

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pencarian informasi di era digital saat ini telah menjadi bagian integral dalam kehidupan sehari-hari. Terutama dengan meningkatnya penggunaan internet dan perkembangan teknologi informasi yang pesat, pencarian artikel berita melalui mesin pencari menjadi salah satu cara utama untuk memperoleh informasi yang relevan. Berita yang relevan dan *up-to-date* sangat dibutuhkan oleh masyarakat untuk mengetahui perkembangan terkini, baik dalam aspek politik, ekonomi, sosial, maupun budaya. Namun, dengan volume informasi yang sangat besar dan terus berkembang, tantangan utama dalam pencarian berita bukan hanya bagaimana menghasilkan hasil yang relevan, tetapi juga bagaimana melakukannya dengan waktu komputasi yang efisien (Hu et al., 2019). Oleh karena itu, penelitian terkait peningkatan efisiensi pencarian artikel berita Indonesia melalui pendekatan klasifikasi kueri pencarian menjadi sangat penting untuk memberikan solusi terhadap permasalahan ini (Ramadhan, 2021).

Latar belakang utama yang mendorong penelitian ini adalah perkembangan pesat dalam dunia digital yang membawa dampak besar terhadap konsumsi informasi masyarakat. Menurut data APJII, jumlah pengguna internet di Indonesia pada tahun 2023 mencapai lebih dari 210 juta orang, dengan sebagian besar mengakses berita melalui internet. Hal ini menciptakan kebutuhan besar terhadap sistem pencarian informasi yang tidak hanya relevan tetapi juga cepat. Permasalahan seperti lambatnya proses pencarian saat dataset berita terus bertambah menjadi keluhan utama pengguna. Dengan demikian, peningkatan efisiensi komputasi dalam pencarian berita menjadi penting untuk meningkatkan kualitas pengalaman pengguna.

Dalam sistem pencarian informasi, relevansi adalah ukuran untuk menentukan seberapa dekat hubungan antara kueri pengguna dengan artikel berita. Klasifikasi kueri merupakan salah satu pendekatan untuk meningkatkan relevansi dan efisiensi pencarian, yang bertujuan untuk mengkategorikan kueri berdasarkan maksud pencariannya (Yu & Litchfield, 2020). Dengan mengetahui kategori kueri

sejak awal, sistem dapat menyaring artikel berita ke dalam *subset* kategori yang relevan sehingga mengurangi ruang pencarian (*search space*). Penyempitan ruang pencarian ini secara signifikan dapat menurunkan kompleksitas komputasi dan mempercepat proses pencarian, terutama pada dataset berita berukuran besar. Penelitian ini berfokus pada penggunaan model BERT (*Bidirectional Encoder Representations from Transformers*) dengan IndoBERT sebagai metode klasifikasi kueri untuk penyaringan awal kategori berita, sehingga sistem hanya membandingkan kueri dengan artikel dalam kategori yang sesuai.

BERT, dengan kemampuannya dalam memahami konteks kata dalam sebuah kalimat, memungkinkan pemrosesan kueri dalam bahasa Indonesia menjadi lebih akurat (Asri et al., 2025). IndoBERT sebagai varian yang telah dilatih menggunakan data bahasa Indonesia menawarkan performa tinggi dalam memahami struktur dan konteks bahasa Indonesia (Priscilya & Girsang, 2024). Dalam penelitian ini, klasifikasi kueri digunakan untuk menentukan kategori awal dari kueri pencarian, sehingga proses *similarity ranking* dengan *Cosine Similarity* dilakukan hanya pada artikel berita dalam kategori yang sesuai. Pendekatan ini tidak hanya mempertahankan relevansi, tetapi juga bertujuan untuk mengurangi waktu komputasi secara signifikan, karena sistem tidak perlu melakukan perhitungan similarity pada seluruh korpus berita.

Metrik evaluasi yang digunakan dalam penelitian ini antara lain MAP (*Mean Average Precision*), nDCG (*Normalized Discounted Cumulative Gain*), Recall, dan Precision. Metrik-metrik ini digunakan untuk mengukur sejauh mana metode klasifikasi kueri dan penyaringan kategori ini dapat meningkatkan efisiensi pencarian tanpa mengorbankan relevansi (Mackie et al., 2023; Pan et al., 2023). MAP dan nDCG mengukur kualitas urutan artikel berdasarkan relevansi, sedangkan Recall dan Precision menggambarkan kemampuan sistem dalam menemukan artikel relevan dalam ruang pencarian yang telah dipersempit.

Penelitian terdahulu terkait relevansi pencarian artikel berita dan klasifikasi kueri telah banyak dilakukan. Garrido-Merchan et al. (Garrido-Merchan et al., 2023) menunjukkan bahwa model BERT unggul dibanding metode tradisional dalam klasifikasi teks. Gusdevi et al. (Gusdevi et al., 2025) menunjukkan bahwa SBERT dengan *Cosine Similarity* efektif dalam mengukur kemiripan semantik.

Sarwar et al. (Sarwar et al., 2022) menunjukkan bahwa penggunaan *Word2Vec* dengan *Cosine Similarity* dapat meningkatkan akurasi artikel serupa. Sarwar et al. (Sarwar et al., 2022) juga membandingkan YAKE dengan algoritma lain untuk ekstraksi keyword pada dataset berita, YAKE menghasilkan skor nDCG yang tinggi (0.94) saat dikombinasikan dengan *Cosine Similarity* dan *Word2Vec*, hanya sedikit di bawah *KP-Miner*, namun lebih baik dari *TopicRank* dan *TextRank* dalam beberapa metrik. Studi oleh Widodo et al. (Widodo & Riana, 2023) menunjukkan IndoBERT efektif untuk memahami teks Indonesia. Namun, sebagian besar penelitian tersebut belum berfokus pada peningkatan efisiensi waktu komputasi melalui penyaringan awal berdasarkan kategori kueri. Oleh karena itu, penelitian ini mengisi celah tersebut dengan menerapkan IndoBERT untuk klasifikasi kueri sebagai langkah penyaringan, sebelum menghitung *Cosine Similarity*.

Tujuan utama dari penelitian ini adalah untuk meningkatkan efisiensi waktu komputasi dalam pencarian artikel berita Indonesia dengan menerapkan penyaringan artikel berdasarkan kategori yang sesuai dengan kategori keri pencarian. Penyaringan ini dilakukan melalui klasifikasi kueri menggunakan IndoBERT sebagai tahap awal, sehingga hanya artikel dalam kategori relevan yang akan diproses dalam perhitungan *similarity*. Dari sisi akademik, penelitian ini memberikan kontribusi dalam bidang NLP, khususnya terkait efisiensi sistem pencarian berbasis IndoBERT. Dari sisi praktis, penelitian ini berpotensi meningkatkan kecepatan dan kualitas pencarian berita bagi pengguna. Ruang lingkup penelitian meliputi pembuatan dataset, klasifikasi kueri, *embedding* teks, penyaringan kategori, perhitungan *Cosine Similarity*, serta evaluasi menggunakan precision.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Pencarian informasi pada artikel berita sering kali tidak efisien dikarenakan volume artikel berita yang besar. Hal ini merupakan masalah penting dalam mencari informasi yang dibutuhkan, masalah yang ditimbulkan ketika melakukan pencarian artikel berita adalah mesin pencari mencocokkan kueri pencarian dengan semua

artikel berita yang ada. Proses tersebut memakan waktu pencarian yang tidak efisien dan hasil artikel yang dicari kurang relevan.

1.2.2 Metode Usulan

Penelitian ini mengusulkan penerapan model *BERT* dengan *IndoBERT*, yang merupakan adaptasi dari model *BERT* untuk bahasa Indonesia, sebagai algoritma klasifikasi kueri untuk menyempitkan ruang pencarian artikel berita Indonesia. Selain itu, solusi yang diusulkan juga mencakup penggunaan *Cosine Similarity* untuk mengukur tingkat kesamaan antara kueri pengguna dengan artikel berita yang tersedia. Dengan menggunakan metode ini, diharapkan mesin pencari dapat memberikan hasil waktu pencarian lebih cepat dan lebih relevan.

1.2.3 Pertanyaan Penelitian

Bagaimana hasil evaluasi kinerja sistem pencarian artikel berita bahasa Indonesia berdasarkan metrik *Precision*, *Recall*, MAP, nDCG, dan waktu komputasi antara skenario pencarian secara menyeluruh dan skenario penyaringan kategori berbasis kueri.

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah melakukan evaluasi dan analisis komparatif terhadap performa sistem pencarian artikel berita Indonesia berdasarkan metrik *Precision*, *Recall*, MAP, nDCG, dan efisiensi waktu komputasi antara penerapan pencarian dengan penyaringan kategori dan pencarian secara menyeluruh.

1.3.2 Manfaat Penelitian

1. Penelitian ini dapat memperkaya literatur dalam bidang *information retrieval* dan *natural language processing* (NLP), khususnya dalam penerapan model pre-trained seperti IndoBERT untuk meningkatkan relevansi pencarian artikel berita dalam bahasa Indonesia.
2. Hasil penelitian ini dapat diterapkan oleh pengembang mesin pencari untuk meningkatkan waktu komputasi dan relevansi pencarian artikel berita Indonesia. Selain itu, penelitian ini juga dapat memberikan panduan bagi

pemangku kepentingan di industri media dan informasi untuk mengoptimalkan sistem pencarian berita dan informasi terkait.

1.4 Batasan Masalah

1. Pengumpulan data artikel berita dilakukan hanya pada satu portal penyedia berita yaitu CNN Indonesia.
2. Data artikel berita yang digunakan hanya terdiri dari judul artikel berita, isi artikel berita, dan kategori yang telah dilabeli oleh portal penyedia berita.
3. Data kueri pencarian merupakan hasil *keyword extraction* dan bukan merupakan hasil riwayat pencarian pengguna.
4. Penelitian ini hanya berfokus pada waktu komputasi *similarity ranking* artikel berita Indonesia dengan 5 kategori yaitu Edukasi, Politik, Ekonomi, Olahraga, dan Travel.
5. Evaluasi *similarity* hanya difokuskan pada tingkat *MAP (Mean Average Precision)*, *nDCG (Normalized Discounted Cumulative Gain)*, *Recall*, *Precision*, dan efisiensi waktu komputasi model.

1.5 Sistematika Penulisan

Adapun sistematika penulisan skripsi yang digunakan terdiri dari beberapa bagian sebagai berikut:

1. Bab I PENDAHULUAN

Bab I membahas mengenai latar belakang penelitian yang menjelaskan alasan dan konteks dilakukannya penelitian, perumusan masalah yang menjadi fokus penelitian, tujuan dan manfaat penelitian, batasan masalah untuk memperjelas ruang lingkup penelitian, serta sistematika penulisan skripsi.

2. Bab II KAJIAN PUSTAKA

Bab II memberikan tinjauan literatur yang berisi konsep-konsep dasar, teori pendukung, dan penelitian terdahulu yang relevan dengan topik penelitian. Bab ini bertujuan untuk membangun landasan teoritis serta menunjukkan posisi penelitian yang dilakukan terhadap penelitian sebelumnya.

3. Bab III METODOLOGI PENELITIAN

Bab III menjelaskan secara menyeluruh metodologi penelitian yang digunakan dalam tugas akhir, mulai dari rancangan dan alur penelitian hingga tahapan pelaksanaan penelitian. Pada bab ini dipaparkan sistem atau model yang dikembangkan, dataset atau sumber data yang digunakan, arsitektur sistem, serta skenario penelitian yang mencakup proses pengumpulan data, preprocessing, perancangan metode, implementasi, dan pengujian. Selain itu, bab ini juga menjelaskan prosedur analisis dan evaluasi hasil menggunakan metrik atau parameter tertentu agar penelitian dapat diukur secara objektif dan dapat direplikasi. Di dalamnya juga dibahas metode penelitian yang lebih spesifik, seperti pendekatan kuantitatif, kualitatif, eksperimen, atau pendekatan lain yang relevan dengan topik penelitian.

4. Bab IV HASIL DAN PEMBAHASAN

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.

5. Bab V KESIMPULAN DAN SARAN

Bab V 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.