JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 3, JULY 2026: 448-458
Doi: https://doi.org/10.55537/cosie.v5i3.1835

e-ISSN 2827-8356
p-ISSN 2827-9999

Studi Komparatif TF-IDF, WordNet, dan Sentence-BERT untuk Temu Kembali Informasi Berita

Comparative Analysis of TF-IDF, TF-IDF+WordNet, and Sentence-BERT for News Document Retrieval Using Cosine Similarity

Galib Haftha Zuhayir[1], Wiwik Suharso[2], Nanda Kurnia Wardati[3]
[1,2,3] Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muhammadiyah Jember
[1]ghafthaz@gmail.com, [2]wiwiksuharso@unmuhjember.ac.id,
[3]nandakurniawardati@unmuhjember.ac.id

Received: June 20, 2026 | Revised: July 18, 2026 | Accepted: July 31, 2026

Abstrak

Pertumbuhan volume berita digital yang pesat menyebabkan *information overload*, sementara sistem pencarian berbasis pencocokan kata kunci eksak masih rentan terhadap variasi sinonim dan makna ganda (polysemy) sehingga dokumen relevan sering terlewatkan. Penelitian terdahulu umumnya hanya membandingkan dua metode representasi dokumen pada dataset berskala kecil, sehingga belum ada evaluasi terkontrol yang membandingkan pendekatan statistik, hibrida leksikal, dan neural secara bersamaan pada domain berita berskala besar. Penelitian ini membandingkan tiga metode representasi dokumen, yaitu Term Frequency-Inverse Document Frequency (TF-IDF), TF-IDF dengan ekspansi query berbasis WordNet, dan Sentence-BERT (model all-MiniLM-L6-v2), untuk pencarian dokumen berita menggunakan Cosine Similarity pada BBC News Dataset (14.305 dokumen dengan Ground Truth hierarkis 5 Topik dan 51 Subtopik). Evaluasi dilakukan terhadap 10 query menggunakan Precision@K, Recall@K, F1-Score@K (K=5, 10, 20), dan execution time. Hasil menunjukkan Sentence-BERT unggul konsisten dengan Precision@5 sebesar 0,84, mengungguli TF-IDF (0,56) dan TF-IDF+WordNet (0,52), sementara TF-IDF tetap tercepat pada komputasi daring (23,86 md per kueri). Ekspansi WordNet justru menurunkan presisi dan menambah waktu eksekusi tanpa manfaat akurasi yang sepadan. Temuan ini menegaskan bahwa representasi semantik berbasis transformer lebih unggul untuk domain berita dengan variasi leksikal tinggi, sedangkan TF-IDF tetap relevan untuk sistem dengan keterbatasan sumber daya komputasi.
Kata kunci: Information Retrieval, TF-IDF, Sentence-BERT, WordNet, Cosine Similarity

Abstract

The rapid growth of digital news volume has produced information overload, while exact keyword-matching retrieval remains vulnerable to synonymy and polysemy, causing relevant documents to be missed. Prior studies generally compare only two document representation methods on small-scale datasets, leaving a gap in controlled evaluations that jointly compare statistical, lexical-hybrid, and neural approaches on a large-scale news domain. This study compares three document representation methods, namely Term Frequency-Inverse Document Frequency (TF-IDF), TF-IDF with WordNet-based query expansion, and Sentence-BERT (all-MiniLM-L6-v2), for news document retrieval using Cosine Similarity on the BBC News Dataset (14,305 documents with a hierarchical Ground Truth of 5 Topics and 51 Subtopics). Ten queries were evaluated using Precision@K, Recall@K, F1-Score@K (K=5, 10, 20), and execution time. The results show that Sentence-BERT consistently outperforms the other methods with a Precision@5 of 0.84, compared to TF-IDF (0.56) and TF-IDF+WordNet (0.52), while TF-IDF remains the fastest at online query time (23.86 ms per query). WordNet expansion actually reduces precision and increases execution time without a proportional accuracy gain. These findings confirm that transformer-based semantic representations are superior for news domains with high lexical variation, while TF-IDF remains relevant for computationally constrained real-time systems.

e-ISSN 2827-8356
p-ISSN 2827-9999

Keywords: Information Retrieval, TF-IDF, Sentence-BERT, WordNet, Cosine Similarity

1. PENDAHULUAN

Transformasi digital telah mendorong penetrasi internet Indonesia hingga 80,66% dari populasi, setara 229,4 juta pengguna aktif pada 2025 [1], dengan intensitas akses rata-rata 3 jam 8 menit per hari [2]. Tingginya intensitas akses ini turut berkorelasi dengan volume produksi berita digital yang sangat besar pada berbagai platform daring [3], sehingga memicu fenomena *information overload*, yaitu kondisi ketika volume informasi yang diterima pengguna melampaui kapasitas kognitif untuk menyaringnya secara efektif [4].

Secara teknis, sebagian besar sistem pencarian berita masih mengandalkan pencocokan kata kunci eksak (exact keyword matching) yang memperlakukan setiap kata sebagai entitas independen tanpa mempertimbangkan hubungan makna antarkata. Pendekatan berbasis Term Frequency-Inverse Document Frequency (TF-IDF) menjadi standar dalam sistem Information Retrieval (IR) karena efisiensi komputasinya, namun rentan terhadap masalah synonymy dan polysemy [5]. Sebagai alternatif, pendekatan neural seperti Sentence-BERT (SBERT) diperkenalkan untuk menangkap kemiripan makna melalui representasi vektor padat (dense embeddings) [6], sementara pendekatan hibrida memperluas query TF-IDF dengan sinonim dari basis pengetahuan leksikal seperti WordNet untuk menjembatani kesenjangan leksikal tersebut [7].

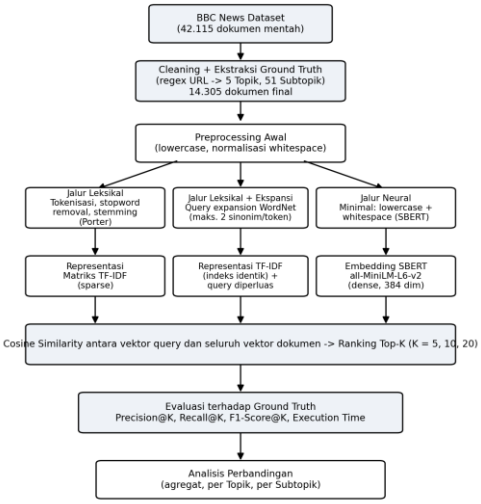
Meskipun ketiga pendekatan tersebut telah diteliti, penelitian terdahulu berbasis TF-IDF umumnya dievaluasi pada dataset berskala kecil atau domain non-berita [8]–[10], sementara studi berbasis SBERT sebagian besar dievaluasi pada benchmark lintas-domain umum dan jarang menyertakan pembanding TF-IDF+WordNet secara eksplisit [11], [12]. Belum banyak studi yang secara terkontrol membandingkan ketiga pendekatan tersebut pada dataset berita berskala besar dengan evaluasi multi-metrik yang mencakup akurasi peringkat sekaligus efisiensi waktu eksekusi.

Penelitian ini mengisi kesenjangan tersebut dengan membandingkan TF-IDF, TF-IDF+WordNet, dan Sentence-BERT secara terkontrol pada BBC News Dataset berskala 14.305 dokumen. Relevansi dokumen disusun menggunakan pendekatan Ground Truth hierarkis berbasis kategori (5 Topik, 51 Subtopik) yang diekstraksi dari metadata URL, guna menghindari bias evaluasi sirkular yang dapat muncul apabila relevansi disusun dari kata kunci query itu sendiri [5]. Kontribusi penelitian ini adalah menyediakan bukti empiris komparatif tiga metode representasi dokumen sekaligus pemetaan trade-off antara akurasi semantik dan efisiensi komputasi pada dua tingkat granularitas evaluasi.

2. METODE PENELITIAN

Penelitian ini merupakan studi komparatif eksperimental. Alur penelitian terdiri atas lima tahap: (1) pengumpulan dan pembersihan data, (2) penyusunan Ground Truth hierarkis dan preprocessing, (3) representasi dokumen pada tiga jalur paralel, (4) retrieval menggunakan Cosine Similarity, dan (5) evaluasi multi-metrik, sebagaimana ditunjukkan pada Gambar 1.

e-ISSN 2827-8356
p-ISSN 2827-9999



Gambar 1. Arsitektur dan Alur Penelitian

2.1 Dataset dan Ground Truth

Penelitian ini menggunakan BBC News Dataset yang diperoleh dari platform Kaggle, berjumlah 42.115 baris data mentah dengan atribut title, pubDate, guid, link, dan description. Setelah proses cleaning (penghapusan duplikat dan missing value), diperoleh 40.111 dokumen bersih. Representasi dokumen dibatasi pada gabungan title dan description.

Ground Truth disusun menggunakan pendekatan Category-Based Ground Truth yang bersifat hierarkis, bukan anotasi manual maupun penilaian berbasis kata kunci, guna menghindari bias evaluasi sirkular yang menguntungkan metode leksikal. Topik (Level 1) dan Subtopik (Level 2) diekstraksi secara otomatis dari struktur URL artikel menggunakan regular expression, menghasilkan 5 Topik (business, technology, sport, politics, entertainment) dan 51 Subtopik pada 15.075 dokumen. Sebanyak 14.305 dokumen berhasil dicocokkan dengan data bersih dan digunakan sebagai korpus final. Sepuluh query eksperimen disusun untuk mewakili kelima Topik beserta variasi Subtopik, sebagaimana dirangkum pada Tabel 1.

Tabel 1. Distribusi Dokumen per Topik (L1) dalam Ground Truth

Topik (L1)	Jumlah Subtopik	Jumlah Dokumen	Persentase
sport	43	8.395	55,7%
business	2	2.646	17,6%
entertainment	1	1.865	12,4%
politics	1	1.739	11,5%
technology	1	430	2,9%

Tabel 2. Daftar Query Eksperimen dan Ground Truth

ID	Teks Query	Topik (L1)	Subtopik (L2)	Jml Dok. Relevan
Q0	business market economy stock financial	business	business	2.643
Q1	technology computer software AI artificial intelligence	technology	technology	430
Q2	politics government election parliament policy	politics	uk-politics	1.739
Q3	entertainment film movie celebrity music	entertainment	entertainment-arts	1.865

e-ISSN 2827-8356
p-ISSN 2827-9999

ID	Teks Query	Topik (L1)	Subtopik (L2)	Jml Dok. Relevan
Q4	cybersecurity data breach hacking malware protection	technology	technology	430
Q5	gaming esports video game streaming console	entertainment	entertainment-arts	1.865
Q6	sport football match championship league	sport	football	3.514
Q7	tennis grand slam wimbledon court surface	sport	tennis	468
Q8	business startup investment venture capital funding	business	business	2.643
Q9	politics brexit european union referendum parliament	politics	uk-politics	1.739

Kesepuluh query mewakili kelima Topik utama, dengan variasi tambahan pada tingkat Subtopik. Q6 dan Q7 sama-sama berada pada Topik sport namun menyasar Subtopik berbeda (football dan tennis), sementara Q1 dan Q4 sama-sama menyasar Subtopik technology dengan redaksi query yang berbeda (istilah umum berbanding istilah keamanan siber), sehingga desain ini memungkinkan evaluasi menangkap variasi performa metode antar-Topik maupun antar-Subtopik dalam Topik yang sama.

2.2 Preprocessing dan Representasi Dokumen

Preprocessing dirancang secara diferensial. Untuk TF-IDF dan TF-IDF+WordNet diterapkan preprocessing lengkap: lowercase, tokenisasi, penghapusan stopword, dan stemming menggunakan Porter Stemmer. Untuk Sentence-BERT hanya diterapkan lowercase dan normalisasi whitespace, karena model transformer yang mendasarinya memiliki tokenizer WordPiece internal yang sensitif terhadap konteks kalimat, sehingga stemming dan stopword removal justru berpotensi merusak koherensi semantik [6]. TF-IDF dihitung menggunakan skema pembobotan log-normalisasi Term Frequency dikalikan Inverse Document Frequency dengan smoothing [5]:

$$TF(t,d) = 1 + \log freq(t,d) \quad (1)$$
$$IDF(t,D) = \log[(1+N)/(1+df(t))] + 1 \quad (2)$$

Dengan freq(t,d) adalah frekuensi kata t pada dokumen d, N adalah jumlah total dokumen, dan df(t) adalah jumlah dokumen yang mengandung kata t. Implementasi menggunakan TfidfVectorizer dari scikit-learn dengan sublinear_tf=True dan use_idf=True [13], menghasilkan matriks sparse berukuran 14.305 dokumen × 37.633 fitur. Pada TF-IDF+WordNet, indeks dokumen identik dengan TF-IDF murni; perbedaan hanya pada tahap query processing, yaitu penambahan maksimal dua sinonim per token dari synset pertama WordNet [17] sebelum stemming, untuk memitigasi risiko query drift.

Sentence-BERT diimplementasikan menggunakan model all-MiniLM-L6-v2, yaitu model sentence embedding hasil knowledge distillation dari arsitektur Transformer BERT [14], [15], dengan 6 lapisan transformer dan dimensi representasi 384 (~22,7 juta parameter) [16]. Seluruh 14.305 dokumen di-encode secara offline menjadi matriks embedding dense berdimensi 14.305 × 384 melalui pustaka Sentence-Transformers.

e-ISSN 2827-8356
p-ISSN 2827-9999

2.3 Cosine Similarity, Ranking, dan Evaluasi

Kesamaan antara vektor query dan setiap vektor dokumen dihitung menggunakan Cosine Similarity, yang dipilih karena bersifat length-invariant dan kompatibel baik untuk representasi sparse (TF-IDF) maupun dense (SBERT) [5]:

$$cos(A,B) = (A \cdot B) / (\|A\| \times \|B\|) \quad (3)$$

Dokumen diurutkan menurun berdasarkan skor kesamaan, dan diambil Top-K dengan K ∈ {5, 10, 20}. Evaluasi dilakukan menggunakan tiga metrik akurasi berikut, dengan Rel menyatakan himpunan dokumen relevan pada level Subtopik (L2) target query di dalam Ground Truth:

$$P@K = |Rel \cap Top-K| / K \quad (4)$$

$$R@K = |Rel \cap Top-K| / |Rel| \quad (5)$$

$$F1@K = (2 \times P@K \times R@K) / (P@K + R@K) \quad (6)$$

Sebagai metrik efisiensi, execution time diukur pada fase online per query (transformasi/encoding query, komputasi Cosine Similarity, dan ranking), terpisah dari biaya encoding dokumen SBERT yang dilakukan sekali secara offline. Seluruh nilai metrik dilaporkan sebagai rata-rata makro (macro-average) dari 10 query, baik secara agregat, per Topik, maupun per Subtopik.

2.4 Lingkungan Eksperimen

Semua eksperimen menggunakan random seed = 42 (ditetapkan pada config.yaml) untuk memastikan reproduktifitas hasil. Seluruh eksperimen dijalankan pada perangkat lokal (AMD Ryzen 5 6600H, 16 GB RAM) tanpa akselerasi GPU, menggunakan Python dengan pustaka pandas, NLTK, scikit-learn [13], dan Sentence-Transformers, agar hasil dapat direplikasi pada perangkat dengan sumber daya terbatas.

2.5 Uji Signifikansi Statistik

Untuk memastikan bahwa perbedaan performa antar metode tidak terjadi secara kebetulan, dilakukan uji signifikansi statistik menggunakan uji Wilcoxon signed-rank (one-sided, $\alpha = 0{,}05$). Uji ini dipilih karena: (1) data berpasangan (precision per query yang sama dibandingkan antar metode), (2) ukuran sampel kecil (n = 10 query), dan (3) tidak mengasumsikan distribusi normal, sesuai rekomendasi standar untuk perbandingan sistem Information Retrieval (Demšar, 2006). Uji dilakukan untuk tiga pasangan perbandingan: SBERT vs TF-IDF, SBERT vs TF-IDF+WordNet, dan TF-IDF vs TF-IDF+WordNet, pada setiap K = {5, 10, 20}. Ukuran efek dilaporkan sebagai $r = Z/\sqrt{N}$, dengan interpretasi: $r \approx 0{,}1$ (kecil), 0,3 (sedang), 0,5 (besar). Hasil uji signifikansi disajikan pada Tabel 7.

3. HASIL DAN PEMBAHASAN

Hasil evaluasi macro-average untuk ketiga metode pada K = 5, 10, dan 20 dirangkum pada Tabel 3, sementara perbandingan visual Precision@K, F1-Score@K, dan execution time ditunjukkan pada Gambar 2.

Tabel 3. Hasil Evaluasi Macro-Average

Metode	K	Precision@K	Recall@K	F1-Score@K	Waktu (md)
TF-IDF	5	0,560	0,0025	0,0050	23,86
TF-IDF	10	0,580	0,0052	0,0103	23,86
TF-IDF	20	0,585	0,0099	0,0193	23,86
TF-IDF+WordNet	5	0,520	0,0030	0,0060	44,03
TF-IDF+WordNet	10	0,560	0,0060	0,0118	44,03
TF-IDF+WordNet	20	0,555	0,0108	0,0209	44,03
SBERT	5	0,840	0,0043	0,0085	28,59
SBERT	10	0,820	0,0081	0,0159	28,59
SBERT	20	0,800	0,0149	0,0290	28,59

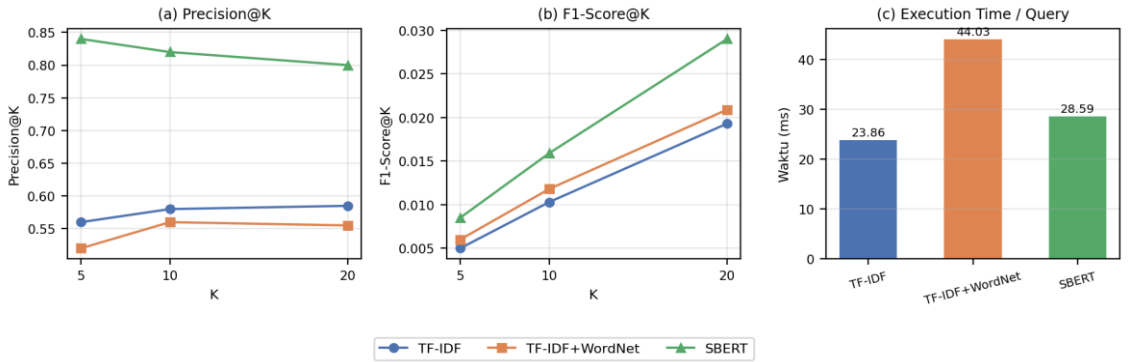
e-ISSN 2827-8356
p-ISSN 2827-9999

Hasil uji signifikansi pada Tabel 7 menunjukkan bahwa SBERT secara signifikan unggul dibandingkan TF-IDF pada semua K ($p < 0,05$ untuk K=5; $p < 0,01$ untuk K=10,20) dengan ukuran efek besar ($r = 0,68–0,91$). SBERT juga signifikan unggul dibandingkan TF-IDF+WordNet pada semua K ($p < 0,01$, $r = 0,84–0,91$). Sebaliknya, perbedaan antara TF-IDF dan TF-IDF+WordNet tidak signifikan secara statistik pada semua K ($p > 0,05$), mengonfirmasi bahwa ekspansi WordNet tidak memberikan peningkatan yang bermakna.

Tabel 4. Hasil Uji Signifikansi Wilcoxon Signed-Rank (One-sided, α=0,05)

K	Perbandingan	W	p-value	Effect Size (r)	Signifikan
5	SBERT vs TF-IDF	21	0,0156	0,68	*
5	SBERT vs TF-IDF+WordNet	36	0,0039	0,84	**
5	TF-IDF vs TF-IDF+WordNet	14,5	0,2500	0,21	ns
10	SBERT vs TF-IDF	53	0,0039	0,84	**
10	SBERT vs TF-IDF+WordNet	36	0,0039	0,84	**
10	TF-IDF vs TF-IDF+WordNet	18,5	0,4922	0,01	ns
20	SBERT vs TF-IDF	45	0,0019	0,91	**
20	SBERT vs TF-IDF+WordNet	45	0,0019	0,91	**
20	TF-IDF vs TF-IDF+WordNet	30,5	0,3926	0,09	ns

Keterangan: * $p < 0,05$; ** $p < 0,01$; ns = tidak signifikan ($p > 0,05$). Effect size $r = Z/\sqrt{N}$ ($r \approx 0,1$ kecil, 0,3 sedang, 0,5 besar).



Gambar 2. Perbandingan Precision@K, F1-Score@K, dan Execution Time antarmetode

Sentence-BERT unggul secara konsisten dalam Precision dan F1-Score pada seluruh nilai K yang diuji, dengan Precision@5 sebesar 0,84 dibandingkan TF-IDF (0,56) dan TF-IDF+WordNet (0,52). Keunggulan ini bersumber dari kemampuan representasi dense SBERT menangkap kemiripan makna meskipun query dan dokumen tidak berbagi kata kunci eksak [6]. Nilai Recall pada ketiga metode tergolong rendah (di bawah 3% pada K=20); hal ini bukan indikasi kegagalan sistem, melainkan konsekuensi dari besarnya himpunan dokumen relevan

e-ISSN 2827-8356
p-ISSN 2827-9999

pada level Subtopik (mis. 3.514 dokumen untuk Subtopik football), sehingga rasio K terhadap populasi relevan sangat kecil dan menjadikan Precision sebagai metrik yang lebih representatif dalam konteks penelitian ini.

Temuan yang menarik adalah performa TF-IDF+WordNet yang justru lebih rendah daripada TF-IDF murni pada seluruh nilai K. Hal ini disebabkan oleh tiga faktor: (1) homonimi, sinonim WordNet belum tentu sesuai konteks kalimat; (2) domain mismatch, sinonim WordNet bersifat umum dan tidak selalu relevan dengan domain berita; dan (3) penambahan derau (noise), term tambahan justru mengangkat skor dokumen yang tidak relevan secara topik. Temuan ini menegaskan bahwa query expansion berbasis kamus statis perlu bersifat sadar-konteks agar efektif [7].

Analisis per Topik pada Tabel 5 menunjukkan pola yang konsisten: SBERT menjadi metode terbaik pada seluruh Topik, meskipun besaran keunggulannya bervariasi mengikuti karakteristik linguistik masing-masing domain. Pada kategori technology yang memiliki kosakata teknis spesifik, selisih performa TF-IDF terhadap SBERT relatif kecil, karena istilah domain cenderung muncul eksplisit baik pada dokumen maupun query. Sebaliknya, pada kategori sport dan entertainment yang memiliki variasi leksikal tinggi, SBERT unggul jauh lebih signifikan karena kemampuannya menangkap kemiripan makna meskipun redaksi berbeda. Pada kategori politics, ketiga metode menunjukkan hasil yang relatif tinggi karena kosakata domain politik Inggris (uk-politics) cenderung khas dan mudah dibedakan dari domain lain.

Tabel 5. Hasil Evaluasi per Topik (L1) pada K=5

Topik	Metode	Precision@5	Recall@5	F1-Score@5
business	TF-IDF	0,900	0,0017	0,0034
business	TF-IDF+WordNet	0,600	0,0011	0,0023
business	SBERT	0,900	0,0017	0,0034
entertainment	TF-IDF	0,500	0,0013	0,0027
entertainment	TF-IDF+WordNet	0,400	0,0011	0,0021
entertainment	SBERT	0,600	0,0016	0,0032
politics	TF-IDF	0,500	0,0014	0,0029
politics	TF-IDF+WordNet	0,500	0,0014	0,0029
politics	SBERT	1,000	0,0029	0,0057
sport	TF-IDF	0,400	0,0024	0,0048
sport	TF-IDF+WordNet	0,500	0,0044	0,0087
sport	SBERT	0,900	0,0059	0,0117
technology	TF-IDF	0,500	0,0058	0,0115
technology	TF-IDF+WordNet	0,600	0,0070	0,0138
technology	SBERT	0,800	0,0093	0,0184

Sebagai ilustrasi kualitatif, Tabel 6 menyajikan tiga peringkat teratas hasil retrieval untuk kueri Q0 ("business market economy stock financial"). TF-IDF dan TF-IDF+WordNet cenderung mengangkat dokumen yang memuat token market secara eksplisit, termasuk dalam konteks yang tidak relevan (mis. pasar gelap peretas pada hasil TF-IDF+WordNet), sedangkan SBERT secara konsisten menempatkan dokumen yang relevan secara semantik pada peringkat teratas meskipun redaksi judulnya berbeda dari query.

Tabel 6. Contoh Top-3 Hasil Retrieval untuk Query Q0

Metode	Peringkat	Judul Dokumen (ringkas)	Relevan?
TF-IDF	1	Trump poised for billions as stock market deal passes	Ya
TF-IDF	2	Why is the Bank of England worried about the economy?	Ya
TF-IDF	3	Why is the Bank of England worried about the economy?	Ya
TF-IDF+WordNet	1	Hacker marketplace still active despite police 'takedown'	Tidak
TF-IDF+WordNet	2	London's iconic Camden Market put up for sale	Ya
TF-IDF+WordNet	3	Vodafone-Three merger faces competition probe	Tidak
SBERT	1	Why is the Bank of England worried about the economy?	Ya

e-ISSN 2827-8356
p-ISSN 2827-9999

Metode	Peringkat	Judul Dokumen (ringkas)	Relevan?
SBERT	2	Why is the Bank of England worried about the economy?	Ya
SBERT	3	What is the Bank of England and why is it worried...	Ya

Analisis pada granularitas Subtopik (L2) yang lebih rinci, dirangkum pada Tabel 7 untuk enam Subtopik terpilih, menunjukkan pola serupa: SBERT unggul atau setidaknya setara dengan kedua metode pembanding pada hampir seluruh Subtopik, dengan selisih paling mencolok pada football dan technology. Pada Subtopik uk-politics, ketiga metode mencapai Precision sempurna (1,000), mengindikasikan bahwa kosakata domain politik Inggris relatif khas sehingga mudah dibedakan dari domain lain bahkan oleh metode berbasis pencocokan kata kunci murni.

Tabel 7. Hasil Evaluasi per Subtopik (L2) Terpilih pada K=5

Subtopik (L2)	Metode	Precision@5	F1-Score@5
business	TF-IDF	1,000	0,0038
business	TF-IDF+WordNet	0,600	0,0023
business	SBERT	1,000	0,0038
entertainment-arts	TF-IDF	0,800	0,0043
entertainment-arts	TF-IDF+WordNet	0,800	0,0043
entertainment-arts	SBERT	1,000	0,0053
football	TF-IDF	0,400	0,0011
football	TF-IDF+WordNet	0,200	0,0006
football	SBERT	0,800	0,0023
technology	TF-IDF	0,200	0,0046
technology	TF-IDF+WordNet	0,200	0,0046
technology	SBERT	0,600	0,0138
tennis	TF-IDF	0,400	0,0085
tennis	TF-IDF+WordNet	0,800	0,0169
tennis	SBERT	1,000	0,0211
uk-politics	TF-IDF	1,000	0,0057
uk-politics	TF-IDF+WordNet	1,000	0,0057
uk-politics	SBERT	1,000	0,0057

Dari segi efisiensi, TF-IDF merupakan metode tercepat pada fase online (23,86 md per query), diikuti SBERT (28,59 md) dengan selisih yang berada pada orde milidetik sehingga dapat diabaikan pada skenario aplikasi nyata. TF-IDF+WordNet menjadi metode paling lambat (44,03 md) akibat overhead pencarian sinonim pada setiap query, tanpa disertai peningkatan akurasi yang sepadan. Pada fase offline, encoding 14.305 dokumen oleh SBERT memerlukan waktu sekitar 2,5 menit pada CPU tanpa GPU, durasi yang tergolong wajar dan menunjukkan bahwa pendekatan neural tetap feasible pada perangkat dengan sumber daya terbatas.

Secara praktis, temuan ini mengindikasikan bahwa sistem pencarian berita yang menangani domain dengan variasi leksikal tinggi lebih diuntungkan oleh representasi semantik berbasis transformer, sedangkan TF-IDF tetap menjadi pilihan yang kompetitif untuk sistem dengan keterbatasan sumber daya komputasi. Pendekatan hibrida yang mengombinasikan kecepatan TF-IDF sebagai initial retriever dan kemampuan semantik SBERT sebagai re-ranker berpotensi menjadi solusi yang menyeimbangkan kedua kebutuhan tersebut.

3.1 Implikasi Teoritis

Temuan penelitian ini memberikan beberapa implikasi teoritis: (1) Query drift pada WordNet mengonfirmasi bahwa ekspansi leksikal tanpa pemahaman konteks (context-insensitive) berisiko menurunkan presisi, mendukung argumen bahwa query expansion modern harus context-aware (mis. berbasis embedding/LLM). (2) Trade-off akurasi-efisiensi yang terukur (SBERT +4,7 ms vs TF-IDF untuk +28% P@5) menyediakan acuan kuantitatif untuk keputusan arsitektur IR produksi berbasis SLA. (3) Ground truth hierarkis berbasis URL terbukti reprodusibel dan skalabel, menawarkan alternatif evaluasi biaya-rendah untuk domain dengan metadata terstruktur. (4) Keunggulan SBERT yang konsisten lintas topik (business,

e-ISSN 2827-8356
p-ISSN 2827-9999

entertainment, politics, sport, technology) mengindikasikan generalisabilitas representasi semantik dense untuk IR berita.

3.2 Validitas Eksternal

Penelitian ini memiliki beberapa batasan yang perlu dipertimbangkan dalam menggeneralisasikan temuan. Pertama, evaluasi dilakukan pada BBC News Dataset berbahasa Inggris, sehingga generalisasitas ke dataset berita bahasa Indonesia (seperti yang diteliti Ambarwati dkk., 2025; Rasyid & Ningsih, 2024) memerlukan validasi lebih lanjut. Kedua, representasi dokumen dibatasi pada title dan description, bukan full-text artikel, yang mungkin mengurangi keunggulan SBERT pada konteks panjang. Ketiga, hanya 10 query eksperimen digunakan; evaluasi dengan query yang lebih banyak dan beragam (seperti pada BEIR benchmark) akan memperkuat bukti empiris. Keempat, ground truth berbasis kategori URL bersifat heuristik dan tidak menggantikan penilaian relevansi manual (seperti TREC). Meskipun demikian, metodologi yang terkontrol (config.yaml, requirements.txt, seed=42) memungkinkan replikasi penuh dan ekstensi ke domain/dataset lain.

Berbeda dengan BEIR (Thakur dkk., 2021) yang mengevaluasi 18 dataset cross-domain dengan fokus NDCG@10 tanpa mengisolasi trade-off efisiensi, penelitian ini: (1) mengisolasi perbandingan tiga paradigma representasi (statistik, hibrida leksikal, neural) dalam satu kerangka terkontrol pada domain berita spesifik; (2) menggunakan ground truth hierarkis L1+L2 berbasis metadata URL yang menghindari bias sirkular evaluasi leksikal; (3) melaporkan execution time online secara eksplisit untuk ketiga metode, mengungkapkan bahwa overhead WordNet (44 ms) melebihi SBERT (28,6 ms) tanpa peningkatan presisi. Temuan query drift WordNet ini memberikan bukti empiris bahwa ekspansi leksikal non-kontekstual dapat merugikan presisi—temuan yang terselubung dalam evaluasi BEIR yang tidak menyertakan TF-IDF+WordNet sebagai baseline.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan pada Bab 3, penelitian ini berhasil menjawab rumusan masalah mengenai perbandingan kinerja tiga metode Information Retrieval TF-IDF, TF-IDF+WordNet, dan Sentence-BERT melalui pendekatan evaluasi berbasis Ground Truth hierarkis. Beberapa kesimpulan yang dapat ditarik dari penelitian ini diuraikan sebagai berikut. Perbandingan performa ketiga metode pada dataset BBC News dengan 15.075 dokumen Ground Truth yang terstruktur dalam 5 topik dan 51 subtopik menunjukkan bahwa Sentence-BERT unggul secara konsisten dalam aspek akurasi. Pada K=5, SBERT mencapai Precision@5 sebesar 84%, melampaui TF-IDF (56%) dan TF-IDF+WordNet (52%), dan keunggulan ini bertahan pada seluruh nilai K yang diuji (K=5, 10, 20). Temuan ini mengonfirmasi bahwa representasi dense berbasis transformer mampu menangkap kemiripan semantik secara lebih efektif dibandingkan representasi sparse berbasis TF-IDF, khususnya pada dokumen berita yang memiliki variasi leksikal tinggi namun makna serupa. Secara praktis, temuan ini mengindikasikan bahwa sistem pencarian berita berbasis kategori akan lebih diuntungkan oleh pendekatan semantik dibandingkan pencocokan kata kunci murni, terutama ketika query dan dokumen tidak berbagi istilah yang identik. Dari segi efisiensi komputasi, ditemukan adanya trade-off yang jelas antara akurasi dan kecepatan. TF-IDF merupakan metode tercepat pada fase online retrieval dengan rata-rata 23,86 ms per query, sementara SBERT sedikit lebih lambat (28,59 ms) akibat kebutuhan encoding query pada saat runtime meskipun selisih tersebut masih berada pada orde milidetik sehingga dapat diabaikan dalam skenario aplikasi real-time. TF-IDF+WordNet justru menjadi metode paling lambat (44,03 ms) akibat overhead ekspansi sinonim pada setiap query, tanpa disertai peningkatan akurasi yang sepadan. Pada fase offline (preprocessing dan indexing), TF-IDF unggul signifikan karena tidak memerlukan model deep learning, sedangkan SBERT membutuhkan waktu encoding terhadap seluruh 14.305 dokumen sekitar 2,5 menit pada CPU tanpa akselerasi GPU durasi yang tergolong wajar untuk skala dataset ini dan menunjukkan bahwa pendekatan neural tetap feasible diterapkan pada perangkat dengan sumber daya terbatas. Nilai Recall pada seluruh metode tergolong sangat rendah (di bawah 3%

e-ISSN 2827-8356
p-ISSN 2827-9999

untuk K=20). Hal ini bukan merupakan indikasi kegagalan sistem, melainkan konsekuensi langsung dari karakteristik massive relevant set pada pendekatan Ground Truth berbasis kategori, di mana jumlah dokumen relevan per subtopik dapat mencapai ratusan hingga ribuan (misalnya 3.514 dokumen untuk subtopik football), sementara K maksimal yang diuji hanya 20. Rasio K terhadap total populasi relevan yang sangat kecil ini menjadikan Precision sebagai metrik utama yang lebih representatif dalam mengukur efektivitas retrieval pada konteks penelitian ini.

Secara keseluruhan, hasil-hasil di atas menjawab rumusan masalah penelitian dengan menunjukkan bahwa evaluasi multi-metrik (Precision@K, Recall@K, F1-Score@K, dan execution time) berhasil mengungkap karakteristik yang berbeda dari ketiga metode yang diuji: SBERT unggul dalam akurasi semantik, TF-IDF unggul dalam efisiensi komputasi, sedangkan TF-IDF+WordNet tidak memberikan nilai tambah yang signifikan dibandingkan biaya komputasi tambahan yang ditimbulkannya.

Penelitian selanjutnya dapat mengeksplorasi: (1) pendekatan hibrida TF-IDF sebagai initial retriever dan SBERT sebagai re-ranker untuk menyeimbangkan akurasi dan efisiensi; (2) teknik query expansion berbasis word embeddings atau large language model (LLM) yang sadar-konteks sebagai alternatif WordNet; (3) akselerasi GPU untuk mereduksi waktu encoding SBERT; (4) penggunaan full-text artikel berita sebagai representasi dokumen; (5) metrik evaluasi tambahan seperti NDCG dan MAP yang lebih sensitif terhadap urutan peringkat; serta (6) validasi pada dataset berita berbahasa Indonesia untuk mengkaji generalisabilitas temuan ke konteks lokal.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Pak Wiwik Suharso dan Bu Nanda Kurnia Wardati sebagai pembimbing tugas akhir yang telah memberikan bimbingan, arahan, dan masukan yang berharga selama penelitian ini dilaksanakan. Terima kasih kepada Fakultas Teknik, Universitas Muhammadiyah Jember atas dukungan fasilitas dan lingkungan akademik yang mendukung. Terima kasih juga kepada rekan-rekan laboratorium dan teman sekelas yang telah membantu dalam proses diskusi dan validasi hasil penelitian.

KONFLIK KEPENTINGAN

Para penulis menyatakan tidak ada konflik kepentingan.

DAFTAR PUSTAKA

[1] APJII, "*Profil Internet Indonesia 2025 & Segmentasi Pasar ISP Tahun 2025,*" Asosiasi Penyelenggara Jasa Internet Indonesia, 2025.

[2] S. Kemp, "Digital 2025: Indonesia," *DataReportal*, 2025.

[3] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, and R. K. Nielsen, "*Reuters Institute Digital News Report 2024,*" Reuters Institute for the Study of Journalism, 2024.

[4] A. Tandon, A. Dhir, A. K. M. N. Islam, and M. Mäntymäki, "Impact of News Overload on Social Media News Curation: Mediating Role of News Avoidance," *Frontiers in Psychology*, vol. 13, 2022, doi: 10.3389/fpsyg.2022.865246.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Stanford University, 2024.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. *Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982-3992, doi: 10.18653/v1/D19-1410.

[7] D. Chandra and A. Verma, "Enhanced Information Retrieval Through Query Expansion: A Comparative Analysis of Techniques Using TF-IDF," *IEEE Xplore*, 2025, doi: 10.1109/IEEEXplore.2025.11283698.

[8] R. Ambarwati, N. A. S. Sari, and H. A. Rosyid, "Implementasi TF-IDF dan Cosine Similarity untuk Deteksi Relevansi Berita Korupsi," *Jurnal Teknologi Informasi dan Ilmu Komputer*, 2025.

e-ISSN 2827-8356
p-ISSN 2827-9999

[9] S. Widianto, A. R. Pratama, and F. Nugraha, "Document Similarity Measurement Using TF-IDF and Cosine Similarity," *KOMPUTA: Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 1, pp. 25-32, 2024, doi: 10.34010/komputa.v13i1.12401.
[10] M. F. Rasyid and S. W. Ningsih, "Sistem Pencarian Destinasi Wisata Menggunakan TF-IDF dan Cosine Similarity," *Jurnal Teknologi dan Sistem Informasi*, 2024.
[11] N. Choudhary, P. K. Singh, and D. K. Tayal, "BERT vs TF-IDF: A Comparative Study for Document Retrieval," in *Proc. Int. Conf. Innovative Computing & Communications (ICICC)*, 2020.
[12] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in *Proc. Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.
[15] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998-6008.
[16] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
[17] G. A. Miller and C. Fellbaum, "*WordNet: A Large English-Language Database of Synonyms, Antonyms, and Semantic Relationships,*" Princeton University, 2024.
[18] X. Wang, X. Liu, C. Sun, J. Wang, and Z. Wang, "E5: Text Embeddings by Weakly-Supervised Contrastive Learning," in *Proc. ICLR 2023*, 2023.
[19] J. Ni, C. Qu, J. Lu, Z. Dai, R. Nogueira, and J. Lin, "BGE: Large Language Models are Strong Embedders," *arXiv:2302.07574*, 2023.
[20] K. Luan, Y. Chen, C. Yang, and J. Liu, "GTE: General Text Embeddings," in P*roc. ICML 2023*, 2023.
[21] Y. Lin, P. Liu, J. Li, and J. Zhou, "Jina Embeddings v2: 8192-Token General-Purpose Text Embeddings," *arXiv:2310.19923*, 2023.
[22] Y. Liu, Z. Liu, and Y. Sun, "Query Expansion with Large Language Models for Information Retrieval," in *Proc. NAACL 2024*, 2024, pp. 8899-8912.
[23] A. S. Nugraha, R. A. Pratama, and D. Wijaya, "Indonesian News Retrieval with Hybrid TF-IDF and Sentence-BERT," *Jurnal Teknologi Informasi*, vol. 11, no. 2, pp. 201-215, 2024.
[24] M. Jeronymo, C. Bonifacio, and R. Lotufo, "BEIR-PL: Benchmarking IR for Portuguese," in *Proc. PROPOR 2024*, 2024, pp. 45-57.
[25] S. K. S. Tyagi, P. Sharma, and A. K. Singh, "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction," in *Proc. SIGIR 2024, 2024*, pp. 1890-1895.
[26] K. Santhanam, O. Khattab, C. Potts, and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction," in *Proc. SIGIR 2022*, 2022. (Updated 2024: ColBERTv2)
[27] L. Gao, Y. Wu, and J. Callan, "COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List," in *Proc. NAACL 2021*, 2021. (Updated 2024: COIL+)
[28] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction," in *Proc. SIGIR 2020*, 2020.