Perbandingan metode lexical dan semantic *retrieval* : BM25, IndoSBERT baseline, dan IndoSBERT fine-tuned pada pencarian dokumen berbahasa Indonesia

Dela Puspita Lasminingrum*, Eva Yulia Puspaningrum, Budi Mukhammad Mulyo

Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
*Email: *22081010209@student.upnjatim.ac.id

Abstrak

Sistem pencarian skripsi pada repositori perguruan tinggi umumnya masih mengandalkan pendekatan lexical berbasis kata kunci, seperti BM25, yang efektif ketika terdapat kecocokan istilah antara query dan dokumen. Namun, pendekatan ini memiliki keterbatasan dalam menangkap kesamaan makna ketika istilah yang digunakan berbeda. Penelitian ini melakukan studi komparatif antara metode lexical BM25 dan metode semantic retrieval berbasis IndoSBERT, baik dalam kondisi pretrained (baseline) maupun setelah fine-tuning, pada repositori skripsi Program Studi Informatika UPN "Veteran" Jawa Timur. Dataset yang digunakan terdiri dari 1.146 dokumen skripsi, dengan evaluasi dilakukan menggunakan ground truth relevansi yang dilabeli secara manual melalui pendekatan pooling. Kinerja sistem dievaluasi menggunakan metrik Precision@5, Recall@5, Mean Average Precision (MAP), dan nDCG@5. Hasil eksperimen menunjukkan bahwa BM25 memiliki performa terbaik pada metrik presisi dan kualitas peringkat untuk query pendek dan eksplisit. Sementara itu, pendekatan semantic retrieval mampu menangkap hubungan makna antar dokumen, tetapi menunjukkan keterbatasan dalam menangani istilah spesifik dan frasa penting. Fine-tuning IndoSBERT memberikan peningkatan pada kualitas pemeringkatan secara keseluruhan, meskipun belum mampu melampaui performa BM25. Temuan ini menunjukkan bahwa efektivitas metode pencarian sangat dipengaruhi oleh karakteristik koleksi dokumen dan sifat query, sehingga pemilihan pendekatan perlu disesuaikan dengan konteks penggunaan sistem pencarian.

Kata Kunci: sistem pencarian; BM25; IndoSBERT; *lexical retrieval*; *semantic retrieval*

A comparative study of lexical and semantic retrieval models: BM25, baseline SBERT, and fine-tuned SBERT for Indonesian document search

Abstract

Dicument retrieval systems in higher education repositories generally still rely on lexical keyword-based approaches, such as BM25, which are effective when there is strong term overlap between the query and documents. However, this approach has limitations in capturing semantic similarity when different terms are used to express the same meaning. This study conducts a comparative analysis between the lexical BM25 method and semantic retrieval approaches based on IndoSBERT, both in the pretrained (baseline) setting and after fine-tuning, on the thesis repository of the Informatics Study Program at UPN "Veteran" East Java. The dataset consists of 1,146 thesis documents, with evaluation performed using manually labeled relevance ground truth obtained through a pooling strategy. System performance is evaluated using Precision@5, Recall@5, Mean Average Precision (MAP), and nDCG@5. Experimental results show that BM25 achieves the best performance in terms of precision and ranking quality for short and explicit queries. Meanwhile, semantic retrieval approaches are able to capture semantic relationships between documents but exhibit limitations in handling specific terms and important phrases. Fine-tuning IndoSBERT improves overall ranking quality, although it does not surpass the performance of BM25. These findings indicate that retrieval effectiveness is highly influenced by document collection characteristics and query properties, suggesting that the choice of retrieval approach should be adapted to the intended search context.

Keywords: *information retrieval; BM25; IndoSBERT; lexical retrieval; semantic retrieval*

1. Pendahuluan

Sistem pencarian dokumen akademik merupakan komponen penting dalam repositori kampus, terutama untuk membantu mahasiswa dan peneliti menemukan skripsi yang relevan berdasarkan topik pencarian tertentu. Sistem pencarian tradisional banyak menggunakan teknik *lexical matching* seperti BM25 dan TF-IDF, yang menghitung relevansi berdasarkan kemunculan kata kunci dalam dokumen

tanpa mempertimbangkan makna semantik dari teks tersebut. Pendekatan ini juga digunakan dalam penelitian sebelumnya pada konteks pencarian skripsi pada repository UPN "Veteran" Jawa Timur, yaitu implementasi algoritma *Weighted Tree Similarity* dan *Content-Based Filtering* yang berfokus pada pencocokan atribut dan struktur dokumen (Matondang et al., 2024). Studi literatur lain menunjukkan BM25 secara konsisten menjadi *baseline* yang kuat di banyak koleksi teks dan bahkan tetap kompetitif pada metrik penting seperti *precision* dan *recall* tanpa memerlukan pelatihan model besar (Farivar, 2025)

Perkembangan teknologi *deep learning* dan *pretrained language models* seperti BERT telah membuka arah baru dalam *semantic retrieval*, di mana *query* dan dokumen dikodekan ke dalam ruang vektor yang menangkap makna semantik di luar kata-kata yang sama. Pendekatan *deep learning* dilaporkan mampu menghasilkan representasi fitur yang lebih baik dibandingkan metode *machine learning* klasik (Mulyo & Widyantoro, 2018). Salah satu kajian awal tentang penggunaan representasi vektor ini adalah *Dense Passage Retrieval* (DPR), yang menunjukkan bahwa model *dense* yang dilatih dengan BERT dapat mengungguli BM25 pada tugas pencarian konteks terbuka seperti pertanyaan jawaban (*question answering*) (Karpukhin et al., 2020). Selain itu, teknik embedding seperti SBERT telah dirancang untuk menghasilkan *sentence embeddings* yang efisien dan efektif bagi tugas pencarian semantik dengan kemiripan kosinus (Reimers & Gurevych, 2019).

Namun, perbandingan antara metode *lexical* dan *semantic retrieval* sangat tergantung pada karakteristik koleksi dokumen dan sifat *query*. Misalnya, dalam sistem pencarian untuk teks medis, SBERT yang di-*finetune* menunjukkan akurasi yang lebih tinggi dibandingkan BM25 pada basis data tertentu (Fujishiro et al., 2023). Di sisi lain, beberapa penelitian lanjutan menemukan bahwa meskipun *dense retrievers* unggul dalam menangkap hubungan semantik dan konteks, mereka sering kalah dalam menangkap entitas atau frasa langka yang secara *lexical* penting bagi relevansi dokumen. Misalnya, pendekatan *Salient Phrase Aware Dense Retrieval* menunjukkan bahwa *dense retrievers* murni masih memiliki keterbatasan bila dibandingkan dengan *sparse retrievers* dalam menghadapi istilah yang penting namun jarang muncul (Chen et al., 2022).

Dalam konteks repositori akademik Indonesia, pengembangan sistem pencarian mulai diarahkan pada pemanfaatan model bahasa berbasis *transformer* untuk menangani keterbatasan pendekatan *lexical*. Salah satu penelitian menerapkan model BERT pra-latih pada tugas *academic expert finding* dan menunjukkan bahwa representasi kontekstual yang dihasilkan mampu meningkatkan kualitas pencocokan dibandingkan metode berbasis kata kunci (Mannix & Yulianti, 2024). Selain itu, penerapan pendekatan *embedding* berbasis *transformer* juga telah dieksplorasi untuk tugas klasifikasi dan rekomendasi pada repositori akademik. Sebuah studi tentang fine-tuning SBERT pada repositori trilingual menemukan bahwa model yang disesuaikan dengan taksonomi domain dapat meningkatkan performa *Precision@5* dan nDCG dalam rekomendasi penelitian dibandingkan model pra-latih tanpa *fine-tuning* (Rashid & Ahmed, 2025).

Di sisi lain, dalam konteks repositori kampus di Indonesia, sistem pencarian skripsi umumnya masih mengandalkan pendekatan *lexical* berbasis kata kunci seperti BM25, dengan evaluasi yang jarang melibatkan perbandingan langsung terhadap model *semantic retrieval* modern atau penggunaan *ground truth* relevansi yang dilabeli secara manual, seperti yang terlihat dari dominasi teknik *lexical* dalam karya-karya sebelumnya di domain ini (Satria et al., 2023). Penelitian literatur pada bidang *information retrieval* menunjukkan bahwa meskipun metode *dense retrieval* berbasis *embedding* terus berkembang, model lexical seperti BM25 masih diposisikan sebagai *baseline* yang kuat dan kompetitif, terutama ketika terdapat *overlap* istilah yang tinggi antara query dan dokumen, serta dalam koleksi dokumen yang repetitif atau terstruktur, sehingga seringkali BM25 dapat mengungguli model *dense* awal tanpa *fine-tuning* pada metrik-metrik standar retrieval (Lafayette et al., 2024). Di lain pihak, studi lain juga menegaskan bahwa hasil perbandingan antara metode *lexical* dan *semantic* memperlihatkan keunggulan model *semantic* yang sangat bergantung pada karakteristik domain dan jenis dokumen, serta tidak selalu secara konsisten mengungguli BM25 pada seluruh metrik evaluasi (Harris, 2025).

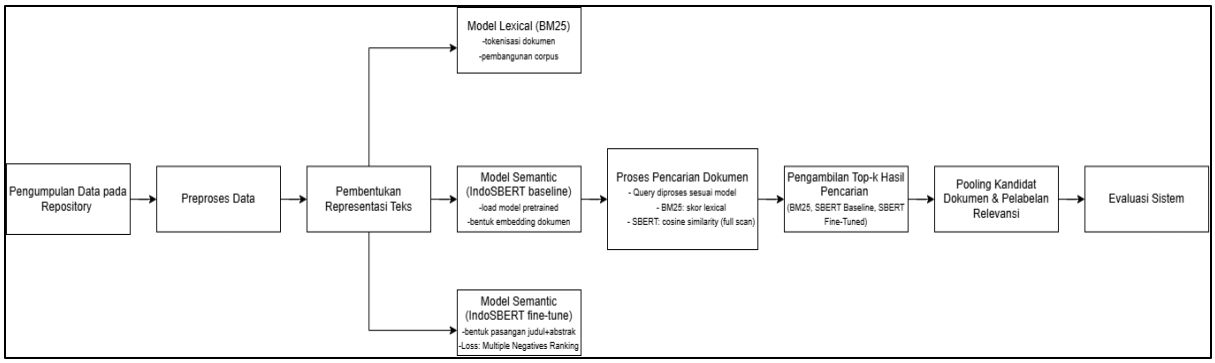
Berdasarkan celah penelitian tersebut, penelitian ini bertujuan untuk melakukan studi komparatif antara metode lexical BM25 dan metode *semantic retrieval* berbasis IndoSBERT baseline serta IndoSBERT *fine-tuned* pada repositori kampus UPN "Veteran" Jawa Timur. Evaluasi dilakukan menggunakan *ground truth* relevansi yang dilabeli secara manual dan metrik standar *information*

retrieval untuk menilai efektivitas masing-masing pendekatan secara objektif dalam mendukung pencarian dokumen akademik di lingkungan perguruan tinggi Indonesia.

2. Metode

2.1. Alur Umum Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan yang terstruktur, mulai dari pengumpulan data hingga evaluasi performa sistem pencarian. Alur umum penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.2. Sumber Data dan Ruang Lingkup

Objek penelitian berupa dokumen skripsi Program Studi Informatika yang diperoleh dari repositori Universitas Pembangunan Nasional "Veteran" Jawa Timur. Pengumpulan data dilakukan melalui proses *web scraping* setelah memperoleh izin resmi dari pengelola repositori. Dataset yang digunakan berjumlah sekitar 1.146 dokumen skripsi dengan rentang tahun publikasi 2019–2026.

Setiap dokumen terdiri atas atribut URL, judul, dan abstrak, yang digunakan sebagai sumber utama dalam pembentukan representasi dokumen. Contoh data dapat dilihat pada Tabel 1:

Tabel 1. Contoh Data dari Repository UPN Jatim

url	judul	abstrak
https://repository.upnjatim.ac.id/48797/	KLASIFIKASI TINGKAT KEPARAHAN KECELAKAAN LALU LINTAS BERBASIS CATBOOST PADA DATA YANG TIDAK SEIMBANG MENGGUNAKAN SMOTENC DAN OPTUNA	Traffic accidents represent a complex issue in Surabaya, having a significant impact on public safety and socio-economic loss. A primary challenge in accident severity classification modeling is the phenomenon of data imbalance, where slight injury cases are considerably more …
https://repository.upnjatim.ac.id/48873/	PERANCANGAN GAME LOKAL "ADIT'S DESTINY : A DARK JOURNEY"	This research provides a comprehensive overview of the Internship Training Program conducted at PT. Kinema Systrans Multimedia, specifically within the Infinite Learning division. Infinite Learning focuses on the development of vocational training courses relevant to ….

2.3. Pra Pemrosesan Teks

Tahap pra-pemrosesan dilakukan untuk memastikan konsistensi, kebersihan, dan keseragaman teks sebelum digunakan pada proses pencarian dokumen, baik oleh model *lexical* maupun *semantic retrieval*. Pra-pemrosesan ini bertujuan untuk mengurangi *noise* pada data teks serta meminimalkan variasi penulisan yang tidak relevan terhadap makna dokumen.

Tahapan pra-pemrosesan yang diterapkan meliputi normalisasi untuk penghapusan elemen non-informatif seperti emoji dan karakter khusus non-alfanumerik, serta *case folding* untuk mengonversi seluruh teks ke huruf kecil. Selanjutnya, dilakukan proses tokenisasi berbasis *regular expression* untuk mempertahankan istilah teknis dan numerik yang umum muncul pada dokumen skripsi bidang informatika. Mengingat model *semantic retrieval* yang digunakan telah dinormalisasi untuk bahasa Indonesia, beberapa abstrak yang ditulis dalam bahasa Inggris diterjemahkan terlebih dahulu ke dalam bahasa Indonesia sebelum tahapan pra-pemrosesan dilakukan. Penerjemahan ini dilakukan dengan formula pada spreadsheet *yaitu =GoogleTranslate([],"en","id")*. Langkah ini bertujuan untuk menjaga konsistensi bahasa antara dokumen dan model *embedding*, serta menghindari degradasi kualitas representasi semantik akibat perbedaan bahasa.

Untuk kebutuhan *semantic retrieval*, judul dan abstrak yang telah melalui proses normalisasi digabungkan menjadi satu representasi teks dokumen. Secara formal, representasi dokumen ke-i didefinisikan sebagai:

$$d_i = judul_i^{norm} + abstrak_i^{norm}$$

Representasi gabungan ini disimpan dalam kolom *text_semantic* dan digunakan secara konsisten sebagai masukan bagi model IndoSBERT, baik pada tahap pembentukan embedding maupun pada proses pencarian dokumen.

2.4. Metode Pencarian Dokumen

2.4.1. BM25 (*Lexical Retrieval*)

BM25 digunakan sebagai metode pencarian berbasis kata kunci (*lexical retrieval*) dan berperan sebagai *baseline* dalam penelitian ini. Pada pendekatan ini, dokumen dan *query* direpresentasikan sebagai kumpulan kata (token), kemudian relevansi dokumen terhadap *query* dihitung berdasarkan kemunculan dan distribusi kata kunci di dalam dokumen.

Secara umum, BM25 memberikan skor yang lebih tinggi pada dokumen yang mengandung kata kunci *query* lebih sering, namun tetap mempertimbangkan panjang dokumen agar dokumen yang terlalu panjang tidak memperoleh keuntungan berlebihan. Dengan mekanisme ini, BM25 mampu menyeimbangkan antara frekuensi istilah dan karakteristik dokumen secara sederhana namun efektif.

Pada penelitian ini, implementasi BM25 dilakukan menggunakan pustaka *rank_bm25*. Proses yang dilakukan meliputi tokenisasi sederhana berbasis spasi, konversi teks ke huruf kecil, serta penghapusan karakter non-alfanumerik. Model BM25 digunakan tanpa proses pelatihan tambahan, karena metode ini bersifat non-parametrik dan tidak memerlukan data latih.

2.4.2 IndoSBERT *Baseline* (*Semantic Retrieval*)

Pada pendekatan semantic retrieval, setiap dokumen skripsi direpresentasikan dalam bentuk vektor numerik menggunakan model *pretrained* IndoSBERT (*firqaaa/indo-sentence-bert-base*). Dokumen yang diproses berasal dari kolom *text_semantic*, yaitu gabungan judul dan abstrak yang telah melalui tahap pra-pemrosesan dan normalisasi teks.

Proses pembentukan embedding dilakukan dengan memasukkan teks dokumen ke dalam transformer encoder IndoSBERT. Pendekatan representasi teks berbasis vektor ini umum digunakan pada tahap ekstraksi fitur untuk merepresentasikan makna teks secara numerik sebelum diterapkan pada berbagai tugas pemrosesan bahasa alami (Putri et al., 2024). Representasi token yang dihasilkan kemudian diringkas menjadi satu vektor dokumen menggunakan *mean pooling*, sehingga setiap dokumen direpresentasikan oleh satu embedding berdimensi 768. Selanjutnya, embedding tersebut dinormalisasi menggunakan *L2 normalization*, sehingga setiap vektor memiliki panjang satu.

$$v\_doc \in \mathbb{R}^{768}$$

Normalisasi L2 diterapkan untuk memastikan bahwa perhitungan kemiripan antar vektor dilakukan secara konsisten menggunakan *cosine similarity*. Dengan kondisi ini, nilai *cosine similarity* antara query dan dokumen dapat dihitung secara langsung melalui operasi *dot product* antar vektor.

2.4.3 IndoSBERT *Fine-Tuned* (*Semantic Retrieval*)

Selain menggunakan model IndoSBERT *pretrained* sebagai *baseline*, penelitian ini juga menerapkan proses *fine-tuning* untuk menyesuaikan representasi *embedding* dengan karakteristik

dokumen skripsi pada repositori kampus. *Fine-tuning* dilakukan dengan tujuan meningkatkan kemampuan model dalam menangkap hubungan semantik yang spesifik terhadap domain akademik, khususnya pada konteks skripsi informatika.

Data fine-tuning disusun dalam bentuk pasangan judul–abstrak dari dokumen skripsi yang telah melalui tahap pra-pemrosesan. Pasangan ini digunakan untuk melatih model agar menghasilkan *embedding* judul dan abstrak yang saling berdekatan dalam ruang vektor apabila berasal dari dokumen yang sama. Pendekatan ini mencerminkan hubungan semantik alami antara judul dan abstrak sebagai representasi ringkas dan penjelas dari satu karya ilmiah.

Proses *fine-tuning* dilakukan menggunakan *Multiple Negatives Ranking Loss*, di mana setiap pasangan judul dan abstrak dianggap sebagai pasangan positif, sementara pasangan lain dalam satu batch diperlakukan sebagai contoh negatif secara implisit. Dengan skema ini, model dilatih untuk memaksimalkan kemiripan antara judul dan abstrak yang relevan, sekaligus membedakannya dari dokumen lain yang tidak berkaitan.

2.5. Proses Pencarian dan Pooling Dokumen

Untuk setiap *query*, sistem menghasilkan daftar dokumen teratas (*top-k*) dari ketiga metode pencarian. Hasil pencarian kemudian digabungkan menggunakan pendekatan *pooling* berbasis *union* untuk membentuk kumpulan kandidat dokumen untuk menghindari bias terhadap satu metode tertentu dan memastikan bahwa dokumen relevan dari berbagai pendekatan memiliki peluang untuk dievaluasi. Jadi untuk setiap model (BM25, IndoSBERT *Baseline*, dan IndoSBERT *Fine-Tune*) diambil top 15 kemudian diacak.

2.6. Pelabelan Relevansi dan *Ground Truth*

Pelabelan relevansi dilakukan secara manual terhadap dokumen hasil *pooling* dari ketiga sistem pencarian yang diuji untuk membangun *ground truth* (*qrels*) yang menjadi dasar evaluasi kuantitatif. *Pooling* dilakukan dengan menggabungkan dokumen teratas dari semua sistem sebagaimana praktik evaluasi *information retrieval* pada benchmark *modern*, di mana *top-ranked documents* dari setiap run dinilai oleh manusia untuk mengurangi beban penilaian terhadap keseluruhan koleksi sekaligus menghasilkan penilaian relevansi yang representatif (*pooling reduces the number of judgments needed while preserving evaluation quality*) (Soboroff, 2021). Setiap pasangan *query* dan dokumen kemudian dinilai menggunakan skema *graded relevance* tiga tingkat: 0 (tidak relevan), 1 (cukup relevan) dan 2 (sangat relevan), agar dapat mencerminkan variasi tingkat kesesuaian dokumen terhadap kebutuhan informasi pengguna dan mendukung metrik evaluasi seperti nDCG yang memperhitungkan perbedaan tingkat relevansi dokumen dalam daftar peringkat.

2.7. Evaluasi Kinerja Sistem

Evaluasi sistem temu kembali informasi dilakukan untuk menilai kemampuan sistem dalam mengembalikan dokumen yang relevan serta kualitas urutan hasil pencarian. Pada penelitian ini digunakan metrik *Precision@K*, *Recall@K*, *Mean Average Precision* (MAP), dan *normalized Discounted Cumulative Gain* (nDCG@K) yang umum dipakai untuk mengevaluasi sistem pencarian berbasis peringkat.

Precision@K mengukur proporsi dokumen relevan di antara K hasil teratas yang dikembalikan oleh sistem. Metrik ini menunjukkan ketepatan sistem dalam menampilkan dokumen relevan pada posisi awal hasil pencarian. *Recall@K* mengukur proporsi dokumen relevan yang berhasil ditemukan oleh sistem dibandingkan dengan seluruh dokumen relevan yang tersedia dalam koleksi. Metrik ini menilai kelengkapan hasil pencarian. MAP mengukur kualitas pemeringkatan dengan memperhitungkan posisi kemunculan dokumen relevan dalam daftar hasil. Nilai MAP diperoleh dari rata-rata *Average Precision* (AP) untuk seluruh query, sehingga mencerminkan kemampuan sistem menempatkan dokumen relevan pada peringkat lebih tinggi. nDCG@K mengevaluasi kualitas peringkat dengan mempertimbangkan tingkat relevansi dokumen dan posisinya dalam hasil pencarian. Dokumen dengan relevansi lebih tinggi dan posisi lebih awal memberikan kontribusi nilai yang lebih besar. Nilai nDCG berada pada rentang 0 hingga 2, di mana nilai yang lebih tinggi menunjukkan kualitas peringkat yang lebih baik (Jadon & Patil, 2024).

3. Hasil dan Pembahasan

3.1. Analisis Hasil Pencarian Berdasarkan Contoh *Query*

Untuk memberikan pemahaman yang lebih konkret mengenai perilaku masing-masing metode pencarian, dilakukan analisis kualitatif menggunakan satu contoh query, yaitu *"rancang bangun sistem informasi"*. Analisis difokuskan pada dokumen peringkat teratas (*top results*) yang dihasilkan oleh tiga pendekatan, yaitu IndoSBERT *baseline*, IndoSBERT *fine-tuned*, dan BM25. Perbandingan ini bertujuan untuk menunjukkan perbedaan karakteristik hasil pencarian secara nyata, baik dari sisi kesesuaian topik maupun pola pemeringkatan. Tabel 2 menunjukkan hasil uji coba pencarian dengan ketiga model:

Tabel 2. Contoh Hasil Pencarian

Model	Rank	Judul	Score
BM25	1	Pengembangan Sistem Informasi Berbasis Website Layanan Pengaduan Masyarakat Dinas Perumahan Rakyat, Kawasan Permukiman Dan Cipta Karya	11.535444
	2	Rancang Bangun Sistem Informasi Penggajian Berbasis Web Pada Pt. Bangun Jaya Power	11.444158
	3	Rancang Bangun Sistem Informasi Penggajian Pegawai Dengan Metode Prorata (Studi Kasus : Sma Wachid Hasyim Pusat Surabaya)	10.629533
IndoSBERT *Baseline*	1	Perancangan Ulang Desain Antarmuka Sistem Informasi Manajemen Pengelolaan Data Bpbd Kota Surabaya	0.622390
	2	Manajemen Proyek Agile Dengan Scrum : Study Kasus Proyek Pembangunan Sistem Dinas Komunikasi Dan Informatika Jawa Timur	0.590866
	3	Sistem Informasi Pengelolahan Data Berbasis Web Di Dinas Tenaga Kerja Dan Trasmigrasi Provinsi Jawa Timur	0.584534
IndoSBERT *Fine-Tune*	1	Simulasi Sistem Kendali Cerdas Berbasis Proportional Integral Derivative (Pid) Dengan Metode Deep Learning	0.310529
	2	Rancang Bangun Sistem Ujian Online Dan Implementasi Algoritma Lcm Dalam Pengacakan Soal Menggunakan Framework Codeigniter	0.309590
	3	Simulasi Sistem Kendali Cerdas Berbasis Proportional Integral Derivative (Pid) Dengan Metode Deep Learning	0.303299

Hasil pencarian BM25 menunjukkan pola yang konsisten, di mana dokumen pada peringkat teratas secara eksplisit mengandung frasa *"rancang bangun"* dan *"sistem informasi"*. Pola ini mencerminkan karakteristik pendekatan *lexical* yang bergantung pada kecocokan istilah dan frekuensi kata, sehingga efektif untuk *query* berbasis kata kunci, namun berpotensi melewatkan dokumen yang relevan secara semantik dengan istilah berbeda.

Sementara itu, Hasil pencarian IndoSBERT *baseline* menunjukkan bahwa dokumen bertopik sistem informasi dan rancang bangun cenderung muncul pada peringkat atas, meskipun tidak selalu mengandung frasa *query* secara eksplisit. Hal ini menandakan bahwa model mampu menangkap kesamaan makna secara semantik, bukan sekadar kecocokan kata. Namun, muncul pula dokumen dengan fokus yang lebih umum, seperti desain antarmuka atau optimasi proses bisnis, yang menunjukkan bahwa model baseline melakukan generalisasi semantik yang cukup luas terhadap dokumen yang masih satu tema.

Berbeda dengan model *baseline*, IndoSBERT *fine-tuned* menampilkan keragaman topik yang lebih tinggi pada peringkat awal, termasuk dokumen terkait *deep learning*, robotika, dan sistem kendali. Temuan ini mengindikasikan bahwa proses *fine-tuning* memperkuat kesamaan representasi antar dokumen secara umum, tetapi belum secara spesifik mengoptimalkan kesesuaian antara query pendek dan konteks dokumen. Nilai *cosine similarity* yang relatif lebih rendah juga menunjukkan adanya perubahan distribusi ruang embedding, yang umum terjadi ketika *fine-tuning* dilakukan tanpa supervisi relevansi berbasis *query*.

3.2. Hasil Evaluasi Kuantitatif

Evaluasi kuantitatif dilakukan menggunakan 10 query uji, dengan kandidat dokumen diperoleh melalui teknik *pooling* dari tiga sistem: BM25, IndoSBERT *baseline*, dan IndoSBERT *fine-tuned*. Dokumen kandidat kemudian dilabeli secara manual menggunakan skema relevansi diskret (0 = tidak relevan, 1 = relevan, 2 = sangat relevan).

Hasil evaluasi berdasarkan metrik Precision@5, Recall@5, MAP, dan nDCG@5 ditunjukkan pada Tabel 3 :

Tabel 3: Hasil Evaluasi Sistem

Metode	Precision@5	Recall@5	MAP	nDCG@5
BM25	0.8067	0.6963	0.6588	0.8866
IndoSBERT Baseline	0.4067	0.2780	0.2587	0.3759
IndoSBERT Fine-Tuned	0.3667	0.2912	0.3260	0.4097

Hasil evaluasi menunjukkan adanya perbedaan karakteristik yang jelas antara pendekatan *lexical* dan *semantic retrieval*. Nilai Precision@5 dan nDCG@5 yang relatif tinggi pada BM25 menunjukkan bahwa metode ini sangat efektif dalam menempatkan dokumen dengan kecocokan istilah yang kuat pada peringkat atas. Hal ini sejalan dengan temuan sebelumnya yang menyatakan bahwa BM25 tetap menjadi *baseline* yang kuat pada berbagai koleksi teks, terutama ketika query bersifat pendek dan eksplisit.

Sebaliknya, performa IndoSBERT *baseline* yang lebih rendah pada metrik presisi menunjukkan keterbatasan *dense retrieval* dalam menangkap istilah spesifik atau frasa penting yang jarang muncul. Model berbasis *embedding* cenderung melakukan generalisasi semantik, sehingga dokumen dengan topik serupa tetapi konteks berbeda masih dapat memperoleh skor kemiripan yang tinggi. Pada IndoSBERT *fine-tuned*, peningkatan terlihat pada metrik MAP dan nDCG dibandingkan model *baseline*. Hal ini menunjukkan bahwa *fine-tuning* membantu model dalam memperbaiki kualitas pemeringkatan dokumen relevan secara keseluruhan. Namun, *fine-tuning* yang dilakukan tanpa data relevansi berbasis *query* dapat menyebabkan model lebih fokus pada kesamaan representasi antar dokumen daripada kesesuaian langsung dengan query pengguna. Fenomena ini juga dilaporkan dalam penelitian yang menunjukkan bahwa dalam *dense retrieval* pengetahuan utama yang mendukung efektivitas retrieval sebagian besar diperoleh dari tahap *pre-training*, dan *fine-tuning* banyak menyesuaikan aktivasi internal model daripada secara signifikan mengoptimalkan kemampuan *retrieval* terhadap relevansi query-dokumen (Yao et al., 2025).

Selain itu, koleksi skripsi yang digunakan dalam penelitian ini umumnya menggunakan bahasa formal dan istilah yang relatif konsisten. Dalam kondisi seperti ini, metode pencarian berbasis kata kunci (*lexical*) cenderung bekerja dengan baik karena istilah penting sering muncul secara langsung pada judul dan abstrak dokumen. Sebaliknya, keunggulan metode pencarian semantik biasanya lebih terlihat pada koleksi dengan variasi bahasa yang tinggi atau pada *query* yang bersifat konseptual dan tidak menyebutkan istilah secara eksplisit.

4. Kesimpulan

Penelitian ini membandingkan metode pencarian berbasis *lexical* dan *semantic* pada repositori UPN "Veteran" Jawa Timur menggunakan BM25, IndoSBERT *baseline*, dan IndoSBERT *fine-tuned*. Hasil evaluasi menunjukkan bahwa BM25 memiliki performa terbaik pada metrik evaluasi yang menandakan kemampuannya dalam menempatkan dokumen dengan kecocokan istilah yang kuat pada peringkat teratas. Hal ini menunjukkan bahwa pendekatan *lexical* masih sangat efektif untuk koleksi dokumen akademik dengan istilah formal dan konsisten serta query yang bersifat pendek dan eksplisit.

Pendekatan *semantic retrieval* berbasis IndoSBERT mampu menangkap kesamaan makna antar dokumen meskipun tidak selalu menggunakan istilah yang sama. Namun, model IndoSBERT *baseline* cenderung melakukan generalisasi semantik yang luas, sehingga dokumen dengan konteks berbeda tetapi masih satu tema dapat memperoleh skor kemiripan tinggi. Proses *fine-tuning* meningkatkan kualitas pemeringkatan secara keseluruhan, meskipun belum mampu secara konsisten melampaui performa BM25 karena tidak menggunakan supervisi relevansi berbasis query.

Secara keseluruhan, hasil penelitian ini menegaskan bahwa tidak terdapat satu metode pencarian yang secara mutlak unggul pada seluruh kondisi. Metode *lexical* seperti BM25 tetap menjadi *baseline* yang kuat untuk repositori akademik dengan karakteristik bahasa yang terstruktur, sementara pendekatan *semantic retrieval* memiliki potensi untuk melengkapi sistem pencarian, terutama pada skenario *query* yang bersifat konseptual. Temuan ini membuka peluang penelitian lanjutan untuk

mengembangkan pendekatan hibrida atau *fine-tuning* berbasis data relevansi *query* dan dokumen guna meningkatkan kinerja sistem pencarian skripsi di lingkungan perguruan tinggi.

5. Ucapan terimakasih

Penulis menyampaikan terima kasih kepada dosen pembimbing dari Program Studi Informatika UPN "Veteran" Jawa Timur yang telah memberikan keilmuan dan wawasan selama masa studi. Ucapan terima kasih turut disampaikan kepada pengelola Repositori Universitas Pembangunan Nasional "Veteran" Jawa Timur atas izin, bantuan, dan dukungan dalam penyediaan data skripsi yang digunakan pada penelitian ini. Selain itu, penulis mengapresiasi pihak-pihak yang terlibat dalam proses penelitian ini..

Daftar Pustaka

Chen, X., Lakhotia, K., Oguz, B., Gupta, A., Lewis, P., Peshterliev, S., Mehdad, Y., Gupta, S., & Yih, W. (2022). Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 250–262). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.19

Farivar, K. (2025). Semantic Search for Information Retrieval. *ArXiv Preprint ArXiv:2508.17694*.

Fujishiro, N., Otaki, Y., & Kawachi, S. (2023). applied sciences Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims. *Applied Sciences 13(6) 4051*. https://doi.org/https://doi.org/10.3390/app13064051

Harris, L. (2025). *Comparing Lexical and Semantic Vector Search Methods When Classifying Medical Documents*.

Jadon, A., & Patil, A. (2024). *A Comprehensive Survey of Evaluation*. 1–25.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Lafayette, W., Mori, L., Lafayette, W., & Ventresca, M. (2024). *Assessing the Performance Gap Between Lexical and Semantic Models for Information Retrieval With Formulaic Legal Language*.

Mannix, I. A., & Yulianti, E. (2024). Academic expert finding using BERT pre-trained language model. *International Journal of Advances in Intelligent Informatics*, *10*(2), 280–295. https://doi.org/10.26555/ijain.v10i2.1497

Matondang, N., Via, Y. V., & Akbar, F. A. (2024). *IMPLEMENTASI ALGORITMA WEIGHTED TREE SIMILARITY DAN CONTENT BASED FILTERING DALAM*. *12*(3).

Mulyo, B. M., & Widyantoro, D. H. (2018). Aspect-Based Sentiment Analysis Approach with CNN. *Proceeding of EECSI*, 16–18.

Putri, D. R., Puspaningrum, E. Y., Maulana, H., Pembangunan, U., Veteran, N., & Timur, J. (2024). KLASIFIKASI SENTIMEN TENTANG PEMINDAHAN IBU KOTA NEGARA INDONESIA DENGAN CONVOLUTIONAL NEURAL NETWORK MENGGUNAKAN GLOVE DAN FASTTEXT. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)*, *12*(3), 2759–2769.

Rashid, H. W., & Ahmed, S. H. (2025). Fine-tuning SBERT for Semantic Research Title Classification in Trilingual University Repository. *Kurdistan Journal of Applied Research*, *10*(2), 119–135. https://doi.org/10.24017/science.2025.2.9

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Satria, R. A., Brawijaya, U., & Korespondensi, P. (2023). *PENGELOMPOKAN HASIL PENCARIAN SKRIPSI BERBAHASA INDONESIA MENGGUNAKAN METODE DBSCAN DENGAN PEMBOBOTAN BM25 THE CLUSTERING OF THESIS SEARCH RESULTS IN INDONESIAN USING THE DBSCAN METHOD WITH BM25 WEIGHTING*. *10*(4), 781–790. https://doi.org/10.25126/jtiik.2023106899

Soboroff, I. (2021). *Overview of TREC 2021*. 1–17.

Yao, Z., Wang, S., & Zuccon, G. (2025). *Pre-training vs. Fine-tuning: A Reproducibility Study on Dense Retrieval Knowledge Acquisition*. 1–22.