**TECHNOVATAR Jurnal Teknologi, Industri, dan Informasi**
Vol. 4, No. 3, Juli 2026, pp. 1 – 11
**Open Access**:
https://journal.awatarapublisher.com/index.php/technovatar
**ISSN: 3031-8874, DOI: 10.61434/technovatar.v4i3.403**

# SISTEM REKOMENDASI TOPIK TUGAS AKHIR MAHASISWA S1 INFORMATIKA BERBASIS CONTENT-BASED FILTERING: STUDI SIMULASI EVALUASI METODE LEKSIKAL DAN SEMANTIK LATEN

**Nur Ariesanto Ramdhan**[1*]
1 Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhadi Setiabudi, Brebes, Jawa Tengah, Indonesia
Email: ariesantoramdhan@gmail.com

**INFO ARTIKEL**
Sejarah Artikel:
Diterima, 11 Juni 2026
Direvisi, Juni 2026
Disetujui, 2 Juli 2026

**KEYWORDS**
Recommender system
Content-based filtering
Thesis topic
BM25
Latent semantic analysis

***ABSTRACT***
*Choosing a final project topic is an early decision that shapes a student's research direction. This study builds and evaluates a prototype recommender for undergraduate Informatics final project topics based on content-based filtering, using each student's interests and course track record as the query. Four methods are compared, namely TF-IDF with cosine similarity, BM25, Latent Semantic Analysis (LSA), and a hybrid of BM25 and LSA, together with a random baseline. The offline evaluation uses 20 synthetic corpora, each with 400 topics from eight Informatics areas and 150 student profiles, and reports Precision@k, MRR, and NDCG@10. Because real data were unavailable and Sentence-BERT could not be downloaded in the research environment, this is a simulation study and does not evaluate Sentence-BERT; its evaluation script is provided. BM25 obtained the highest Precision@5 (0.927), followed by the hybrid (0.914), TF-IDF (0.895), LSA (0.876), and random (0.209); all pairwise differences in Precision@5 remained significant after Holm correction. The hypothesis that latent semantic methods are more robust to term variation was not supported, because LSA declined the most as term variation increased. These findings are preliminary and bounded by the assumptions of the data generator; validation on real final project repositories with expert relevance judgments is still required.*

*This work is licensed under a Creative Commons Attribution 4.0 International License.*

**KATA KUNCI**
Sistem rekomendasi
Content-based filtering
Topik tugas akhir
BM25
Latent semantic analysis

**CORRESPONDING AUTHOR**
Nur Ariesanto Ramdhan
Program Studi Teknik Informatika,
Universitas Muhadi Setiabudi
Jl. Pangeran Diponegoro KM 2,
Pesantunan, Wanasari, Brebes 52252, Jawa Tengah, Indonesia
Email: ariesantoramdhan@gmail.com

**ABSTRAK**
Pemilihan topik tugas akhir merupakan keputusan awal yang menentukan arah penelitian mahasiswa. Penelitian ini membangun dan mengevaluasi prototipe sistem rekomendasi topik tugas akhir mahasiswa S1 Informatika berbasis content-based filtering yang memakai minat dan rekam jejak mata kuliah mahasiswa sebagai kueri. Empat metode dibandingkan, yaitu TF-IDF dengan cosine similarity, BM25, Latent Semantic Analysis (LSA), dan hybrid BM25 dengan LSA, bersama pembanding acak. Evaluasi luring memakai 20 korpus sintetis, masing-masing berisi 400 topik dari delapan bidang Informatika dan 150 profil mahasiswa, dengan metrik Precision@k, MRR, dan NDCG@10. Karena data nyata tidak tersedia dan model Sentence-BERT tidak dapat diunduh pada lingkungan penelitian, penelitian ini adalah studi simulasi dan tidak mengevaluasi Sentence-BERT; skrip evaluasinya disediakan. BM25 memperoleh Precision@5 tertinggi (0,927), diikuti hybrid (0,914), TF-IDF (0,895), LSA (0,876), dan acak (0,209); seluruh selisih Precision@5 antarmetode tetap signifikan setelah koreksi Holm. Dugaan bahwa metode semantik laten lebih tahan terhadap variasi istilah tidak terdukung, karena LSA justru menurun paling tajam ketika variasi istilah meningkat. Hasil ini bersifat awal dan terbatas pada asumsi generator data; validasi pada repositori tugas akhir nyata dengan penilaian relevansi oleh pakar masih diperlukan.

## PENDAHULUAN

Pemilihan topik tugas akhir merupakan keputusan awal yang menentukan arah penelitian mahasiswa. Sejumlah penelitian membangun sistem rekomendasi untuk membantu keputusan ini, baik berupa rekomendasi topik tugas akhir (Kusuma dan Musdholifah, 2021) maupun rekomendasi dosen pembimbing berdasarkan kemiripan teks (Damayanti et al., 2025; Hairani dan Mujahid, 2022; Syabilla et al., 2025). Tinjauan sistematis tentang sistem rekomendasi pendidikan (Urdaneta-Ponte et al., 2021) dan pemilihan mata kuliah (Algarni dan Sheldon, 2023) menunjukkan bahwa topik ini aktif diteliti, tetapi pada pemilihan mata kuliah pendekatan berbasis konten masih jarang dipakai. Algarni dan Sheldon (2023) menemukan hanya 2 dari 35 studi yang berbasis konten, sebagian besar studi memakai data privat (hanya 3 yang memakai data publik), dan masalah cold start serta keragaman hasil yang terbatas masih menjadi tantangan.

Content-based filtering mencocokkan profil pengguna dengan isi item tanpa memerlukan riwayat interaksi pengguna lain, sehingga cocok untuk mahasiswa baru yang belum memiliki riwayat pemilihan topik. Representasi teks yang umum adalah TF-IDF dengan cosine similarity, seperti pada penelitian rekomendasi dosen pembimbing berbahasa Indonesia (Damayanti et al., 2025). Pendekatan yang lebih baru memakai embedding kalimat. Sentence-BERT (SBERT) memetakan kalimat ke vektor yang dapat dibandingkan dengan cosine similarity dan menurunkan waktu pencarian pasangan paling mirip dari sekitar 65 jam menjadi sekitar 5 detik pada 10.000 kalimat (Reimers dan Gurevych, 2019). Model multibahasa dilatih dengan distilasi pengetahuan (Reimers dan Gurevych, 2020), dan keluarga model lain seperti LaBSE (Feng et al., 2022) serta multilingual-E5 (Wang et al., 2024) tersedia untuk puluhan bahasa. Pada rekomendasi publikasi ilmiah, encoder kalimat msmarco-distilbert-base-tas-b memberikan hasil terbaik pada tahap rekomendasi (Guesmi et al., 2023).

Bukti keunggulan embedding atas metode leksikal belum seragam. Nabila et al. (2025) melaporkan bahwa pendekatan BERT dan Word2Vec mencapai hit ratio 1,00 pada rekomendasi dosen pembimbing dan lebih baik daripada TF-IDF serta bag of words. Sebaliknya, Saputra et al. (2026) melaporkan pada Wikipedia berbahasa Indonesia bahwa BM25 (P@10 = 0,973; MRR = 0,917) jauh mengungguli model paraphrase-multilingual-MiniLM-L12-v2 (P@10 = 0,325; MRR = 0,243), dan Thakur et al. (2021) menunjukkan bahwa BM25 merupakan pembanding yang kuat pada 18 dataset pengujian temu kembali. Selain itu, protokol evaluasi pada studi rekomendasi topik dan pembimbing tidak seragam: akurasi, hit ratio, F1, atau skor tujuan dipakai pada sampel kecil, sedangkan kerangka evaluasi sistem rekomendasi menekankan bahwa evaluasi yang memadai memerlukan beberapa perspektif (Zangerle dan Bauer, 2022) dan bahwa kebocoran data dapat mengubah hasil evaluasi luring (Ji et al., 2023). Berdasarkan literatur yang berhasil diverifikasi dalam penelitian ini, belum ditemukan studi rekomendasi topik tugas akhir S1 Informatika berbahasa Indonesia yang membandingkan metode berbasis konten secara terkendali dengan metrik berbasis peringkat.

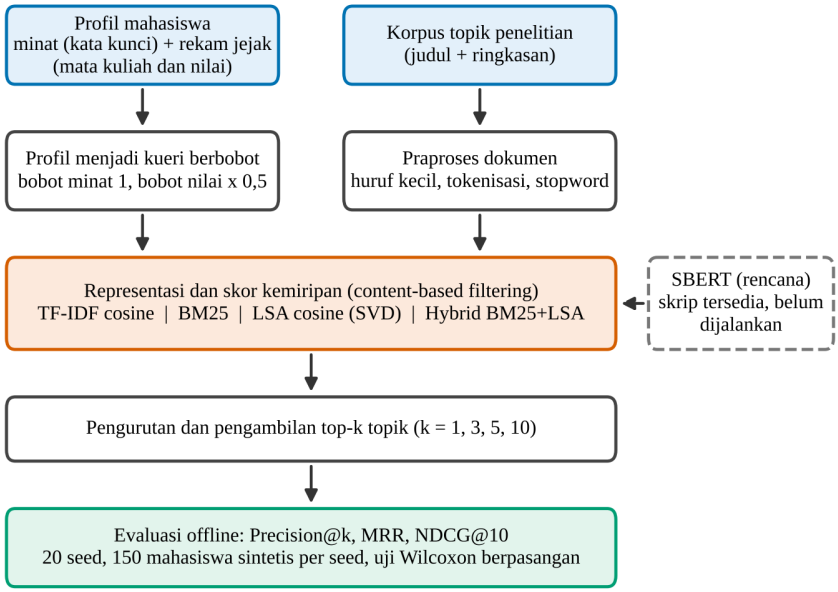
Penelitian ini bertujuan (1) membangun prototipe sistem rekomendasi topik tugas akhir berbasis content-based filtering yang memakai minat dan rekam jejak mata kuliah mahasiswa sebagai kueri, (2) membandingkan TF-IDF, BM25, Latent Semantic Analysis (LSA), dan hybrid BM25 dengan LSA memakai Precision@k, MRR, dan NDCG@10, serta (3) menguji sensitivitas hasil terhadap variasi istilah, kebisingan minat, dan dimensi LSA. Pertanyaan penelitiannya adalah: RQ1, seberapa baik masing-masing metode merekomendasikan topik yang sesuai dengan profil mahasiswa; RQ2, bagaimana kontribusi minat dan rekam jejak mata kuliah terhadap kualitas rekomendasi; dan RQ3, bagaimana sensitivitas metode terhadap variasi leksikal, kebisingan minat, dan dimensi LSA. Hipotesis kerja yang diuji adalah bahwa metode semantik laten (LSA dan hybrid) lebih tahan terhadap variasi istilah daripada metode leksikal.

Ruang lingkup penelitian perlu dinyatakan sejak awal. Data tugas akhir nyata yang dapat dipakai secara terbuka tidak tersedia pada penelitian ini, dan model SBERT terlatih tidak dapat diunduh pada lingkungan komputasi yang dipakai karena akses ke repositori model diblokir. Oleh karena itu, penelitian ini merupakan studi simulasi pada korpus dan profil mahasiswa sintetis, serta tidak mengevaluasi SBERT. Skrip evaluasi SBERT yang memakai protokol dan metrik yang sama disediakan sebagai paket kode pendamping dan belum dijalankan. Kontribusi penelitian ini adalah perancangan protokol benchmarking luring yang terkontrol dan dapat direproduksi, pembuktian ketangguhan metode leksikal BM25 sebagai garis dasar (baseline) komputasi efisien, serta identifikasi batas kinerja LSA yang harus dilampaui oleh model embedding kalimat mutakhir pada repositori naskah akademik riil.

## METODE PENELITIAN

### Desain Penelitian

Penelitian ini adalah eksperimen komputasi luring (offline) dengan pembangkitan data simulasi. Alur sistem ditunjukkan pada Gambar 1. Profil mahasiswa (minat dan rekam jejak mata kuliah) diubah menjadi kueri berbobot, korpus topik diproses menjadi representasi vektor, setiap topik diberi skor kemiripan terhadap kueri, dan k topik teratas dievaluasi terhadap label relevansi.




**Gambar 1. Alur sistem rekomendasi topik tugas akhir dan protokol evaluasi. Kotak putus-putus menandai perluasan SBERT yang belum dijalankan.**

### Pembangkitan Korpus dan Profil Sintetis

Korpus mencakup delapan bidang Informatika: kecerdasan buatan dan pembelajaran mesin, visi komputer, pemrosesan bahasa alami, rekayasa perangkat lunak, sistem informasi, jaringan komputer dan Internet of Things, keamanan siber, serta interaksi manusia-komputer dan multimedia. Setiap bidang memiliki 10 konsep, dan setiap konsep memiliki satu sampai tiga bentuk permukaan, misalnya "jaringan saraf tiruan", "neural network", dan "artificial neural network", untuk meniru campuran istilah Indonesia dan Inggris pada judul skripsi. Setiap bidang memiliki lima subtopik dan setiap subtopik memiliki 10 topik, sehingga korpus berisi 400 topik. Topik terdiri atas judul dan satu kalimat ringkasan yang memuat dua konsep inti subtopik dan, dengan peluang 0,5, satu konsep tambahan dari bidang yang sama, satu domain penerapan dari 12 domain bersama (misalnya

kesehatan, pendidikan, UMKM), dan kata umum seperti "rancang bangun" atau "implementasi". Sebanyak 20 persen topik bersifat penghubung (memuat satu konsep dari bidang lain) dan diberi label dua bidang. Parameter lengkap tercantum pada Tabel 1.

**Tabel 1. Parameter simulasi pada pengaturan utama**

| Parameter | Nilai |
|---|---|
| Bidang, subtopik per bidang, topik per subtopik | 8, 5, 10 (total 400 topik) |
| Konsep per bidang, bentuk permukaan per konsep | 10, 1 sampai 3 |
| Peluang dokumen memakai bentuk non-baku | 0,4 |
| Peluang kueri memakai bentuk non-baku (variasi leksikal) | 0,5 (divariasikan 0 sampai 0,8) |
| Proporsi kata kunci minat di luar bidang utama (kebisingan) | 0,3 (divariasikan 0 sampai 0,6) |
| Proporsi topik penghubung dua bidang | 0,2 |
| Mahasiswa per seed, jumlah seed | 150, 20 |
| Peluang mahasiswa memiliki bidang sekunder | 0,4 |
| Kata kunci minat per mahasiswa | 3 bidang utama, 1 sekunder (bila ada), 1 domain |
| Mata kuliah per mahasiswa (rata-rata) | sekitar 11, nilai A sampai D |
| Bobot istilah nilai A, B, C, D | 1,0; 0,75; 0,4; 0,1 (dikalikan 0,5) |
| Dimensi LSA | 50 (divariasikan 10 sampai 200) |

Setiap mahasiswa sintetis memiliki satu bidang utama dan, dengan peluang 0,4, satu bidang sekunder. Kueri minat berisi tiga kata kunci bidang utama (masing-masing dapat berasal dari bidang lain dengan peluang kebisingan), satu kata kunci bidang sekunder bila ada, dan satu domain penerapan. Rekam jejak berupa daftar mata kuliah beserta nilai: 4 sampai 6 mata kuliah bidang utama dengan nilai cenderung tinggi, 1 sampai 2 mata kuliah bidang sekunder, 2 sampai 3 mata kuliah bidang lain, dan tiga mata kuliah umum (algoritma dan pemrograman, struktur data, matematika diskrit). Label relevansi ditetapkan dari bidang: topik bernilai relevansi 2 bila memuat bidang utama mahasiswa, bernilai 1 bila hanya memuat bidang sekunder, dan 0 untuk lainnya. Pada rata-rata, 20,4 persen topik relevan bagi seorang mahasiswa (seed 0), sehingga Precision@k acak bernilai sekitar 0,2. Setiap seed membangkitkan korpus dan mahasiswa baru, sehingga 20 seed adalah 20 dataset independen.

Asumsi peluang bentuk non-baku (0,4 pada dokumen dan 0,5 pada kueri) dirancang secara terukur untuk mengaproksimasi campuran istilah dwibahasa (Indonesia-Inggris) yang lazim ditemukan pada judul skripsi mahasiswa Informatika. Parameter pengaturan utama ditetapkan sebelum seluruh eksperimen dijalankan. Uji awal pada sampel kecil dengan variasi leksikal 0,2 dan kebisingan 0,15 menunjukkan skor sempurna pada semua metode non-acak (efek langit-langit), sehingga tidak dapat membedakan metode; parameter kemudian dinaikkan ke nilai pada Tabel 1. Pemilihan ini bertujuan menghindari efek langit-langit dan tidak disesuaikan untuk menguntungkan metode tertentu. Korpus dan profil seed 0 disimpan sebagai berkas CSV dan JSON pada paket kode pendamping.

**Representasi dan Skor Kemiripan**

Teks diubah menjadi huruf kecil, dipecah dengan ekspresi reguler menjadi token, dan 16 kata fungsi dibuang. Tidak dilakukan stemming. Kueri berbobot dibentuk dari token minat (bobot 1) dan token mata kuliah (bobot nilai dikali 0,5). Lima metode dibandingkan. (1) TF-IDF: frekuensi istilah sublinear $1 + \ln(tf)$, $idf = \ln((1 + N) / (1 + df)) + 1$, vektor dinormalkan L2, dan skor adalah cosine similarity, yaitu $\cos(q, d) = (q \cdot d) / (\|q\| \|d\|)$. (2) BM25 dengan $k1 = 1{,}5$, $b = 0{,}75$, dan $idf = \ln(1 + (N - df + 0{,}5) / (df + 0{,}5))$; bobot kueri dikalikan skor istilah. (3) LSA: TruncatedSVD (scikit-learn) pada matriks TF-IDF dengan 50 komponen utama (ditentukan berdasarkan hasil analisis sensitivitas dimensi E5), kueri diproyeksikan ke ruang semantik laten yang sama, lalu dihitung cosine similarity. Stemming tidak diterapkan guna mempertahankan integritas istilah teknis komputasi. (4) Hybrid: 0,5 kali skor z BM25 ditambah 0,5 kali skor z LSA per mahasiswa. (5) Acak sebagai batas bawah.

SBERT direncanakan sebagai metode keenam dengan model paraphrase-multilingual-MiniLM-L12-v2, yang menurut kartu model Hugging Face menghasilkan vektor 384 dimensi dengan panjang masukan maksimum 128 token dan mendukung 50 bahasa. Karena batas 128 token, masukan yang panjang akan terpotong. Skrip sbert_extension.py pada paket kode menjalankan model ini pada dataset yang sama dan menghitung metrik yang sama, tetapi tidak dijalankan pada penelitian ini.

**Evaluasi dan Analisis Statistik**

Metrik yang dipakai adalah Precision@k = (jumlah topik relevan pada k teratas) / k untuk k = 1, 3, 5, dan 10 dengan relevansi biner (nilai relevansi minimal 1); Mean Reciprocal Rank (MRR) = rata-rata 1 / (peringkat topik relevan pertama); dan NDCG@10 dengan gain 2^rel - 1 dan diskon 1 / log2(peringkat + 1) terhadap peringkat ideal. Skor seri diurutkan acak. Recall@k tidak dilaporkan karena jumlah topik relevan per mahasiswa (sekitar 80 dari 400) jauh melebihi k. Setiap seed menghasilkan rata-rata metrik atas 150 mahasiswa. Hasil dilaporkan sebagai rata-rata atas 20 seed dengan interval kepercayaan 95 persen berbasis distribusi t. Perbandingan antarmetode memakai uji Wilcoxon berpasangan pada tingkat seed (scipy), dengan koreksi Holm atas enam perbandingan per metrik.

Lima eksperimen dijalankan: (E1) perbandingan utama dengan masukan minat dan rekam jejak; (E2) ablasi masukan (minat saja, rekam jejak saja, gabungan); (E3) variasi leksikal kueri 0, 0,2, 0,4, 0,6, dan 0,8; (E4) kebisingan minat 0, 0,15, 0,3, 0,45, dan 0,6; serta (E5) dimensi LSA 10, 25, 50, 100, dan 200. Seluruh kode ditulis dengan Python 3.13, scikit-learn 1.9.1, NumPy 2.5.3, SciPy 1.18.1, dan pandas 3.0.5.

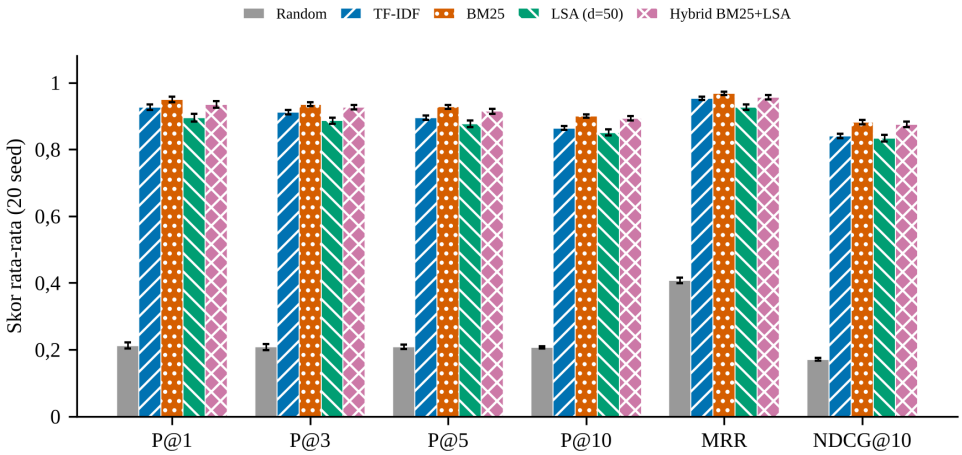
## HASIL DAN PEMBAHASAN

**Perbandingan Metode (RQ1)**

**Tabel 2. Rata-rata metrik atas 20 seed pada pengaturan utama (minat dan rekam jejak); NDCG diukur pada k = 10**

| Metode | P@1 | P@3 | P@5 | P@10 | MRR | NDCG |
|---|---|---|---|---|---|---|
| Acak | 0,212 | 0,208 | 0,209 | 0,207 | 0,408 | 0,171 |
| TF-IDF | 0,926 | 0,911 | 0,895 | 0,864 | 0,953 | 0,840 |
| BM25 | 0,949 | 0,935 | 0,927 | 0,900 | 0,968 | 0,882 |
| LSA | 0,895 | 0,886 | 0,876 | 0,851 | 0,927 | 0,833 |
| Hybrid | 0,934 | 0,926 | 0,914 | 0,893 | 0,957 | 0,875 |

Hasil pada Tabel 2 dan Gambar 2 menunjukkan bahwa semua metode berbasis konten jauh melampaui acak (Precision@5 sekitar 0,21). BM25 memperoleh skor tertinggi pada seluruh metrik (Precision@5 = 0,927, NDCG@10 = 0,882), diikuti hybrid (0,914), TF-IDF (0,895), dan LSA (0,876). Simpangan baku antarseed berkisar 0,012 sampai 0,020 untuk keempat metode tersebut pada Precision@5.




**Gambar 2. Perbandingan metrik antarmetode (rata-rata 20 seed, batang galat interval kepercayaan 95 persen).**

**Tabel 3. Uji berpasangan pada Precision@5 (20 seed)**

| Pembandingan | Selisih rata-rata | CI 95 persen | p Wilcoxon | p Holm |
|---|---|---|---|---|
| BM25 - TF-IDF | 0,032 | 0,028; 0,036 | < 0,001 | < 0,001 |
| BM25 - LSA | 0,051 | 0,045; 0,057 | < 0,001 | < 0,001 |
| TF-IDF - LSA | 0,019 | 0,012; 0,025 | < 0,001 | < 0,001 |
| Hybrid - BM25 | -0,013 | -0,017; -0,010 | < 0,001 | < 0,001 |
| Hybrid - LSA | 0,038 | 0,033; 0,043 | < 0,001 | < 0,001 |
| Hybrid - TF-IDF | 0,019 | 0,015; 0,023 | < 0,001 | < 0,001 |

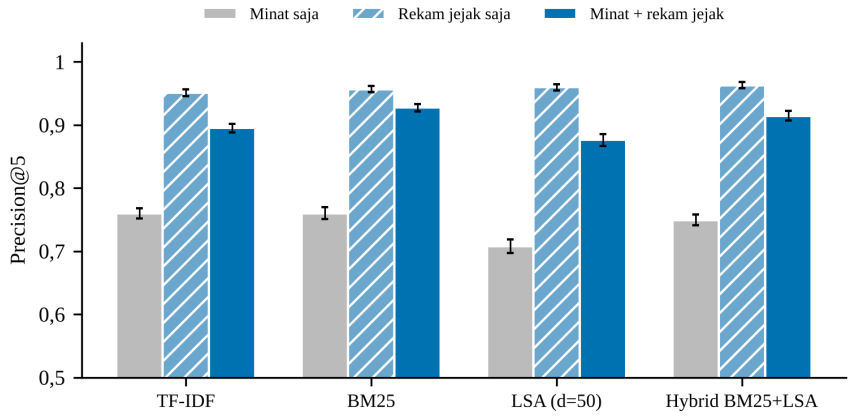
Seluruh enam selisih pada Precision@5 signifikan setelah koreksi Holm (Tabel 3). BM25 unggul atas TF-IDF sebesar 0,032 dan atas LSA sebesar 0,051, sedangkan hybrid berada di bawah BM25 sebesar 0,013. Pola yang sama muncul pada NDCG@10 (seluruh p Holm di bawah 0,05; selisih TF-IDF dan LSA paling kecil, 0,007, p Holm = 0,030). Selisih absolut berukuran kecil (kurang dari 0,06) dan hanya berlaku pada data simulasi ini.

**Peran Minat dan Rekam Jejak (RQ2)**

**Tabel 4. Precision@5 menurut masukan profil (rata-rata 20 seed)**

| Metode | Minat saja | Rekam jejak saja | Minat + rekam jejak |
|---|---|---|---|
| TF-IDF | 0,760 | 0,951 | 0,895 |
| BM25 | 0,760 | 0,957 | 0,927 |
| LSA | 0,708 | 0,960 | 0,876 |
| Hybrid | 0,749 | 0,963 | 0,914 |

Pada Tabel 4, masukan minat saja menghasilkan Precision@5 sebesar 0,760 (BM25), sedangkan rekam jejak saja menghasilkan 0,957 dan gabungan 0,927. Penambahan rekam jejak meningkatkan kualitas rekomendasi dibandingkan minat saja pada semua metode (Gambar 3). Namun gabungan tidak melampaui rekam jejak saja. Hasil ini adalah konsekuensi asumsi generator: kata kunci minat dibangkitkan dengan kebisingan 0,3, sedangkan mata kuliah bidang utama dibangkitkan lebih bersih dan lebih banyak (sekitar 4 sampai 6 mata kuliah dibanding 3 kata kunci), serta bobot rekam jejak 0,5 tidak disetel. Hasil ini tidak dapat dibaca sebagai bukti bahwa rekam jejak lebih informatif daripada minat pada mahasiswa nyata.

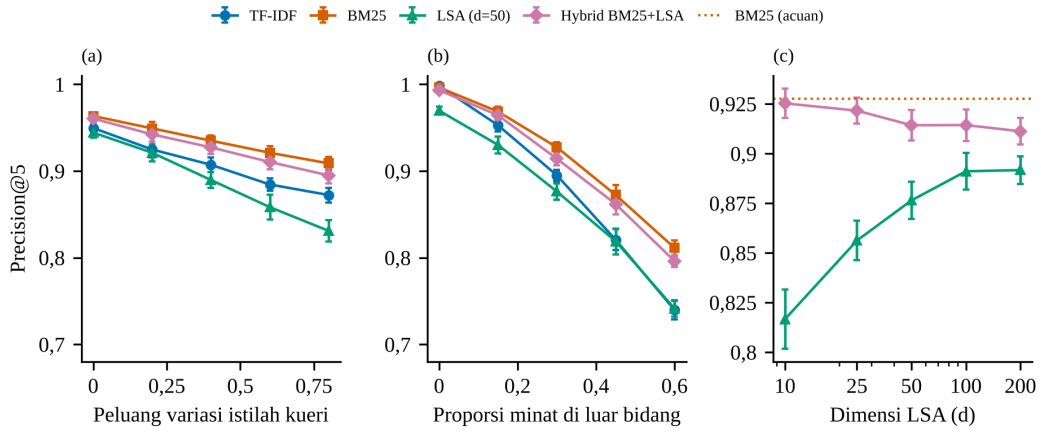



**Gambar 3. Precision@5 menurut masukan profil dan metode (rata-rata 20 seed, batang galat interval kepercayaan 95 persen).**

### Sensitivitas terhadap Variasi Istilah, Kebisingan, dan Dimensi LSA (RQ3)

Tiga analisis sensitivitas diringkas pada Gambar 4. Pada panel (a), peningkatan variasi istilah kueri dari 0 ke 0,8 menurunkan Precision@5 BM25 dari 0,963 ke 0,909, TF-IDF dari 0,950 ke 0,872, dan LSA dari 0,944 ke 0,831. Penurunan LSA (sekitar 0,113) paling besar, dua kali penurunan BM25 (sekitar 0,055). Dengan demikian hipotesis bahwa metode semantik laten lebih tahan terhadap variasi istilah tidak didukung pada data ini; arah hasilnya justru berlawanan.

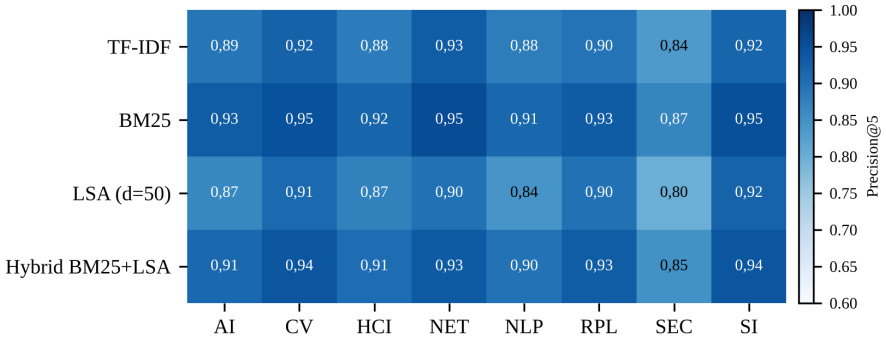
Pada panel (b), kebisingan minat menurunkan semua metode. Dari kebisingan 0 ke 0,6, Precision@5 BM25 turun dari 0,996 ke 0,811, TF-IDF dari 0,998 ke 0,740, dan LSA dari 0,970 ke 0,741. Pada kebisingan tinggi (0,6), LSA dan TF-IDF hampir sama (0,741 dan 0,740) dan BM25 tetap tertinggi. Pada panel (c), Precision@5 LSA naik dari 0,817 (d = 10) ke 0,892 (d = 200) tetapi tidak melampaui acuan BM25 (0,927); hybrid hanya berubah sedikit terhadap dimensi (0,925 pada d = 10 dan 0,911 pada d = 200). Penyebab LSA tertinggal belum diuji; dugaan penulis adalah korpus yang kecil (400 dokumen, 315 istilah unik) dan bentuk permukaan setara yang jarang muncul bersama dalam satu dokumen, sehingga kesetaraannya sulit dipelajari dari ko-okurensi.




**Gambar 4. Sensitivitas Precision@5 terhadap (a) variasi istilah kueri, (b) kebisingan minat, dan (c) dimensi LSA (rata-rata 20 seed, batang galat interval kepercayaan 95 persen).**

### Hasil per Bidang dan Contoh Kasus

Precision@5 per bidang mahasiswa ditunjukkan pada Gambar 5. Bidang SI (jaringan komputer dan IoT) paling mudah (BM25 0,95) dan bidang SEC (keamanan siber) paling sulit pada keempat metode (BM25 0,87). Penyebabnya belum diuji; salah satu dugaan adalah tumpang tindih istilah umum seperti "aplikasi" atau "web" dengan bidang lain.




**Gambar 5. Precision@5 menurut bidang utama mahasiswa dan metode (rata-rata 20 seed).**

Contoh untuk satu mahasiswa sintetis (seed 0, bidang utama pemrosesan bahasa alami, bidang sekunder interaksi manusia-komputer) diberikan pada Tabel 5. Kelima rekomendasi teratas BM25 dan LSA semuanya relevan, tetapi seluruhnya berasal dari bidang utama dan tidak ada yang berasal dari bidang sekunder. Satu kasus ini hanya ilustrasi, namun sejalan dengan keterbatasan keragaman hasil pada sistem berbasis konten (Algarni dan Sheldon, 2023).

**Tabel 5. Lima topik teratas untuk satu mahasiswa sintetis (seed 0, mahasiswa ke-7)**

| Peringkat | BM25 | LSA (d = 50) |
|---|---|---|
| 1 | studi komparatif opini publik berbasis pengenalan entitas pada perbankan | studi komparatif ringkasan teks otomatis menggunakan temu kembali informasi pada perbankan |
| 2 | rancang bangun representasi kata berbasis opini publik pada pemerintahan daerah | analisis tanya jawab otomatis dengan pendekatan analisis sentimen pada perpustakaan |
| 3 | studi komparatif ringkasan teks otomatis menggunakan temu kembali informasi pada perbankan | pengembangan tanya jawab otomatis dengan pendekatan representasi kata pada e-commerce |
| 4 | pengembangan tanya jawab otomatis dengan pendekatan representasi kata pada e-commerce | rancang bangun pemodelan topik dengan pendekatan representasi kata pada perpustakaan |
| 5 | analisis tanya jawab otomatis dengan pendekatan analisis sentimen pada perpustakaan | rancang bangun representasi kata berbasis opini publik pada pemerintahan daerah |

## PEMBAHASAN

Temuan utama adalah bahwa pada data simulasi ini metode leksikal (BM25 dan TF-IDF) sekurang-kurangnya setara dengan LSA, dan BM25 terbaik. Arah ini sejalan dengan Saputra et al. (2026) yang menemukan BM25 jauh di atas model embedding kalimat pada Wikipedia Indonesia, serta dengan Thakur et al. (2021) yang menyebut BM25 sebagai pembanding yang kuat. Namun penelitian ini tidak dapat menyatakan apa pun tentang SBERT, karena SBERT tidak dievaluasi. Bukti yang bertentangan dari Nabila et al. (2025), yang menemukan BERT dan Word2Vec lebih baik daripada TF-IDF, juga tidak diuji ulang di sini. Implikasi praktisnya adalah bahwa BM25 direkomendasikan sebagai pembanding utama maupun komponen penyaring tahap awal pada arsitektur rekomendasi dua tahap (two-stage retriever-reranker) di perguruan tinggi, di mana hybrid BM25 dan LSA menunjukkan kinerja yang kompetitif (selisih terhadap BM25 hanya sekitar 0,013 pada Precision@5) (sekitar 0,013 pada Precision@5) sementara pada dimensi LSA rendah tidak jauh dari BM25.

Terdapat beberapa keterbatasan yang membatasi generalisasi. Pertama, seluruh data bersifat sintetis dan label relevansi ditentukan oleh bidang yang sama dengan pembentuk kosakata; hal ini menguntungkan metode leksikal karena tumpang tindih istilah antara kueri dan topik relevan dijamin

oleh generator. Kedua, variasi leksikal hanya berupa pergantian di antara 1 sampai 3 bentuk tetap, bukan parafrasa bebas yang menjadi keunggulan embedding terlatih. Ketiga, relevansi bersifat biner per bidang dan tidak memodelkan kecocokan dengan dosen pembimbing, kesulitan topik, atau ketersediaan data. Keempat, bobot rekam jejak 0,5 dan parameter BM25 tidak disetel, dan hasil ablasi dipengaruhi asumsi kebisingan generator (lihat pembahasan RQ2). Kelima, tidak ada studi pengguna, dan tidak ada pengukuran keragaman atau kebaruan. Keenam, metode statistik memperlakukan 20 seed sebagai sampel independen dari generator yang sama, sehingga kesimpulan hanya berlaku untuk generator ini.

## KESIMPULAN

Penelitian ini membangun dan mengevaluasi prototipe rekomendasi topik tugas akhir berbasis content-based filtering pada korpus sintetis 400 topik dan 150 mahasiswa per seed dengan 20 seed. Untuk RQ1, BM25 memperoleh Precision@5 tertinggi (0,927), diikuti hybrid (0,914), TF-IDF (0,895), dan LSA (0,876), jauh di atas acak (0,21). Untuk RQ2, penambahan rekam jejak mata kuliah meningkatkan hasil dibandingkan minat saja, tetapi hasil ini bergantung pada asumsi generator. Untuk RQ3, peningkatan variasi istilah menurunkan LSA paling tajam, sehingga hipotesis keunggulan metode semantik laten tidak didukung, dan LSA mendekati tetapi tidak melampaui BM25 saat dimensi dinaikkan.

Saran untuk penelitian lanjutan adalah menjalankan skrip SBERT yang telah disediakan, dengan model paraphrase-multilingual-MiniLM-L12-v2 dan pembanding multilingual-E5, pada repositori tugas akhir nyata yang telah mendapat izin dan memakai penilaian relevansi oleh dosen; menguji hybrid BM25 dengan SBERT; menyetel bobot rekam jejak; serta menambahkan pengukuran keragaman dan studi pengguna. Seluruh klaim tentang efektivitas SBERT untuk mahasiswa Informatika Indonesia baru dapat dinyatakan setelah langkah-langkah tersebut dilakukan.

## UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih yang sebesar-besarnya kepada Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhadi Setiabudi atas dukungan sarana komputasi, fasilitasi riset, dan iklim akademik yang mendukung selama penyusunan studi simulasi sistem rekomendasi ini.

## KONTRIBUSI PENULIS

Nur Ariesanto Ramdhan: Conceptualization, Methodology, Investigation, Formal Analysis, Writing - Original Draft, Writing - Review and Editing. Penulis telah membaca dan menyetujui naskah versi akhir.

## PERNYATAAN KETERSEDIAAN DATA

Data yang dianalisis adalah data sintetis yang dibangkitkan oleh skrip simulasi dengan seed 0 sampai 19 (korpus dan profil seed 0 disertakan pada paket kode). Paket skrip simulasi generator korpus (seed 0 sampai 19), dataset sintetis lengkap, parameter uji, dan skrip evaluasi SBERT pendamping (sbert_extension.py) tersedia dari penulis atas permintaan yang wajar untuk kepentingan verifikasi dan pengembangan riset lanjutan.

## PERNYATAAN BEBAS KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa tidak ada konflik kepentingan finansial maupun hubungan pribadi yang dapat mempengaruhi hasil yang dilaporkan dalam naskah ini.

## PERNYATAAN ETIKA PENELITIAN DAN PUBLIKASI

Penelitian ini tidak melibatkan subjek manusia maupun data pribadi karena seluruh profil mahasiswa dan topik bersifat sintetis. Para penulis menyatakan bahwa penulisan naskah ini mematuhi standar etika penelitian dan publikasi serta bebas dari plagiasi.

## PERNYATAAN PEMANFAATAN TEKNOLOGI ASISTIF

Rancangan simulasi, kode analisis, gambar, dan draf awal naskah ini disusun dengan bantuan asisten kecerdasan buatan generatif (Claude, Anthropic), termasuk pencarian dan penelaahan literatur. Seluruh angka hasil eksperimen dalam naskah berasal dari keluaran kode yang dijalankan, dan setiap rujukan telah dicek metadata bibliografisnya pada Crossref atau situs penerbit. Para penulis bertanggung jawab penuh atas isi naskah dan wajib meninjau serta memverifikasi seluruh konten sebelum naskah dikirimkan.

## REFERENSI


Algarni, S., dan Sheldon, F. (2023). Systematic review of recommendation systems for course selection. *Machine Learning and Knowledge Extraction, 5*(2), 560-596. https://doi.org/10.3390/make5020033

Damayanti, A., Purwani, F., dan Kadafi, M. (2025). A content-based thesis supervisor recommendation system based on research interest clustering and cosine similarity. *JUSIFO (Jurnal Sistem Informasi), 11*(2), 111-120. https://doi.org/10.19109/jusifo.v11i2.27605

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., dan Wang, W. (2022). Language-agnostic BERT sentence embedding. Dalam *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (hlm. 878-891). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Guesmi, M., Chatti, M. A., Kadhim, L., Joarder, S., dan Ain, Q. U. (2023). Semantic interest modeling and content-based scientific publication recommendation using word embeddings and sentence encoders. *Multimodal Technologies and Interaction, 7*(9), 91. https://doi.org/10.3390/mti7090091

Hairani, H., dan Mujahid, M. (2022). Recommendations of thesis supervisor using the cosine similarity method. *SISTEMASI, 11*(3), 646. https://doi.org/10.32520/stmsi.v11i3.2003

Ibrahim, O. A. S., Younis, E. M. G., Mohamed, E. A., dan Ismail, W. N. (2025). Revisiting recommender systems: An investigative survey. *Neural Computing and Applications, 37*(4), 2145-2173. https://doi.org/10.1007/s00521-024-10828-5

Ji, Y., Sun, A., Zhang, J., dan Li, C. (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems, 41*(3), 1-27. https://doi.org/10.1145/3569930

Kreutz, C. K., dan Schenkel, R. (2022). Scientific paper recommendation systems: A literature review of recent publications. *International Journal on Digital Libraries, 23*(4), 335-369. https://doi.org/10.1007/s00799-022-00339-w

Kusuma, H. S., dan Musdholifah, A. (2021). Recommendation system for thesis topics using content-based filtering. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 15*(1), 65. https://doi.org/10.22146/ijccs.62716

Nabila, Q., Luthfiarta, A., Syabilla, M., Ahmad, A., dan Riyanto, R. (2025). Comparative analysis of vectorization methods for academic supervisor recommendations. *Jurnal Teknologi Informasi dan Terapan, 11*(2), 115-123. https://doi.org/10.25047/jtit.v11i2.438

Reimers, N., dan Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Dalam *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

*Processing (EMNLP-IJCNLP)* (hlm. 3982-3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reimers, N., dan Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. Dalam *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (hlm. 4512-4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Saputra, T., Julianto, E., Widjonarko, A., dan Tjahjono, B. (2026). Comparative evaluation of sparse, dense, and hybrid retrieval models on Indonesian Wikipedia. *Jurnal Teknik Informatika (JUTIF), 7*(3), 2920-2934. https://doi.org/10.52436/1.jutif.2026.7.3.5776

Syabilla, M., Zeniarja, J., dan Nabila, Q. (2025). Implementasi LDA, TF-IDF, dan BERT dalam sistem rekomendasi dosen pembimbing untuk mahasiswa. *Building of Informatics, Technology and Science (BITS), 6*(4), 2175-2183. https://doi.org/10.47065/bits.v6i4.6499

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., dan Gurevych, I. (2021). *BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models* (arXiv:2104.08663). arXiv. https://arxiv.org/abs/2104.08663

Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., dan Oleagordia-Ruiz, I. (2021). Recommendation systems for education: Systematic review. *Electronics, 10*(14), 1611. https://doi.org/10.3390/electronics10141611

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., dan Wei, F. (2024). *Multilingual E5 text embeddings: A technical report* (arXiv:2402.05672). arXiv. https://arxiv.org/abs/2402.05672

Zangerle, E., dan Bauer, C. (2022). Evaluating recommender systems: Survey and framework. *ACM Computing Surveys, 55*(8), 1-38. https://doi.org/10.1145/3556536