlah 83.967 dokumen berita yang merupakan hasil akhir dari proses pengumpulan dan pembersihan data.
6. Kueri evaluasi dibatasi sebanyak delapan kueri yang masing-masing merepresentasikan satu kategori berita. Setiap kueri dibentuk dari 20 istilah dengan nilai rerata TF-IDF (*mean TF-IDF*) tertinggi yang diperoleh dari dokumen pada masing-masing kategori, yang selanjutnya digabungkan menjadi satu kueri representatif.

7. Evaluasi relevansi (*ground truth*) dibatasi menggunakan teknik *Depth-10 Pooling*. Penilaian manual hanya dilakukan terhadap dokumen yang termasuk dalam *candidate pool* hasil penggabungan *Top-10* BM25 dan IndoSBERT, sedangkan dokumen di luar *candidate pool* tidak menjadi bagian dari proses penilaian relevansi dalam penelitian ini.
8. Evaluasi sistem secara kuantitatif dilakukan menggunakan metrik pemeringkatan tingkat atas, yaitu *Precision@5*, *Precision@10*, *NDCG@5*, dan *NDCG@10*, serta ditinjau dari aspek efisiensi waktu respons komputasi (*latency*).

