BAB 1
PENDAHULUAN

1.1 Latar Belakang

Transformasi digital telah mengubah secara mendasar pola konsumsi dan akses informasi masyarakat global. Pada awal tahun 2025, jumlah pengguna internet dunia mencapai 5,56 miliar orang atau sekitar 67,9% dari populasi global, dengan pertambahan 136 juta pengguna sepanjang tahun 2024. Meskipun demikian, sebanyak 2,63 miliar orang masih belum memiliki akses internet. Kondisi tersebut menunjukkan bahwa internet telah menjadi infrastruktur utama dalam distribusi dan pencarian informasi, sekaligus mendorong pertumbuhan volume konten digital yang tersedia bagi pengguna (DataReportal, 2025). Kelimpahan informasi tersebut memberikan kemudahan dalam memperoleh berita, tetapi pada saat yang sama menimbulkan tantangan berupa *information overload*, yaitu kondisi ketika jumlah informasi yang tersedia atau diterima melebihi kemampuan pengguna untuk memprosesnya secara efektif.

Permasalahan tersebut semakin relevan dalam konsumsi berita digital. *Reuters Institute Digital News Report* (2025) menunjukkan bahwa kelompok usia muda semakin bergeser menuju pola konsumsi berita berbasis media sosial dan platform video. Di antara kelompok usia 18–24 tahun, media sosial menjadi sumber utama berita bagi 39% responden pada 2025, meningkat dari 21% pada 2015. Platform visual seperti Instagram, YouTube, dan TikTok juga semakin berperan dalam distribusi berita kepada kelompok tersebut (Newman dkk., 2025). Di sisi lain, fenomena *news avoidance* menunjukkan bahwa kelimpahan informasi tidak selalu menghasilkan konsumsi informasi yang lebih efektif. Pada 2024, sekitar 39% responden secara global menyatakan bahwa mereka terkadang atau sering menghindari berita, meningkat 10 poin persentase dibandingkan 2017. Selain itu, sekitar 39% responden merasa lelah atau kewalahan dengan jumlah berita yang tersedia (Newman dkk., 2024). Kondisi ini menunjukkan pentingnya sistem *Information Retrieval* (IR) yang mampu membantu pengguna menemukan dokumen yang relevan secara lebih efektif di tengah koleksi berita yang semakin besar.

Dalam sistem IR, salah satu tantangan utama terletak pada bagaimana dokumen dan *query* direpresentasikan sehingga hubungan relevansi dapat dikenali secara tepat. Pendekatan berbasis pencocokan kata secara literal dapat mengalami kegagalan ketika dokumen dan *query* menggunakan kata yang berbeda tetapi memiliki makna yang serupa. Variasi tersebut merupakan salah satu bentuk *synonymy*, sedangkan satu kata yang memiliki lebih dari satu makna dalam konteks berbeda berkaitan dengan polysemy. Oleh karena itu, pencocokan berbasis kata semata belum sepenuhnya mampu merepresentasikan hubungan semantik dalam dokumen berita.

Salah satu metode statistik yang umum digunakan dalam IR adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini memberikan bobot berdasarkan tingkat kepentingan suatu kata dalam dokumen relatif terhadap keseluruhan koleksi, kemudian kemiripan antara *query* dan dokumen dapat dihitung menggunakan *Cosine Similarity* (Jurafsky & Martin, 2024). TF-IDF memiliki keunggulan berupa implementasi yang sederhana, kebutuhan komputasi relatif rendah, serta kemampuan menangani koleksi dokumen berukuran besar. Namun, representasi berbasis frekuensi kata tidak secara langsung memahami hubungan makna antarkata. Kata yang memiliki makna serupa tetapi berbeda secara leksikal tetap direpresentasikan sebagai fitur yang berbeda sehingga dokumen yang secara semantik relevan berpotensi tidak memperoleh peringkat yang tinggi.

Salah satu upaya untuk mengatasi keterbatasan tersebut adalah memperluas query menggunakan sumber pengetahuan leksikal seperti *WordNet*. Melalui pendekatan TF-IDF + *WordNet*, kata-kata pada *query* dapat diperluas dengan sinonim yang relevan sebelum proses pencarian dilakukan. Pendekatan ini mempertahankan keunggulan TF-IDF dalam hal efisiensi komputasi, sekaligus meningkatkan cakupan pencarian terhadap variasi kosakata. Dengan demikian, WordNet dapat diposisikan sebagai pendekatan perantara antara pencocokan berbasis kata dan representasi semantik, karena perluasan dilakukan berdasarkan hubungan leksikal yang telah didefinisikan dalam suatu *lexical database*. Namun, pendekatan tersebut masih memiliki keterbatasan karena hubungan sinonim yang digunakan belum tentu mampu merepresentasikan konteks makna suatu kalimat secara menyeluruh.

Perkembangan *deep learning* dan arsitektur *transformer* kemudian memberikan alternatif melalui representasi teks berbasis semantik. Salah satu pendekatan yang relevan adalah *Sentence*-BERT (SBERT), yang dikembangkan untuk menghasilkan *sentence embeddings* yang memiliki representasi semantik dan dapat dibandingkan menggunakan *Cosine Similarity* (Reimers & Gurevych, 2019). Berbeda dengan TF-IDF yang merepresentasikan dokumen berdasarkan bobot kata dan TF-IDF + *WordNet* yang memperluas kosakata melalui sinonim, SBERT menghasilkan representasi dense yang mempertimbangkan konteks keseluruhan kalimat. Dengan demikian, SBERT berpotensi menemukan hubungan relevansi meskipun *query* dan dokumen tidak memiliki kesamaan kata secara langsung. Akan tetapi, kemampuan semantik tersebut disertai kebutuhan sumber daya komputasi yang lebih besar dibandingkan metode berbasis TF-IDF.

Perbedaan karakteristik tersebut menunjukkan adanya *trade-off* antara kualitas representasi semantik dan efisiensi komputasi. TF-IDF menawarkan efisiensi tinggi tetapi memiliki keterbatasan dalam memahami makna, TF-IDF + *WordNet* berupaya meningkatkan cakupan leksikal dengan tetap mempertahankan efisiensi, sedangkan SBERT menawarkan kemampuan representasi semantik yang lebih kuat dengan konsekuensi kebutuhan komputasi yang lebih besar. Meskipun berbagai penelitian telah menerapkan metode-metode tersebut dalam sistem IR, masih diperlukan evaluasi komparatif yang menguji ketiganya dalam kondisi eksperimen dan dataset yang sama. Terlebih lagi, evaluasi yang hanya berfokus pada kualitas hasil pencarian belum cukup untuk menentukan metode yang paling sesuai bagi sistem berita berskala besar karena efisiensi waktu eksekusi juga menjadi faktor penting.

Berdasarkan kesenjangan tersebut, penelitian ini dirancang untuk membandingkan kinerja TF-IDF, TF-IDF + *WordNet*, dan *Sentence*-BERT dalam pencarian dokumen berita menggunakan BBC *News* Dataset sebagai *benchmark* dengan jumlah dokumen lebih dari 40.000. *Ground truth* relevansi disusun secara hierarkis berdasarkan metadata URL yang merepresentasikan Topik (Level 1) dan Subtopik (Level 2), sehingga kualitas retrieval dapat dievaluasi pada dua tingkat granularitas. Evaluasi dilakukan menggunakan *Precision*@K, *Recall*@K, dan *F1-score*@K dengan K = 5, 10, dan 20, serta pengukuran *execution time*. Untuk

menjaga keseragaman fungsi kemiripan, *Cosine Similarity* digunakan sebagai dasar pengukuran kesamaan antara representasi *query* dan dokumen.

Penelitian ini diharapkan dapat memberikan bukti empiris mengenai perbedaan performa ketiga pendekatan tersebut serta menunjukkan *trade-off* antara akurasi pencarian dan efisiensi komputasi. Meskipun eksperimen menggunakan dataset berita berbahasa Inggris, kerangka metodologi dan evaluasi yang dikembangkan memiliki potensi untuk diadaptasi pada koleksi berita berbahasa Indonesia. Relevansi tersebut semakin penting seiring meningkatnya konektivitas digital di Indonesia, yang pada 2025 telah mencapai tingkat penetrasi internet sebesar 80,66% atau sekitar 229,4 juta pengguna (APJII, 2025). Dengan demikian, hasil penelitian diharapkan tidak hanya memberikan kontribusi akademis dalam kajian *Information Retrieval*, tetapi juga menjadi dasar pertimbangan dalam pemilihan metode representasi dokumen untuk pengembangan sistem pencarian berita yang akurat dan efisien.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah bagaimana perbandingan efektivitas dan efisiensi komputasi (*execution time*) antara pendekatan statistik (TF-IDF), pendekatan hibrida leksikal (TF-IDF + *WordNet*), dan pendekatan neural (*Sentence*-BERT) dalam sistem *retrieval* dokumen berita berdasarkan perbandingan metode pembobotan dan *Cosine Similarity*.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan efektivitas dan efisiensi komputasi (*execution time*) antara pendekatan statistik (TF-IDF), pendekatan hibrida leksikal (TF-IDF + *WordNet*), dan pendekatan neural (*Sentence*-BERT) dalam sistem *retrieval* dokumen berita berdasarkan perbandingan metode pembobotan dan *Cosine Similarity*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoretis maupun praktis, antara lain:

1. Memberikan bukti empiris mengenai efektivitas ekspansi leksikal menggunakan *WordNet* dan *dense embedding* menggunakan SBERT dalam mengatasi masalah *sparsity* (kelangkaan kata) dan rendahnya nilai *Cosine Similarity* pada metode TF-IDF tradisional, khususnya pada *Dataset* berita yang memiliki struktur hierarkis.
2. Memperkaya literatur mengenai perilaku pencarian informasi (*user search intent*) dengan mendemonstrasikan bagaimana variasi panjang dan tingkat spesifisitas *query* (Level 1 yang umum vs Level 2 yang spesifik) berdampak langsung pada performa *retrieval* dari ketiga metode yang diuji.
3. Memberikan rekomendasi teknis bagi *software engineer* atau praktisi NLP dalam memilih model pencarian yang paling optimal berdasarkan skenario penggunaan: apakah untuk kebutuhan browsing umum (di mana TF-IDF+*WordNet*/SBERT lebih unggul) atau pencarian entitas spesifik (di mana TF-IDF murni dengan kata kunci tepat sudah memadai).
4. Menghasilkan metodologi dan desain *query benchmark* yang dapat diadaptasi oleh peneliti selanjutnya untuk mengevaluasi model NLP pada *Dataset* yang memiliki kategori induk (L1) dan subkategori (L2), sehingga pengujian *Preprocessing* evaluasi menjadi lebih terukur dan realistis.

1.5 Batasan Penelitian

Agar penelitian ini dapat dilaksanakan secara terfokus, terukur, dan hasilnya dapat direplikasi oleh peneliti lain, perlu ditetapkan batasan-batasan yang jelas terhadap ruang lingkup eksperimen. Dengan batasan ini, perbedaan performa antar metode dapat diatribusikan secara sah pada perbedaan metode representasi dokumen semata. Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan dalam penelitian ini adalah BBC *News Dataset* berbahasa Inggris yang diperoleh dari platform Kaggle, dengan jumlah data awal sebanyak 42.115 dokumen. Setelah proses *cleaning* (penghapusan duplikat dan *missing value*), diperoleh 40.111 dokumen bersih.

2. Atribut yang digunakan sebagai representasi dokumen dibatasi pada *title dan description*, yang digabungkan menjadi satu teks. Penelitian ini tidak menggunakan isi artikel secara lengkap (*full text*).
3. Tahap *Preprocessing* dibedakan berdasarkan metode yang digunakan, di mana untuk metode TF-IDF dan TF-IDF + *WordNet* dilakukan *Preprocessing* berupa *lowercase*, *tokenization*, penghapusan *stopword*, dan *stemming*, sedangkan pada metode *Sentence-BERT* hanya dilakukan *lowercase* dan normalisasi *whitespace* tanpa *stemming* maupun *stopword removal*.
4. Pada metode TF-IDF + *WordNet*, proses *query expansion* dilakukan pada setiap kata dalam *query* menggunakan *WordNet*.
5. Metrik kesamaan yang digunakan pada seluruh metode adalah *Cosine Similarity*, sehingga perbedaan hasil yang diperoleh hanya disebabkan oleh perbedaan metode representasi dokumen.
6. Penentuan relevansi dokumen (*Ground truth*) dilakukan menggunakan pendekatan berbasis kategori hierarkis (*hierarchical category-based relevance*), bukan penilaian leksikal maupun anotasi manual. Label kategori diekstraksi secara otomatis dari struktur URL pada atribut *link* menggunakan pendekatan berbasis regular *expression* (*regex*), menghasilkan dua tingkat label: Topik (Level 1/L1) sebanyak 5 kategori (*business, technology, sport, politics, dan entertainment*) dan Subtopik (Level 2/L2) sebanyak 53 subkategori.
7. Proses ekstraksi *Ground truth* berhasil memetakan 15.075 dokumen ke dalam struktur Topik/Subtopik. Dari jumlah tersebut, 14.305 dokumen berhasil dicocokkan dengan *Dataset* bersih dan digunakan sebagai korpus final dalam pembangunan representasi dokumen maupun proses *retrieval*, sementara 770 dokumen sisanya tidak diikutsertakan karena tidak ditemukan kecocokan akibat perbedaan normalisasi URL antara *Dataset* mentah dan berkas *Ground truth*.
8. Evaluasi sistem dilakukan menggunakan metrik *Precision*@K, *Recall*@K, dan *F1-score*@K dengan nilai K yang digunakan adalah 5, 10, dan 20, serta ditambahkan pengukuran *execution time* sebagai aspek efisiensi. Nilai

*Recall@*K dihitung menggunakan denominator berupa jumlah dokumen relevan sesuai tingkat pencocokan.

9. Jumlah *query* yang digunakan dalam pengujian dibatasi sebanyak 10 *query* yang disusun untuk mewakili Topik (L1) sebanyak 5 *query* dan Subtopik (L2) sebanyak 5 *query* yang representatif dalam *Dataset*, mencakup kategori *business, technology, politics, entertainments*, dan *sport*s.
10. Penelitian ini difokuskan pada tahap evaluasi performa metode *Information retrieval* dan tidak mencakup aspek implementasi sistem secara produksi maupun optimasi skala besar.