ANALISIS KINERJA DAN PENGEMBANGAN MODEL HYBRID TF-IDF DAN SENTENCE EMBEDDING (SBERT/E5) UNTUK MENINGKATKAN AKURASI SIMILARITY TEKS PADA REPOSITORY INSTITUSI BERBAHASA INDONESIA

(Studi Kasus: eprints.amikom.ac.id)

Disusun oleh:

Nama : **Ero Wahyu Pratomo**
NIM : **24.55.1583**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 PJJ INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

TESIS

ANALISIS KINERJA DAN PENGEMBANGAN MODEL HYBRID TF-IDF DAN SENTENCE EMBEDDING (SBERT/E5) UNTUK MENINGKATKAN AKURASI SIMILARITY TEKS PADA REPOSITORY INSTITUSI BERBAHASA INDONESIA

(Studi Kasus: eprints.amikom.ac.id)

ANALYSIS OF PERFORMANCE AND DEVELOPMENT OF A HYBRID TF-IDF AND SENTENCE EMBEDDING (SBERT/E5) MODEL TO ENHANCE TEXT SIMILARITY ACCURACY IN INDONESIAN-LANGUAGE INSTITUTIONAL REPOSITORIES

(Case Study: eprints.amikom.ac.id)

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi PJJ Informatika

Disusun oleh:

Nama : **Ero Wahyu Pratomo**
NIM : **24.55.1583**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 PJJ INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

ANALISIS KINERJA DAN PENGEMBANGAN MODEL HYBRID TF-IDF DAN SENTENCE EMBEDDING (SBERT/E5) UNTUK MENINGKATKAN AKURASI SIMILARITY TEKS PADA REPOSITORY INSTITUSI BERBAHASA INDONESIA

ANALYSIS OF PERFORMANCE AND DEVELOPMENT OF A HYBRID TF-IDF AND SENTENCE EMBEDDING (SBERT/E5) MODEL TO ENHANCE TEXT SIMILARITY ACCURACY IN INDONESIAN-LANGUAGE INSTITUTIONAL REPOSITORIES

Dipersiapkan dan Disusun oleh

Ero Wahyu Pratomo

24.55.1583

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 6 Januari 2026

Dosen Pembimbing,

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

HALAMAN PENGESAHAN

ANALISIS KINERJA DAN PENGEMBANGAN MODEL HYBRID TF-IDF DAN SENTENCE EMBEDDING (SBERT/E5) UNTUK MENINGKATKAN AKURASI SIMILARITY TEKS PADA REPOSITORY INSTITUSI BERBAHASA INDONESIA

ANALYSIS OF PERFORMANCE AND DEVELOPMENT OF A HYBRID TF-IDF AND SENTENCE EMBEDDING (SBERT/E5) MODEL TO ENHANCE TEXT SIMILARITY ACCURACY IN INDONESIAN-LANGUAGE INSTITUTIONAL REPOSITORIES

Dipersiapkan dan Disusun oleh

Ero Wahyu Pratomo

24.55.1583

Telah dipertahankan di depan Dewan Penguji
pada tanggal 6 Januari 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Dr. Kumara Ari Yuana, S.T., M.T. **NIK. 190302575**	
Robert Marco, S.T., M.T., Ph.D. **NIK. 190302228**	
Prof. Dr. Ema Utami, S.Si., M.Kom. **NIK. 190302037**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 6 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ero Wahyu Pratomo
NIM : 24.55.1583
Konsentrasi : Digital Transformation Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Analisis Kinerja dan Pengembangan Model Hybrid TF-IDF dan Sentence Embedding (SBERT/E5) untuk Meningkatkan Akurasi Similarity Teks pada Repository Institusi Berbahasa Indonesia

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.St., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 6 Januari 2026
Yang Menyatakan,

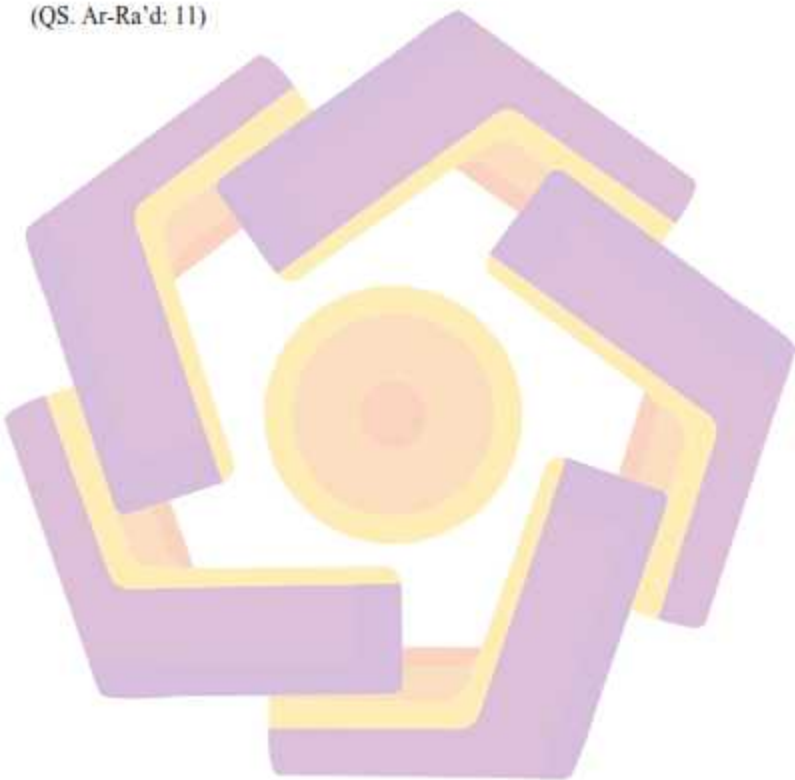
Ero Wahyu Pratomo

HALAMAN PERSEMBAHAN

Tesis ini dipersembahkan dengan penuh rasa syukur dan hormat kepada istri tercinta atas doa, kesabaran, pengertian, serta dukungan moril dan materiil yang senantiasa diberikan selama proses penyusunan tesis ini; kepada anak tersayang yang menjadi sumber semangat, inspirasi, dan motivasi dalam setiap langkah perjuangan akademik; kepada orang tua tercinta atas kasih sayang, doa, bimbingan, serta nilai-nilai kehidupan yang menjadi landasan dalam menempuh pendidikan; serta kepada para dosen dan pembimbing atas arahan, ilmu, waktu, dan dedikasi yang diberikan dengan penuh tanggung jawab dan profesionalisme. Semoga karya ini menjadi wujud tanggung jawab akademik dan memberikan manfaat bagi pengembangan ilmu pengetahuan serta masyarakat luas.

HALAMAN MOTTO

"Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum sampai mereka mengubah keadaan yang ada pada diri mereka sendiri." (QS. Ar-Ra'd: 11)



KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga tesis ini dapat diselesaikan dengan baik sebagai salah satu syarat untuk memperoleh gelar akademik pada program studi yang ditempuh.

Penulis menyampaikan terima kasih dan penghargaan yang setinggi-tingginya kepada Dosen Pembimbing, Prof. Dr. Ema Utami, S.Si., M.Kom., atas segala arahan, bimbingan, waktu, dan perhatian yang diberikan dengan penuh kesabaran dan tanggung jawab selama proses penyusunan tesis ini. Ucapan terima kasih juga penulis sampaikan kepada Tim Dosen Penguji, Dr. Kumara Ari Yuana, S.T., M.T., dan Robert Marco, S.T., M.T., Ph.D., atas masukan, koreksi, serta saran konstruktif yang sangat berharga bagi penyempurnaan tesis ini.

Penulis juga menyampaikan terima kasih kepada seluruh dosen dan tenaga kependidikan di lingkungan Universitas Amikom Yogyakarta atas dukungan akademik dan administratif selama masa studi. Penghargaan dan rasa hormat yang setinggi-tingginya penulis sampaikan kepada kedua orang tua tercinta atas kasih sayang, doa yang tiada henti, dukungan moral maupun spiritual, serta nilai-nilai kehidupan yang telah ditanamkan sehingga menjadi landasan dalam menempuh dan menyelesaikan pendidikan ini. Secara khusus, penulis menyampaikan terima kasih yang tulus kepada istri tercinta atas kesabaran, pengertian, doa, serta dukungan yang tidak pernah terputus selama proses penyusunan tesis ini. Kepada anak

tersayang, penulis menyampaikan terima kasih atas keceriaan, semangat, dan motivasi yang menjadi sumber energi dalam menyelesaikan studi ini.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi penyempurnaan karya ilmiah di masa mendatang. Semoga tesis ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan pihak-pihak yang memerlukan.

Yogyakarta, 6 Januari 2026

Ero Wahyu Pratomo

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

1. **Terminologi Model & Arsitektur**
   a. **TF-IDF (*Term Frequency-Inverse Document Frequency*):** Metode statistik yang digunakan untuk mengukur seberapa penting sebuah kata dalam suatu dokumen terhadap kumpulan dokumen (korpus). Dalam penelitian ini, digunakan sebagai filter leksikal.
   b. **Sentence Embedding:** Representasi numerik dari satu kalimat utuh ke dalam bentuk vektor berdimensi tinggi yang menangkap makna semantik.
   c. **Transformer:** Arsitektur *deep learning* yang menggunakan mekanisme *attention* untuk memahami hubungan konteks antar kata dalam teks.
   d. **SBERT (*Sentence-BERT*):** Modifikasi model BERT menggunakan struktur *Siamese Network* untuk menghasilkan *sentence embedding* yang optimal untuk perbandingan kemiripan teks.
   e. **E5 (*EmbEddings from bidirectional Encoder representations*):** Model *embedding* berbasis kontras yang dilatih secara masif untuk tugas retrival informasi dan kemiripan teks.
   f. **Hybrid Model:** Pendekatan yang menggabungkan dua atau lebih metode berbeda (dalam hal ini leksikal TF-IDF dan semantik *Transformer*) untuk mencapai keseimbangan antara akurasi dan efisiensi.
   g. **Reranking:** Proses pengurutan ulang kandidat dokumen hasil filter awal menggunakan model yang lebih kompleks (semantik) untuk meningkatkan presisi.
2. **Metrik Evaluasi & Validasi**
   a. **ARI (*Adjusted Rand Index*):** Metrik evaluasi eksternal yang mengukur kesesuaian antara hasil *clustering* dengan *ground truth* (label subjek), yang telah disesuaikan dengan faktor kebetulan.
   b. **NMI (*Normalized Mutual Information*):** Metrik untuk mengukur seberapa banyak informasi yang dibagikan antara hasil *clustering* dan label *ground truth*.
   c. **Silhouette Score:** Metrik evaluasi internal untuk mengukur seberapa mirip sebuah dokumen dengan klasternya sendiri dibandingkan dengan klaster lainnya (mengukur separasi).
   d. **Davies-Bouldin Index (DBI):** Metrik untuk mengevaluasi kualitas *clustering* berdasarkan rasio jarak di dalam klaster dengan jarak antar klaster. Skor lebih rendah menunjukkan hasil yang lebih baik.
   e. **Ground Truth:** Referensi kebenaran absolut yang digunakan sebagai standar pembanding, dalam penelitian ini menggunakan atribut **Subject** dari metadata repositori.

f. **Lost Ratio:** Rasio jumlah dokumen yang dieliminasi pada tahap filtrasi awal terhadap total populasi dokumen dalam dataset.

3. **Pemrosesan Data & Dimensi**
   a. **Cosine Similarity:** Metrik untuk mengukur kesamaan antara dua vektor dengan menghitung kosinus sudut di antara keduanya.
   b. **PCA (*Principal Component Analysis*):** Teknik reduksi dimensi yang digunakan untuk memproyeksikan data berdimensi tinggi (vektor *embedding*) ke dalam ruang dua dimensi untuk visualisasi.
   c. **t-SNE (*t-Distributed Stochastic Neighbor Embedding*):** Algoritma reduksi dimensi non-linear yang sangat efektif untuk memvisualisasikan struktur klaster yang kompleks.
   d. **Curse of Dimensionality:** Fenomena di mana ruang data menjadi sangat luas seiring bertambahnya jumlah dimensi (seperti pada metode *hybrid*), yang dapat menyebabkan penurunan kinerja metrik jarak seperti *Silhouette*.

INTISARI

Penelitian ini mengembangkan dan menganalisis performa model *hybrid* antara metode TF-IDF dan *sentence embedding* berbasis *transformer* (SBERT dan E5) untuk meningkatkan akurasi pengukuran kemiripan (*text similarity*) pada dokumen akademik berbahasa Indonesia. Permasalahan utama yang diangkat adalah keterbatasan TF-IDF dalam menangkap konteks semantik serta tingginya biaya komputasi apabila model *embedding* diterapkan secara penuh pada repositori berskala besar (28.575 dokumen).

Untuk mengatasi hal tersebut, penelitian ini mengusulkan arsitektur *hybrid* dua tahap: TF-IDF digunakan sebagai filter leksikal agresif dengan *lost ratio* sebesar 0,9965 untuk mereduksi ruang pencarian, diikuti oleh *reranking* semantik menggunakan SBERT dan E5. Hasil eksperimen menunjukkan bahwa model SBERT unggul dalam stabilitas struktur klaster dengan skor ARI tertinggi (0,3444) dan Davies-Bouldin Index terendah (2,7506), sedangkan E5 unggul dalam aspek kelengkapan semantik dengan NMI sebesar 0,5183.

Dari sisi efisiensi, arsitektur *hybrid* berhasil memangkas waktu pemrosesan secara signifikan. Metode Hybrid TF-IDF + E5 (3,71 detik) terbukti 47,7% lebih cepat dibandingkan Hybrid TF-IDF + SBERT (7,10 detik). Secara keseluruhan, pendekatan *hybrid* mampu meningkatkan kualitas *similarity* dan struktur klaster dibandingkan penggunaan model tunggal. Model ini terbukti mampu menyeimbangkan efisiensi komputasi dan ketepatan semantik, sehingga sangat relevan untuk diimplementasikan pada sistem rekomendasi dokumen akademik berskala besar di repositori institusi.

Kata kunci: TF-IDF, SBERT, E5, *text similarity*, *hybrid model*, *ground truth*, repositori akademik.

ABSTRACT

This study develops and analyzes a hybrid text similarity model combining TF-IDF and transformer-based sentence embeddings (SBERT and E5) to enhance similarity measurement for Indonesian academic documents. The primary challenges addressed are the lexical limitations of TF-IDF and the high computational overhead of applying transformer models to a large-scale repository of 28,575 documents.

To address these issues, a two-stage hybrid architecture is proposed: TF-IDF serves as an aggressive lexical filter with a lost ratio of 0.9965 to reduce the search space, followed by semantic reranking using SBERT and E5. Experimental results demonstrate that SBERT outperforms other models in clustering stability, achieving the highest Adjusted Rand Index (ARI) of 0.3444 and the lowest Davies-Bouldin Index (2.7506). Conversely, E5 excels in semantic completeness with a Normalized Mutual Information (NMI) score of 0.5183.

In terms of efficiency, the hybrid architecture significantly reduces processing time. The Hybrid TF-IDF + E5 (3.71 seconds) approach is 47.7% faster than the Hybrid TF-IDF + SBERT (7.10 seconds). Overall, the hybrid approach enhances similarity quality and clustering structure compared to single-model methods. This model effectively balances computational efficiency and semantic precision, making it highly suitable for large-scale academic repository recommendation systems.

Keyword: TF-IDF, SBERT, E5, text similarity, hybrid model, ground truth, academic repository.

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Mendeteksi kesamaan merupakan salah satu teknik utama dalam analisis teks yang bertujuan menentukan skor kemiripan antara dua teks berdasarkan kesamaan konten atau makna yang dikandungnya. Kesamaan dapat muncul baik secara leksikal, yaitu melalui kemiripan urutan karakter atau kata, maupun secara semantik, yakni melalui kesamaan makna atau konteks penggunaan kata dalam kalimat. Dalam pendekatan leksikal, metode TF-IDF (Term Frequency–Inverse Document Frequency) menjadi salah satu teknik paling populer karena kesederhanaan dan efisiensinya dalam mengukur kemiripan antar dokumen berdasarkan distribusi kata. TF-IDF bekerja dengan menilai frekuensi kemunculan kata pada suatu dokumen dibandingkan keseluruhan korpus, sehingga mampu menonjolkan kata kunci yang bersifat unik. Namun, metode ini memiliki keterbatasan karena tidak memahami konteks atau makna di balik kata tersebut. Akibatnya, dua kalimat yang memiliki arti serupa tetapi menggunakan kata berbeda dapat dianggap tidak mirip, sedangkan dua kalimat yang memiliki kata serupa namun maknanya berbeda justru dianggap mirip (Chaichulee et al., 2022; Mansoor et al., 2020; Yu et al., 2024).

Hal ini, berbeda dengan pendekatan semantik berupaya memahami konteks dan makna yang terkandung dalam teks melalui representasi vektor (embedding) yang merepresentasikan hubungan semantik antar kata atau kalimat. Model

semantik seperti Word2Vec dan GloVe mempelajari hubungan antar kata berdasarkan konteks kemunculannya, namun model-model ini masih terbatas karena tidak memperhitungkan arah konteks secara penuh. Perkembangan besar dalam bidang pemrosesan bahasa alami (Natural Language Processing) terjadi dengan hadirnya BERT (Bidirectional Encoder Representations from Transformers), yang memanfaatkan mekanisme attention dua arah untuk memahami makna kata dalam konteks keseluruhan kalimat. Dengan arsitektur transformer-nya, BERT mampu menghasilkan embedding yang kaya makna dan relevan dengan konteks linguistik yang kompleks (Ibrahem Al-Obaydy et al., 2022).

Beberapa peneliti sebelumnya telah melakukan penelitian dalam bidang text similarity untuk mendapatkan kinerja terbaik, namun hasil akurasi yang dicapai masih belum memuaskan. Hal ini disebabkan oleh berbagai kendala dalam bidang text processing, seperti keterbatasan jumlah kata dan makna yang dapat direpresentasikan secara kontekstual, ketidakmampuan model menangani sinonim atau variasi morfologi bahasa Indonesia, serta tingginya kebutuhan komputasi pada model berbasis deep learning. Beberapa algoritma populer yang sering digunakan dalam pengukuran kemiripan teks meliputi TF-IDF, Word2Vec, BERT, SBERT, dan E5. Di antara algoritma tersebut, SBERT dan E5 memiliki kekuatan utama dalam menangkap hubungan semantik antar kalimat dan memahami konteks secara mendalam melalui representasi vektor yang bermakna. SBERT mampu mengenali kemiripan semantik meskipun struktur kalimat berbeda, sementara E5 unggul dalam mengolah instruksi dan memahami tujuan teks melalui pendekatan

instruction-tuning. Namun demikian, kedua model ini memiliki kelemahan dalam efisiensi komputasi, membutuhkan memori besar, serta waktu inferensi yang lama ketika diterapkan pada dataset berskala besar seperti dokumen akademik di repository institusi (Bergman et al., 2023; Ibrahem Al-Obaydy et al., 2022).

Berdasarkan permasalahan tersebut, maka dalam penelitian ini diusulkan pendekatan hybrid yang mengombinasikan kekuatan model leksikal dan semantik. Meskipun model berbasis embedding seperti SBERT dan E5 memiliki keunggulan dalam memahami konteks dan makna teks, kelemahannya terletak pada konsumsi sumber daya yang tinggi. Oleh karena itu, penelitian ini menambahkan pendekatan TF-IDF sebagai tahap penyaringan awal untuk mengurangi jumlah dokumen yang perlu diproses oleh model semantik. Pendekatan ini bertujuan untuk mengatasi masalah efisiensi tanpa mengorbankan akurasi. Dengan mengintegrasikan kelebihan TF-IDF dalam identifikasi kata kunci dan kekuatan SBERT serta E5 dalam memahami makna kontekstual, diharapkan kombinasi model ini dapat meningkatkan performa pengukuran similarity sekaligus menyeimbangkan efisiensi leksikal dan akurasi semantik. Melalui strategi hybrid ini, sistem rekomendasi dokumen akademik diharapkan menjadi lebih cepat, akurat, dan adaptif terhadap karakteristik bahasa Indonesia.

Justifikasi penggunaan model hybrid menjadi semakin kuat ketika mempertimbangkan bahwa TF-IDF unggul dalam efisiensi dan mampu menyaring dokumen berdasarkan kata kunci dominan, sementara model seperti SBERT dan E5 unggul dalam memahami kedekatan makna melalui representasi semantik. Kedua pendekatan ini memiliki kekuatan yang saling melengkapi TF-IDF

memberikan seleksi awal dokumen dengan biaya komputasi sangat rendah, sedangkan SBERT dan E5 memberikan penilaian akhir yang kaya makna namun mahal secara komputasi. Dengan mengintegrasikan kedua jenis representasi tersebut, model hybrid mampu mengurangi beban proses embedding secara drastis sekaligus meningkatkan akurasi pengukuran similarity. Oleh karena itu, pendekatan hybrid dipilih sebagai solusi optimal untuk konteks dokumen akademik berbahasa Indonesia yang berskala besar dan kaya variasi linguistik.

1.2 Rumusan Masalah

Penelitian ini dirancang untuk mengkaji dan memberikan solusi terhadap permasalahan-permasalahan berikut:

1. Bagaimana performansi metode TF-IDF dan model word embedding dalam mengukur kemiripan teks Bahasa Indonesia pada dokumen akademik, serta apa keterbatasan masing-masing pendekatan dalam menangkap konteks semantik?
2. Apakah penerapan model hybrid yang menggabungkan TF-IDF dan embedding dapat meningkatkan akurasi serta efisiensi pengukuran kemiripan teks?

Rumusan masalah ini akan menjadi dasar untuk pengembangan model, analisis performansi, serta evaluasi hasil penelitian demi menghasilkan sistem similarity teks yang lebih akurat dan aplikatif untuk repositori institusi berbahasa Indonesia.

1.3 Batasan Masalah

a. Penelitian ini hanya mengkaji pengukuran similarity teks Bahasa Indonesia yang terdapat pada repository institusi eprints.amikom.ac.id sebagai studi kasus.
b. Model yang dikembangkan adalah model hybrid antara TF-IDF dan word embedding, khususnya menggunakan embedding BERT dan E5, sehingga tidak membahas metode NLP lainnya secara mendalam seperti topic modeling atau transformer murni selain BERT.
c. Analisis performansi difokuskan pada akurasi pengukuran similarity dan efisiensi komputasi tanpa memperluas pada aspek user interface atau integrasi sistem pencarian yang operasional.
d. Data yang digunakan terbatas pada dokumen akademik jenis tugas akhir dan skripsi yang tersedia dalam repository tersebut, tidak termasuk jenis dokumen lain seperti jurnal atau artikel populer.
e. Eksperimen model embedding lebih memprioritaskan metode embedding BERT dan E5 yang relevan untuk Bahasa Indonesia, tanpa memasukkan model embedding yang khusus untuk domain lain.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Menganalisis performansi metode TF-IDF dan model embedding (SBERT, E5) dalam mengukur kemiripan teks Bahasa Indonesia pada dokumen akademik, serta mengidentifikasi keterbatasan

masing-masing metode dalam menangkap konteks semantik dan variasi linguistik Bahasa Indonesia.

2. Mengembangkan dan mengevaluasi model hybrid TF-IDF + embedding untuk meningkatkan akurasi dan efisiensi perhitungan similarity, sehingga model dapat diimplementasikan secara praktis pada sistem pencarian dan rekomendasi dokumen di repository institusi, khususnya yang berbasis EPrints.

1.5 Manfaat Penelitian

a. Manfaat Ilmiah dalam Ilmu Pengetahuan

1. Menambah khasanah pengetahuan di bidang Natural Language Processing (NLP), khususnya dalam penerapan teknik hybrid TF-IDF dan word embedding (BERT dan E5) untuk pengukuran text similarity pada teks berbahasa Indonesia.
2. Memberikan kontribusi konseptual berupa model hybrid baru yang menggabungkan kekuatan metode statistik dan representasi semantik, sehingga meningkatkan akurasi pengukuran kemiripan teks pada dokumen akademik.
3. Menjadi referensi bagi penelitian lanjutan dalam pengembangan metode hybrid untuk keperluan text similarity, clustering, serta sistem rekomendasi dokumen berbasis semantik.
4. Menghasilkan analisis empiris dan evaluasi performa yang memperkaya literatur ilmiah terkait penerapan model embedding

BERT dan E5 dalam konteks bahasa Indonesia, yang masih relatif terbatas.

b. Manfaat terhadap Sistem yang Dikembangkan

1. Sistem yang dikembangkan mampu mengukur kemiripan teks dengan tingkat akurasi lebih tinggi dibandingkan penggunaan metode tunggal seperti TF-IDF atau embedding secara terpisah.
2. Model hybrid yang diusulkan memungkinkan proses pencarian dan pengelompokan dokumen akademik menjadi lebih relevan dan efisien, karena mengombinasikan analisis leksikal dan semantik secara adaptif.
3. Sistem ini dapat diintegrasikan untuk meningkatkan efektivitas sistem rekomendasi dokumen serupa di repositori akademik, membantu pengguna menemukan referensi yang kontekstual secara lebih cepat.
4. Memberikan dukungan terhadap digitalisasi dan pengelolaan informasi ilmiah dengan memanfaatkan teknologi NLP terkini yang bersifat fleksibel dan hemat sumber daya.

c. Manfaat Praktis bagi Institusi atau Pengguna

1. Institusi akademik seperti perguruan tinggi dan perpustakaan digital dapat meningkatkan kualitas layanan pencarian dan rekomendasi dokumen, khususnya dalam repositori institusi.
2. Mempermudah dosen, peneliti, dan mahasiswa dalam menemukan sumber referensi yang relevan dengan topik penelitian mereka,

sehingga dapat meningkatkan produktivitas akademik dan publikasi ilmiah.

3. Mengurangi waktu dan biaya pencarian manual dokumen yang serupa secara signifikan melalui otomatisasi berbasis semantic similarity.
4. Memperkuat posisi institusi dalam pengelolaan repositori digital dengan penerapan teknologi NLP modern yang kompetitif dan berdaya saing tinggi di bidang riset dan pendidikan.

BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Berbagai penelitian sebelumnya telah mengkaji pendekatan dalam pengukuran text similarity menggunakan metode statistik maupun representasi semantik. Salah satu metode klasik yang masih banyak digunakan adalah Term Frequency–Inverse Document Frequency (TF-IDF). Metode ini terbukti efektif dalam mengukur kesamaan teks, terutama pada dokumen yang panjang seperti artikel ilmiah atau tugas akhir, karena mampu memberikan bobot yang proporsional terhadap kata-kata penting dalam teks. Keunggulan utama TF-IDF terletak pada kecepatan pemrosesan dan kemudahan interpretasi hasil, sehingga sering dijadikan dasar bagi sistem pencarian informasi berbasis konten. Namun demikian, kelemahan mendasar dari TF-IDF adalah ketidakmampuannya dalam menangkap konteks semantic, metode ini hanya menghitung frekuensi kemunculan kata tanpa memahami hubungan makna antar kata dalam kalimat (Hassan & Ahmed, 2023).

Untuk mengatasi keterbatasan tersebut, muncul pendekatan berbasis embedding kontekstual, salah satunya adalah BERT (Bidirectional Encoder Representations from Transformers). Model ini mampu mempelajari konteks kata dalam dua arah secara bersamaan, sehingga menghasilkan representasi vektor yang merefleksikan makna semantik secara lebih mendalam. Penelitian-penelitian terdahulu menunjukkan bahwa BERT memberikan hasil pengukuran similarity yang lebih akurat dibandingkan metode tradisional. Namun, tantangan utama dari

pendekatan ini adalah kebutuhan sumber daya komputasi yang tinggi serta waktu inferensi yang relatif lama, terutama ketika diterapkan pada dataset besar atau sistem yang membutuhkan pemrosesan real-time (Babić et al., 2020).

Selanjutnya, sejumlah studi mulai mengembangkan model hybrid yang menggabungkan keunggulan metode statistik dan semantik. Model ini mengombinasikan fitur lokal yang dihasilkan TF-IDF dengan fitur kontekstual dari embedding seperti BERT dan E5, menghasilkan peningkatan akurasi yang signifikan serta efisiensi yang lebih baik dibandingkan pendekatan tunggal. Kelebihan utama pendekatan hybrid terletak pada kemampuannya untuk menyeimbangkan presisi semantik dan efisiensi perhitungan. Meskipun demikian, sebagian besar penelitian tersebut masih berfokus pada teks berbahasa Inggris, sehingga ruang pengembangan untuk adaptasi pada Bahasa Indonesia masih terbuka luas (Lan, 2022).

Dari sisi evaluasi metode, penelitian sebelumnya juga membandingkan berbagai teknik pengukuran kemiripan teks, seperti Cosine Similarity, Jaccard Similarity, dan ukuran jarak lainnya. Cosine Similarity secara konsisten menunjukkan performa yang baik dalam mengukur kesamaan antar dokumen menggunakan TF-IDF. Akan tetapi, sebagian besar penelitian tersebut belum memasukkan representasi semantik ke dalam perhitungan similarity, sehingga hasil pengukurannya masih terbatas pada aspek leksikal. Kondisi ini memperkuat justifikasi perlunya pengembangan pendekatan hybrid yang menggabungkan analisis leksikal dan semantik untuk menghasilkan penilaian kemiripan yang lebih kontekstual dan akurat (de Vos et al., 2022).

Selain itu, terdapat penelitian yang menerapkan ekstraksi kata kunci berbasis embedding, misalnya melalui algoritma KeyBERT, yang menggunakan representasi BERT untuk menghasilkan kata kunci paling relevan dalam dokumen akademik. Hasil penelitian tersebut menunjukkan peningkatan relevansi kata kunci dibandingkan metode berbasis frekuensi kata semata, membuktikan keunggulan embedding dalam menangani konteks linguistik dan makna kalimat (Khan et al., 2022).

Beberapa studi lain juga menyoroti integrasi TF-IDF dengan algoritma pembelajaran mesin seperti Support Vector Machine (SVM) dalam klasifikasi teks. Pendekatan ini menunjukkan hasil klasifikasi yang baik pada teks dengan struktur sederhana. Namun, karena tidak mempertimbangkan aspek semantik, performanya menurun pada teks yang mengandung variasi bahasa dan konteks kompleks. Hal ini menunjukkan perlunya penggabungan teknik semantik seperti embedding untuk menghasilkan model yang lebih adaptif terhadap variasi konteks bahasa (Cahyani & Patasik, 2021).

Secara keseluruhan, hasil tinjauan pustaka menunjukkan bahwa kombinasi antara TF-IDF dan embedding (BERT atau E5) berpotensi memberikan keseimbangan antara akurasi semantik dan efisiensi komputasi. Keterbatasan penelitian terdahulu yang masih berfokus pada Bahasa Inggris dan penggunaan model tunggal menjadi dasar penting bagi penelitian ini untuk melakukan pengembangan model hybrid TF-IDF dan embedding dalam konteks teks berbahasa Indonesia, khususnya pada dokumen akademik yang terdapat di repositori institusi.

2.2 Keaslian Penelitian

Tabel 2.1 Matriks Literature Review dan Posisi Penelitian

Analisis Performasi Dan Pengembangan Model Hybrid Tf-Id Dan Word Embedding Dalam Meningkatkan Akurasi Similarity Teks Bahasa Indonesia Pada Repository Institusi

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Evaluating Efficacy of Semantic Similarity Methods for Comparison of Academic Thesis and Dissertation Texts (Hassan & Ahmed, 2023)	Hassan R, Ahmed N, Science Journal of University of Zakho, 2023	Membandingkan metode TF-IDF, Doc2Vec, SBERT, dan BERT dalam mengukur kesamaan semantik dokumen akademik	TF-IDF memberikan akurasi tertinggi (97%) dan efisiensi waktu lebih baik dari metode embedding	Keterbatasan pada data Bahasa Inggris dan penggunaan threshold yang sensitif	Penelitian ini mengembangkan model hybrid TF-IDF dan embedding BERT & E5 di konteks Bahasa Indonesia untuk mengatasi batasan tersebut.
2	Deep Learning on Small Datasets without Pre-Training using Cosine Loss (Barz & Denzler, n.d.)	Barz B, Denzler J, IEEE, 2020	Mengoptimalkan deep learning pada dataset kecil tanpa pra-pelatihan menggunakan fungsi loss cosine	Metode deep learning mampu bekerja efektif tanpa pra-pelatihan dengan akurasi cukup baik	Terbatas pada dataset kecil dan domain tertentu klinis	Penelitian ini fokus pada pengujian metode hybrid dalam dokumen akademis dengan dataset yang lebih besar
3	Construction and Study of Textual Association Network Based on Cosine Similarity Algorithm	Chen Q,Zhang O, Institute of Electrical and Electronics	Membuat jaringan asosiasi kata berbasis cosine similarity	Teknik ini membantu menganalisis tren dan asosiasi kata dalam dokumen	Ketergantungan pada data dan algoritma tertentu	Penelitian ini menambahkan embedding semantik ke analisis similarity sehingga lebih kaya konteks

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	(Chen & Zhang, 2023)	Engineers Inc., 2023				
4	Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain (Chen et al., 2021)	Chen Q, Rankine A, Peng Y, Aghaarabi ELu Z, JMIR Medical Informatics, 2021	Mengevaluasi model deep learning untuk pengukuran similarity dalam domain klinis	Ada gap antara skor model dan anotasi manual sehingga perlu pengembangan dataset lebih baik	Model memerlukan data anotasi yang luas dan berkualitas tinggi	Penelitian ini menyesuaikan pada dokumen akademik dan menggunakan hybrid TF-IDF dan embedding yang lebih ringan
5	Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic (Lan, 2022)	Lan F, Advances in Multimedia, 2022	Mengembangkan algoritma hybrid untuk mengukur similarity teks secara kombinasi statistik dan semantik	Algoritma hybrid meningkatkan performa secara signifikan dibanding tunggal	Fokus penelitian masih di Bahasa Inggris dan dataset kecil	Penelitian ini mengadaptasi model hybrid pada teks Bahasa Indonesia dengan dataset akademik yang berbeda
6	Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports (Jiang et al., 2021)	Jiang Z, Gao B, He Y, Han Y, Doyle P, Zhu Q, Mathematical Problems in Engineering, 2021	Pengembangan TF-IDF untuk klasifikasi teks berita internet	Skema weighting yang diusulkan meningkatkan akurasi klasifikasi	Belum menangani konteks semantik secara mendalam	Penelitian ini menggunakan embedding untuk melengkapi TF-IDF sehingga menangkap konteks lebih baik

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
7	Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine (Gifari et al., 2022)	Gifari O, Adha M, Rifky Hendrawan I, Freddy F, Durrand S, JIFOTECH (Journal of Information Technology), 2022	Mengklasifikasi sentimen teks ulasan film	Metode TF-IDF dan SVM memberikan akurasi cukup baik	Hanya mengandalkan fitur statistik dan tidak mempertimbangkan konteks kalimat	Penelitian ini mengembangkan fitur embedding untuk memperbaiki representasi teks
8	Cosine Similarity - A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words (Januzaj & Luma, 2022)	Januzaj Y, Luma A, International Journal of Emerging Technologies in Learning, 2022	Mencocokkan similarity antara program pendidikan dan pasar kerja dengan cosine similarity	Cosine similarity efektif namun terbatas pada metode sederhana	Tidak menggunakan embedding untuk konteks semantik	Penelitian ini menggunakan embedding untuk meningkatkan pemahaman konteks kalimat secara mendalam
9	Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity	Wilianto D, Girsang A, TEM Journal, 2023	Menerapkan similarity teks untuk penilaian jawaban singkat pada e-learning	Metode embedding membantu meningkatkan akurasi penilaian	Dibatasi pada jawaban singkat dan domain pendidikan menengah	Penelitian ini mengalihkan fokus ke dokumen akademik lengkap dengan pendekatan hybrid

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Methods (Wilianto & Girsang, 2023)					

2.3 Landasan Teori

2.3.1 Leksikal

Dalam pemrosesan teks, aspek leksikal berfokus pada pengolahan dan representasi kata-kata sebagai unit dasar bahasa. Representasi leksikal tradisional biasanya menggunakan model sparse seperti bag-of-words, tf-idf, atau BM25 yang merepresentasikan teks dalam bentuk vektor berdimensi tinggi dengan banyak nilai nol. Model SPLADE yang mengadopsi pendekatan sparse lexical representation berbasis BERT MLM, dimana pembobotan dan ekspansi term dilakukan secara kontekstual untuk mengatasi masalah ketidaksesuaian kosa kata. Selain itu, model DeLADE memperkenalkan peningkatan densitas vektor leksikal guna memperbaiki efisiensi dan stabilitas komputasi, tanpa mengorbankan kualitas representasi (Lin & Lin, 2023). Oleh karena itu, pemrosesan leksikal masih menekankan pada pencocokan literal kata-token dan tetap relevan dalam sistem pencarian teks yang membutuhkan interpretasi eksplisit.

2.3.2 Semantik

Aspek semantik dalam pemrosesan teks lebih menitikberatkan pada pemahaman makna dan konteks kata di dalam kalimat atau dokumen, melampaui sekadar kesamaan token literal. Representasi semantik yang dominan menggunakan dense embeddings hasil dari model transformer pretrained seperti BERT, yang menyediakan vektor berdimensi rendah namun kaya makna dan mampu merepresentasikan konteks secara

menyeluruh (Lin & Lin, 2023). Selain itu, penggunaan Word2Vec sebagai metode distribusional semantic representation yang handal dalam mengukur kesamaan makna dan membantu proses ekstraksi keyphrase serta pengelompokan dokumen (Sarwar et al., 2022). Pendekatan semantik terbukti efektif memperbaiki relevansi dan kualitas retrieval teks walaupun tantangan masih ada terutama pada bahasa dengan sumber daya terbatas seperti bahasa daerah.

2.3.3 Representasi Teks dalam Pemrosesan Bahasa Alami (Natural Language Processing)

Representasi teks merupakan komponen fundamental dalam bidang Natural Language Processing (NLP) yang bertujuan mengubah teks mentah menjadi bentuk numerik atau vector representation agar dapat diproses oleh algoritma komputasi. Representasi ini digunakan dalam berbagai tugas seperti klasifikasi teks, pencarian informasi (information retrieval), sistem rekomendasi, serta clustering dokumen.

Secara umum, representasi teks terbagi menjadi dua pendekatan utama, yaitu:

1. Pendekatan Statistik (Bag-of-Words/TF-IDF), yang berfokus pada frekuensi kemunculan kata tanpa memperhatikan konteks semantik.
2. Pendekatan Pembelajaran Mendalam (Embedding), yang menghasilkan vektor representasi dengan mempertimbangkan makna dan hubungan antar kata atau kalimat.

Pendekatan klasik seperti TF-IDF unggul dalam hal kesederhanaan dan efisiensi komputasi, tetapi lemah dalam memahami makna kontekstual. Sebaliknya, pendekatan modern seperti BERT, SBERT, dan E5 memanfaatkan arsitektur Transformer untuk memahami hubungan semantik antar kata dalam konteks kalimat. Perkembangan ini secara signifikan meningkatkan akurasi sistem pencarian dan rekomendasi dokumen akademik (Bergman et al., 2023).

Dalam penelitian ini, representasi teks berfungsi untuk mengukur kesamaan antar judul dan abstrak tugas akhir mahasiswa, dengan tujuan mengelompokkan dokumen yang memiliki topik serupa serta memberikan rekomendasi berdasarkan kedekatan makna semantik.

2.3.4 Metode TF-IDF (Term Frequency–Inverse Document Frequency)

Metode TF-IDF merupakan salah satu teknik representasi teks paling populer untuk mengukur pentingnya suatu kata dalam dokumen relatif terhadap seluruh korpus.

Secara matematis, bobot TF-IDF dihitung menggunakan rumus berikut:

$$TF - IDF(t,d) = (f\_(t,d) / \Sigma f\_(k,d)) \times log(N / DF(t))$$

di mana:

- TF(t, d) adalah frekuensi kemunculan kata t pada dokumen d,
- DF(t) adalah jumlah dokumen yang memuat kata t, dan
- N adalah total jumlah dokumen pada korpus.

Penelitian oleh Ibrahem et al. (2022) menunjukkan bahwa kombinasi TF-IDF dan K-Means Clustering efektif dalam klasifikasi dokumen berbasis topik (Ibrahem Al-Obaydy et al., 2022). Metode ini mampu menonjolkan istilah teknis atau kata kunci unik sebagai pembeda topik.

Namun, Bergman et al. (2023) menegaskan bahwa pendekatan berbasis frekuensi seperti TF-IDF tidak mempertimbangkan konteks semantik, sehingga dua dokumen yang memiliki arti serupa tetapi kata berbeda dapat dinilai tidak mirip (Bergman et al., 2023).

Dalam penelitian ini, TF-IDF digunakan sebagai lapisan awal untuk mengidentifikasi kemiripan leksikal, sebelum hasilnya diperkuat dengan analisis semantik menggunakan embedding SBERT dan E5.

2.3.5 Sentence-BERT (SBERT)

Sentence-BERT (SBERT) merupakan pengembangan dari arsitektur BERT yang dioptimalkan untuk menghasilkan sentence embedding secara efisien. Model ini menggunakan arsitektur siamese dan triplet network, yang memungkinkan pembandingan antar kalimat secara langsung melalui metrik seperti Cosine Similarity.

SBERT mengekstraksi konteks kata dari dua arah (kiri dan kanan) dan mengubah kalimat menjadi vektor berdimensi tetap, sehingga dapat mengenali kesamaan makna antar kalimat dengan struktur yang berbeda.

Penelitian Gatto et al. (2022) membuktikan efektivitas SBERT dalam menangkap hubungan semantik pada teks medis yang kompleks (Gatto et

al., 2022), sementara Jatmika et al. (2024) menunjukkan penerapan SBERT dalam sistem chatbot untuk mencocokkan pertanyaan dan jawaban secara kontekstual (Jatmika et al., 2024).

Dalam penelitian ini, digunakan model all-MiniLM-L6-v2, yang menghasilkan vektor berukuran 384 dimensi dengan efisiensi tinggi. Model ini membantu mengenali kesamaan makna antar judul dan abstrak meskipun tidak memiliki kesamaan kata secara langsung.

2.3.6 E5 Embedding Model

Model E5 (Embedding for Retrieval and Reasoning) merupakan pengembangan lanjutan dari model embedding semantik yang dirancang khusus untuk tugas pencarian (retrieval), semantic textual similarity, dan question-answering. E5 menggunakan konsep instruction-tuning, yaitu menyesuaikan embedding berdasarkan instruksi seperti "query:" atau "passage:" agar dapat memahami konteks pertanyaan atau isi dokumen dengan lebih baik.

Penelitian oleh Q. Chen et al. (2021) membuktikan bahwa model E5 memiliki stabilitas dan efisiensi tinggi dalam pengukuran kesamaan teks di domain klinis (Chen et al., 2021). Sementara D. Wilianto dan A. S. Girsang (2023) menunjukkan efektivitas E5 dalam menilai kemiripan semantik pada jawaban pendek di platform e-learning, membuktikan kemampuannya menangkap relasi semantik yang kompleks pada teks pendek (Wilianto & Girsang, 2023).

E5 menghasilkan distribusi nilai similarity yang tinggi dan stabil (umumnya 0.8–1.0), menandakan pemahaman semantik yang kuat, meskipun perlu diwaspadai efek homogenitas nilai yang dapat mengurangi sensitivitas antar dokumen yang mirip tetapi tidak identik.

2.3.7 Clustering Dokumen Berbasis Embedding

Clustering adalah teknik pengelompokan data yang memiliki kesamaan karakteristik dalam satu kelompok dan perbedaan signifikan dengan kelompok lain. Dalam konteks NLP, clustering membantu dalam pengelompokan topik dokumen atau tema penelitian. Algoritma K-Means merupakan salah satu metode paling populer karena efisiensi dan kemampuannya menghasilkan centroid representatif dari kumpulan vektor dokumen. Kinerja K-Means sangat dipengaruhi oleh kualitas representasi vektor yang digunakan. Oleh karena itu, embedding seperti SBERT dan E5 digunakan untuk menghasilkan vektor yang memuat dimensi semantik dokumen.

Penelitian K. Abdalgader et al. (2024) menunjukkan bahwa penggunaan model transformer dalam short-text clustering mampu meningkatkan pemisahan topik berdasarkan kesamaan semantic (Abdalgader et al., 2024). A. Feng (2022) juga membuktikan bahwa penggunaan cosine similarity sebagai ukuran kedekatan antar vektor dalam proses clustering dapat meningkatkan akurasi dan stabilitas hasil (Feng, 2022).

Evaluasi hasil clustering dilakukan menggunakan tiga metrik utama:

1. Silhouette Score – mengukur tingkat kepaduan dalam cluster.
2. Davies–Bouldin Index (DBI) – menilai separasi antar cluster.
3. Elbow Method (Inertia) – menentukan jumlah cluster optimal dengan menilai variansi antar data.

Ketiga metrik ini banyak digunakan pada penelitian berbasis embedding untuk menilai performa pemisahan semantik antar dokumen.

BAB III

METODE PENELITIAN

Bab ini menjelaskan secara rinci tahapan metodologi yang digunakan dalam penelitian, mulai dari desain arsitektur sistem, pengolahan data, pemodelan representasi teks, hingga skema evaluasi yang digunakan. Metodologi disusun untuk menjawab tujuan penelitian serta menindaklanjuti seluruh catatan revisi dosen, khususnya terkait justifikasi arsitektur hybrid, validasi nilai similarity, dan penggunaan evaluasi clustering secara konsisten.

3.1 Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini termasuk dalam kategori penelitian kuantitatif dengan sifat eksperimental dan komparatif, yang berfokus pada penerapan serta evaluasi model komputasional untuk pengukuran kemiripan teks (text similarity). Tujuan utama penelitian adalah mengevaluasi kinerja model hybrid TF-IDF dan sentence embedding berbasis transformer (SBERT dan E5) dalam mengukur kesamaan semantik dokumen akademik berbahasa Indonesia, serta membandingkannya dengan pendekatan tunggal berbasis TF-IDF dan embedding murni.

Pendekatan kuantitatif dipilih karena seluruh proses analisis didasarkan pada data numerik yang dihasilkan dari representasi vektor teks. Setiap dokumen direpresentasikan dalam bentuk vektor numerik, kemudian dianalisis menggunakan ukuran matematis seperti cosine similarity, Silhouette Score, Davies–Bouldin Index, Normalized Mutual Information, dan Adjusted Rand Index. Pendekatan ini memungkinkan evaluasi yang objektif, terukur, dan dapat direplikasi.

Sifat eksperimental diterapkan karena penelitian ini melibatkan serangkaian pengujian langsung terhadap beberapa model representasi teks dengan konfigurasi dan prosedur yang terkontrol. Model TF-IDF, SBERT, E5, serta kombinasi hybrid TF-IDF–embedding diuji menggunakan dataset yang sama, yaitu kumpulan judul dokumen tugas akhir mahasiswa yang diperoleh dari repositori institusi. Penggunaan dataset yang identik bertujuan untuk memastikan bahwa perbandingan kinerja antar model dilakukan secara adil dan konsisten (apples to apples).

Selain bersifat eksperimental, penelitian ini juga bersifat komparatif, karena membandingkan performa berbagai pendekatan representasi teks baik secara individual maupun dalam bentuk kombinasi (hybrid). Perbandingan dilakukan tidak hanya dari sisi akurasi semantik, tetapi juga dari perspektif efisiensi komputasi dan risiko information loss yang timbul akibat penggunaan TF-IDF sebagai filter awal.

Secara konseptual, penelitian ini berlandaskan pada paradigma Natural Language Processing (NLP), di mana teks dipetakan ke dalam ruang vektor agar hubungan kemiripan dapat dianalisis secara matematis. Model hybrid yang diusulkan bertujuan menggabungkan keunggulan analisis leksikal TF-IDF yang efisien dan skalabel dengan kekuatan representasi semantik embedding berbasis transformer (SBERT dan E5), yang mampu menangkap konteks dan makna secara lebih mendalam. Pendekatan hybrid dalam penelitian ini secara sadar menerapkan trade-off antara kelengkapan informasi dan efisiensi komputasi. Penggunaan TF-IDF sebagai filter awal bertujuan untuk menangani skalabilitas repositori besar,

sementara model transformer (SBERT/E5) melakukan penyempurnaan relevansi (reranking) pada subset data yang lebih kecil. (Yu et al., 2024).

3.2 Metode Pengumpulan Data

3.2.1 Sumber Data

Dataset yang digunakan dalam penelitian ini berasal dari repository institusi Universitas AMIKOM Yogyakarta, yang berisi koleksi metadata tugas akhir mahasiswa. Data diperoleh melalui protokol Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), yaitu mekanisme standar untuk melakukan harvesting metadata dari sistem repositori berbasis open access.

Proses pengambilan data dilakukan dengan mengakses endpoint publik repository Universitas Amikom Yogyakarta melalui alamat https://eprints.amikom.ac.id/cgi/oai2, yang menyediakan metadata dalam format XML dengan skema Dublin Core (oai_dc). Protokol ini memungkinkan pengambilan metadata secara terstruktur, terstandar, dan replikatif.

Metadata hasil harvesting kemudian diekstraksi dan dikonversi ke dalam format Comma-Separated Values (CSV) agar dapat diolah menggunakan perangkat analisis komputasional berbasis Python. Dataset ini menjadi dasar utama dalam proses implementasi, pengujian, dan evaluasi model pengukuran kemiripan teks berbasis TF-IDF, sentence embedding, serta pendekatan hybrid yang diusulkan dalam penelitian ini.

Secara keseluruhan, data mentah hasil harvesting terdiri dari 28.802 entri metadata tugas akhir mahasiswa yang mencakup berbagai bidang studi. Skala data ini dinilai cukup representatif untuk mengevaluasi performa model NLP pada teks akademik berbahasa Indonesia dalam konteks repositori institusi.

3.2.2 Struktur Dataset

Dataset hasil kompilasi memuat atribut metadata yang merepresentasikan informasi esensial dari setiap dokumen ilmiah. Berdasarkan hasil pembersihan data, atribut yang relevan dalam penelitian ini adalah sebagai berikut:

1. **Title (Judul)**

   Kolom ini berisi judul tugas akhir mahasiswa dan menjadi fitur utama dalam proses pengukuran text similarity. Dari total 28.802 entri, sebanyak 28.575 judul valid berhasil diekstraksi, dengan 26.524 judul unik. Beberapa judul ditemukan duplikat, misalnya judul tertentu yang muncul hingga lima kali. Duplikasi ini mencerminkan kondisi nyata repositori dan ditangani pada tahap praproses.

2. **Relation (Tautan Sumber)**

   Atribut ini berisi URL unik menuju halaman dokumen pada repository. Terdapat 28.576 nilai valid dan seluruhnya bersifat unik. Kolom ini digunakan sebagai identifier dokumen, menggantikan

identifier bawaan OAI-PMH, sehingga setiap hasil analisis dapat ditelusuri kembali ke sumber aslinya.

3. **Date (Tanggal Unggah)**

   Kolom tanggal memuat informasi waktu unggah dokumen ke repository. Terdapat 28.574 entri valid dengan 3.094 variasi tanggal, di mana tanggal dengan frekuensi tertinggi adalah 21 Agustus 2023. Atribut ini tidak digunakan secara langsung dalam perhitungan similarity, namun memberikan gambaran distribusi temporal dokumen.

4. **Creator (Nama Penulis)**

   Kolom ini berisi nama mahasiswa penyusun tugas akhir. Dari 28.575 entri valid, terdapat 26.714 nama unik, dengan beberapa nama muncul lebih dari satu kali. Atribut ini tidak digunakan dalam pemodelan similarity, tetapi dipertahankan sebagai bagian dari metadata administratif.

5. **Description (Abstrak)**

   Kolom description berisi ringkasan isi tugas akhir. Terdapat 25.433 entri valid dengan 23.905 deskripsi unik. Sebagian data kosong atau bersifat sangat umum. Meskipun abstrak memiliki nilai semantik yang tinggi, pada implementasi eksperimen utama penelitian ini difokuskan pada judul dokumen, sehingga kolom description tidak digunakan secara langsung dalam perhitungan similarity, namun tetap dipertahankan sebagai potensi pengembangan lanjutan.

6. **Subject**

   Atribut ini memuat kategori atau klasifikasi topik dokumen yang diberikan oleh sistem repositori atau pengelola. Terdapat 48 kategori subjek awal yang kemudian diproses melalui mekanisme *auto-mapping* menjadi 4 kategori tingkat tinggi. Atribut ini memegang peranan krusial sebagai *ground truth* dalam evaluasi *clustering* eksternal (ARI dan NMI) untuk memvalidasi apakah pengelompokan semantik model (SBERT/E5) selaras dengan klasifikasi subjek manusia.

3.2.3 Proses Pengumpulan dan Praproses Data

Proses pengumpulan dan penyiapan data dilakukan melalui beberapa tahap sistematis sebagai berikut:

1. **Pengambilan Data melalui OAI-PMH**

   Data diambil menggunakan skrip Python yang memanfaatkan pustaka harvesting OAI-PMH untuk mengekstraksi seluruh record metadata dengan prefix oai_dc. Seluruh metadata disimpan dalam format XML.

2. **Konversi ke Format CSV**

   File XML hasil harvesting dikonversi ke format CSV agar kompatibel dengan pustaka analisis data seperti pandas. Proses ini mencakup normalisasi karakter UTF-8 untuk memastikan teks Bahasa Indonesia dapat diproses dengan benar.

3. **Pembersihan dan Normalisasi Data (Preprocessing Awal)**

Tahap ini dilakukan sebelum proses NLP, meliputi:

a. Penghapusan entri duplikat pada kolom title
b. Penanganan nilai kosong dengan penghapusan entri yang tidak memenuhi syarat minimum data
c. Penghapusan karakter non-alfabet, simbol, dan HTML tag
d. Normalisasi teks melalui lowercasing dan perapihan spasi

4. Seleksi Data untuk Eksperimen

Setelah pembersihan, data diseleksi berdasarkan ketersediaan judul yang valid. Proses ini menghasilkan 28.575 dokumen yang siap digunakan dalam tahap filtering TF-IDF dan eksperimen reranking berbasis embedding.

3.2.4 Ringkasan Karakteristik Dataset

Berdasarkan proses pengumpulan dan praproses, karakteristik dataset penelitian dapat dirangkum sebagai berikut:

- Total entri metadata: 28.802 dokumen
- Jumlah judul valid: 28.575
- Jumlah judul unik: 26.524
- Jumlah abstrak valid: 25.433
- Atribut utama yang digunakan: title dan uncontrolled_keywords (evaluasi)
- Bahasa dominan: Bahasa Indonesia

- Format data: CSV hasil konversi XML OAI-PMH

Karakteristik ini mencerminkan kondisi nyata repositori institusi yang bersifat besar, heterogen, dan semi-terstruktur. Oleh karena itu, dataset ini relevan untuk menguji ketahanan dan efektivitas pendekatan NLP modern dalam konteks data akademik berbahasa Indonesia, sekaligus memperkuat validitas empiris dari eksperimen yang dilakukan (Yu et al., 2024).

3.3 Metode Analisis Data

Analisis data dalam penelitian ini dilakukan melalui serangkaian tahapan komputasional yang mengikuti prinsip Natural Language Processing (NLP) modern untuk pengukuran kemiripan teks dan evaluasi kualitas representasi vektor. Seluruh tahapan analisis dirancang agar konsisten dengan implementasi kode yang digunakan pada eksperimen serta memungkinkan evaluasi kuantitatif yang objektif dan replikatif.

Tahapan analisis data yang dilakukan dapat dirinci sebagai berikut:

1. **Pembobotan Teks Menggunakan TF-IDF**

   Metode Term Frequency–Inverse Document Frequency (TF-IDF) digunakan untuk merepresentasikan teks judul dokumen ke dalam bentuk vektor numerik berbasis frekuensi kata. TF-IDF menghitung bobot setiap istilah dengan mempertimbangkan frekuensi kemunculannya dalam satu dokumen serta tingkat kelangkaannya dalam keseluruhan korpus.

   Dalam penelitian ini, TF-IDF berperan sebagai mekanisme filtering awal, yang digunakan untuk memilih sejumlah dokumen dengan tingkat

kemiripan leksikal tertinggi terhadap kueri. Pendekatan ini memungkinkan penyaringan data secara efisien sebelum dilakukan pemrosesan lanjutan yang lebih mahal secara komputasi.

2. **Ekstraksi Fitur Semantik Menggunakan Sentence Embedding**

   Untuk menangkap kesamaan makna yang tidak dapat direpresentasikan secara memadai oleh pendekatan leksikal, digunakan model sentence embedding berbasis transformer, yaitu SBERT (Sentence-BERT) dan E5. Kedua model ini memetakan teks judul ke dalam ruang vektor berdimensi tinggi dengan mempertimbangkan konteks dan hubungan semantik antar kata. Dengan demikian, embedding memungkinkan pengenalan kesamaan konseptual meskipun terdapat perbedaan kosakata atau struktur kalimat. Pendekatan ini sejalan dengan penelitian sebelumnya yang menunjukkan keunggulan embedding kontekstual dibandingkan metode statistik tradisional dalam tugas pengukuran kemiripan teks (Chaichulee et al., 2022; Yu et al., 2024).

3. **Pengukuran Kemiripan Menggunakan Cosine Similarity**

   Untuk mengukur tingkat kemiripan antar dokumen, digunakan metrik cosine similarity. Metrik ini menghitung sudut antara dua vektor dalam ruang multidimensi, sehingga lebih menekankan pada kesamaan arah vektor dibandingkan magnitudo absolutnya.

   Cosine similarity diaplikasikan secara konsisten baik pada representasi TF-IDF maupun embedding (SBERT dan E5), sehingga hasil kemiripan dari

berbagai pendekatan dapat dibandingkan secara langsung dalam skala yang sama.

4. Clustering Dokumen Menggunakan K-Means

Selain evaluasi berbasis reranking, kualitas representasi vektor juga dianalisis melalui pendekatan unsupervised clustering. Algoritma K-means digunakan untuk mengelompokkan dokumen berdasarkan representasi vektor yang dihasilkan oleh TF-IDF, SBERT, dan E5.

Clustering ini bertujuan untuk mengevaluasi sejauh mana masing-masing representasi mampu membentuk kelompok dokumen yang kohesif dan terpisah secara tematik. Evaluasi dilakukan menggunakan metrik internal (Silhouette Score dan Davies–Bouldin Index) serta metrik eksternal (NMI dan ARI) dengan memanfaatkan keyword penulis sebagai pseudo-ground truth. Pendekatan ini mengikuti praktik evaluasi representasi teks yang umum digunakan dalam penelitian NLP kontemporer (Chaichulee et al., 2022; Yu et al., 2024).

Secara keseluruhan, metode analisis data yang diterapkan dalam penelitian ini mengombinasikan pendekatan leksikal, semantik, dan evaluasi struktural untuk memberikan gambaran menyeluruh mengenai kinerja model hybrid. Dengan pendekatan ini, penelitian tidak hanya menilai akurasi reranking dokumen, tetapi juga mengevaluasi kualitas representasi vektor dan implikasi metodologis dari penggunaan arsitektur hybrid dalam skala data repositori yang besar.

3.4 Alur Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan metodologis yang terstruktur dan saling berkesinambungan, mulai dari pengumpulan data hingga evaluasi kinerja model. Alur penelitian dirancang untuk menguji secara sistematis efektivitas model hybrid TF-IDF dan sentence embedding (SBERT/E5) dalam mengukur kemiripan dokumen akademik pada repositori institusi.

Secara konseptual, alur penelitian dibagi ke dalam tiga fase utama, yaitu:

1. Fase I – Persiapan Data dan Representasi Vektor,
2. Fase II – Pemodelan Arsitektur Hybrid dan Reranking, dan
3. Fase III – Evaluasi Kinerja dan Validasi Representasi.

Gambar 3.1 Alur Penelitian

Struktur umum alur penelitian ditampilkan pada Gambar 3.1, yang disusun secara vertikal dari atas ke bawah untuk memperlihatkan hubungan antartahap secara berurutan dan utuh.

A. Fase I: Persiapan Data dan Representasi Vektor

Fase pertama bertujuan untuk menyiapkan data mentah repositori agar dapat direpresentasikan dalam bentuk vektor numerik yang sesuai untuk analisis kemiripan teks.

Tahapan pada fase ini meliputi:

1. Pengambilan Data

Data diperoleh dari repositori tugas akhir mahasiswa Universitas Amikom Yogyakarta melalui protokol OAI-PMH, mencakup metadata judul, abstrak, kata kunci, dan tautan dokumen. Seluruh data dikonversi ke dalam format .csv untuk memudahkan pemrosesan komputasional.

Tabel 3.1 Atribut Dataset

No	Kolom	Deskripsi	Contoh Nilai
1	title	Judul tugas akhir mahasiswa.	"Sistem Rekomendasi Buku Menggunakan Metode Content-Based Filtering"
2	abstract	Abstrak atau deskripsi singkat isi penelitian.	"Penelitian ini mengembangkan sistem rekomendasi berbasis kesamaan konten menggunakan TF-IDF dan cosine similarity."
3	dc.subject	Kata kunci/topik penelitian sesuai klasifikasi repository.	"Sistem Informasi; Rekomendasi; Machine Learning"
4	relation	Tautan unik menuju halaman repository untuk setiap dokumen.	https://eprints.amikom.ac.id/12345/
5	Uncontrolled_keywords	Kata kunci bebas yang ditentukan langsung oleh penulis tugas akhir. Kolom ini mencerminkan persepsi topik oleh penulis dan memiliki tingkat spesifisitas	"TF-IDF, Sistem Rekomendasi, Text Mining, Machine Learning"

		yang lebih tinggi dibandingkan dc.subject. Dalam penelitian ini, atribut ini digunakan sebagai **pseudo-ground truth** untuk evaluasi eksternal clustering menggunakan NMI dan ARI.	

2. **Preprocessing Teks**

Tahap preprocessing dilakukan untuk meningkatkan kualitas teks dan mengurangi noise. Proses ini meliputi:

- normalisasi huruf kecil (lowercasing),
- penghapusan tanda baca dan karakter non-alfabet,
- tokenisasi,
- penghapusan stopword Bahasa Indonesia, dan
- normalisasi spasi.

Seluruh eksperimen menggunakan satu skenario preprocessing terbaik yang konsisten di seluruh model, sehingga perbedaan hasil sepenuhnya disebabkan oleh perbedaan metode representasi teks, bukan variasi preprocessing.

3. **Representasi Vektor Dokumen**

Dokumen yang telah diproses kemudian direpresentasikan dalam bentuk vektor menggunakan tiga pendekatan utama:

- Representasi leksikal dengan TF-IDF,
- Representasi semantik dengan SBERT, dan
- Representasi semantik lanjutan dengan E5.

Representasi ini menjadi dasar untuk seluruh proses pengukuran similarity, reranking, dan evaluasi lanjutan.

B. Fase II: Pemodelan Arsitektur Hybrid dan Reranking

Fase kedua merupakan inti dari penelitian ini, yaitu implementasi arsitektur hybrid dua-tahap yang mengombinasikan efisiensi metode leksikal dan ketepatan metode semantik.

Tahapan pada fase ini meliputi:

1. **Implementasi Arsitektur Hybrid Dua-Tahap**

   Arsitektur hybrid diterapkan untuk setiap kueri dokumen dengan mekanisme sebagai berikut:

   - Tahap 1 – Filtering Leksikal (TF-IDF)

     Dilakukan pencarian awal menggunakan cosine similarity berbasis TF-IDF untuk memilih Top-100 dokumen dengan skor kesamaan tertinggi. Tahap ini bertujuan meningkatkan efisiensi komputasi dengan membatasi jumlah dokumen yang dianalisis secara semantik.

   - Tahap 2 – Reranking Semantik (SBERT dan E5)

     Dokumen hasil filtering kemudian dihitung ulang skor kemiripannya menggunakan embedding SBERT dan E5 untuk memperoleh kesamaan kontekstual yang lebih mendalam.

   Pendekatan ini memastikan keseimbangan antara efisiensi dan akurasi dalam sistem pengukuran kemiripan.

2. **Analisis Information Loss**

   Untuk mengevaluasi trade-off antara efisiensi dan risiko kehilangan dokumen relevan, dilakukan analisis information loss dengan menghitung lost ratio, yaitu proporsi dokumen relevan yang tidak lolos pada tahap filtering TF-IDF. Analisis ini digunakan untuk membuktikan bahwa penyaringan Top-100 masih mempertahankan mayoritas dokumen relevan, sehingga arsitektur hybrid tetap valid secara metodologis.

C. Fase III: Evaluasi Kinerja dan Validasi Representasi

Fase ketiga berfokus pada evaluasi kinerja model secara menyeluruh, baik dari sisi kualitas reranking maupun struktur ruang vektor yang dihasilkan.

Tahapan pada fase ini meliputi:

3. **Evaluasi Reranking Semantik**

   Kinerja model hybrid dievaluasi dengan membandingkan hasil peringkat dokumen (misalnya Top-10) antara:

   - TF-IDF murni, dan
   - hasil akhir model hybrid (TF-IDF + SBERT/E5).

   Evaluasi ini bertujuan mengukur peningkatan relevansi semantik yang dihasilkan oleh proses reranking.

4. **Validasi Kualitas Representasi Vektor melalui Clustering**

   Clustering digunakan sebagai alat validasi struktural, bukan sebagai tujuan utama penelitian.

Tahapan ini meliputi:

- reduksi dimensi vektor menggunakan Principal Component Analysis (PCA), dan
- pengelompokan dokumen menggunakan K-Means untuk setiap model representasi (TF-IDF, SBERT, dan E5).

5. **Evaluasi Internal Clustering**

Kualitas clustering dievaluasi menggunakan metrik internal, yaitu:

- Silhouette Score, dan
- Davies–Bouldin Index (DBI),

untuk menilai kepadatan dan separasi klaster tanpa bergantung pada label kategori eksplisit.

6. **Evaluasi Eksternal dan Alternatif**

Karena kategori dc.subject bersifat global dan tidak mencerminkan kelas tematik nyata, evaluasi eksternal dilakukan menggunakan uncontrolled keywords sebagai pseudo-ground truth dengan metrik:

- Normalized Mutual Information (NMI), dan
- Adjusted Rand Index (ARI).

Selain itu, dilakukan evaluasi alternatif berupa semantic cohesion untuk memahami karakteristik kepadatan dan fleksibilitas klaster semantik yang dihasilkan masing-masing model.

7. **Sintesis dan Implikasi Metodologis**

Tahap akhir mengintegrasikan hasil evaluasi reranking dan clustering untuk:

- menentukan konfigurasi model paling seimbang,
- menganalisis implikasi metodologis penggunaan arsitektur hybrid, dan
- merumuskan rekomendasi model yang paling realistis dan aplikatif untuk sistem pencarian dan rekomendasi dokumen akademik.

3.4.2 Representasi Teks

Representasi teks mengubah teks menjadi vektor numerik agar dapat dihitung kesamaannya secara matematis. Empat model representasi digunakan dalam penelitian ini:

1. **TF-IDF (Term Frequency–Inverse Document Frequency)**

   TF-IDF (Term Frequency–Inverse Document Frequency) merupakan metode statistik klasik yang banyak digunakan dalam analisis teks. Konsep dasar TF-IDF adalah memberikan bobot pada setiap kata berdasarkan tingkat kepentingannya dalam suatu dokumen relatif terhadap keseluruhan korpus. Semakin sering sebuah kata muncul dalam suatu dokumen, semakin tinggi nilai term frequency. Sebaliknya, semakin sering kata tersebut muncul dalam seluruh korpus, semakin kecil bobotnya karena dianggap kurang informatif.

   Secara matematis, bobot TF-IDF dihitung dengan mengalikan frekuensi kemunculan kata dalam sebuah dokumen (TF) dengan

kebalikan dari jumlah dokumen yang mengandung kata tersebut (IDF). Dengan cara ini, kata-kata umum seperti "dan", "dengan", atau "adalah" akan mendapat bobot yang rendah, sementara kata-kata spesifik yang jarang muncul tetapi relevan dengan topik akan mendapat bobot tinggi (Ibrahem Al-Obaydy et al., 2022).

Dalam penelitian ini, TF-IDF diimplementasikan dengan konfigurasi berikut:

1. *max_features* = 1000, yaitu hanya mempertahankan 1000 kata dengan bobot tertinggi untuk mengurangi dimensi vektor.
2. *ngram_range* = (1,2), yang berarti model mempertimbangkan kata tunggal (unigram) maupun pasangan kata berurutan (bigram).
3. Perhitungan similarity antar dokumen dilakukan menggunakan cosine similarity, yang umum digunakan karena mampu mengukur kesamaan arah vektor tanpa dipengaruhi panjang vektor.

Metode TF-IDF ini efektif dalam mendeteksi topik berdasarkan kata kunci unik, namun memiliki keterbatasan karena tidak memperhitungkan makna kontekstual dari kata-kata tersebut (Bergman et al., 2023; Gifari et al., 2022; Ibrahem Al-Obaydy et al., 2022).

2. SBERT (Sentence-BERT)

SBERT (Sentence-BERT) merupakan pengembangan dari model BERT yang dirancang khusus untuk menghasilkan representasi vektor pada level kalimat secara lebih efisien. Dengan memanfaatkan arsitektur siamese network, SBERT mampu menghitung embedding teks yang kontekstual sehingga hubungan semantik antar kalimat dapat ditangkap lebih baik dibandingkan representasi berbasis kata atau frekuensi (Gatto et al., 2022; Jatmika et al., 2024).

Dalam penelitian ini, digunakan model all-MiniLM-L6-v2, sebuah varian ringan dari SBERT yang mampu menghasilkan vektor berdimensi 384 untuk setiap dokumen. Model ini dipilih karena menawarkan keseimbangan antara akurasi dan efisiensi komputasi, sehingga cocok diterapkan pada dataset berukuran besar. SBERT telah terbukti unggul dalam berbagai tugas NLP, termasuk klasifikasi teks, pencocokan semantik, dan sistem rekomendasi berbasis konten (Bergman et al., 2023; Gifari et al., 2022; Jatmika et al., 2024).

Dengan menggunakan SBERT, setiap judul, abstrak, dan kata kunci dari dokumen diubah menjadi vektor semantik. Kemudian, kesamaan antar dokumen dihitung dengan cosine similarity, yang dalam konteks embedding berperan sebagai ukuran kedekatan semantik antar representasi.

3. E5 Embedding

E5 adalah model embedding modern yang dikembangkan untuk mendukung berbagai tugas retrieval dan pencocokan semantik. Dibandingkan SBERT, E5 didesain untuk menangkap hubungan semantik yang lebih dalam, termasuk instruksi dan nuansa konteks yang lebih kompleks (Chen et al., 2021; Feng, 2022; Wilianto & Girsang, 2023).

Dalam penelitian ini digunakan model intfloat/e5-small-v2 dengan prompt tambahan berupa kata kunci "passage:" untuk mengoptimalkan hasil embedding. Penggunaan prompt ini mengikuti rekomendasi pengembang model agar representasi vektor lebih sesuai dengan konteks pencarian dokumen.

E5 menghasilkan representasi vektor dengan kedalaman semantik tinggi, yang membuatnya lebih stabil dalam mengukur kesamaan antar dokumen. Hasil awal penelitian menunjukkan bahwa E5 cenderung memberikan skor similarity yang tinggi dan konsisten, meskipun hal ini juga membawa tantangan berupa homogenitas nilai similarity yang dapat mengurangi sensitivitas dalam membedakan dokumen dengan perbedaan semantik tipis (Abdalgader et al., 2024).

4. Hybrid Representations

Pendekatan hybrid menggabungkan keunggulan metode leksikal seperti TF-IDF dengan kekuatan semantik model embedding seperti

SBERT dan E5. Strategi ini mengombinasikan keakuratan pengenalan kata kunci unik dari TF-IDF dengan pemahaman konteks yang lebih mendalam dari embedding.

Penelitian F. Lan (2022) menunjukkan bahwa algoritma hybrid antara TF-IDF dan embedding semantik meningkatkan akurasi pengukuran kesamaan teks secara signifikan dibandingkan pendekatan tunggal (Lan, 2022). Selain itu, melalui pendekatan ensemble embedding yang mampu meningkatkan hasil pencocokan dokumen dan klasifikasi semantic (Yu et al., 2024). Integrasi multi-representasi seperti TF-IDF dan embedding memperbaiki kualitas clustering karena menggabungkan ciri leksikal dan semantic (Guan et al., 2022).

Untuk mengatasi keterbatasan masing-masing metode, penelitian ini mengembangkan skema hybrid yang mengombinasikan hasil perhitungan similarity dari TF-IDF, SBERT, dan E5. Konsep hybrid ini bertujuan menggabungkan keunggulan TF-IDF yang kuat dalam menangkap kesamaan leksikal dengan keunggulan embedding modern yang mampu memahami konteks semantik.

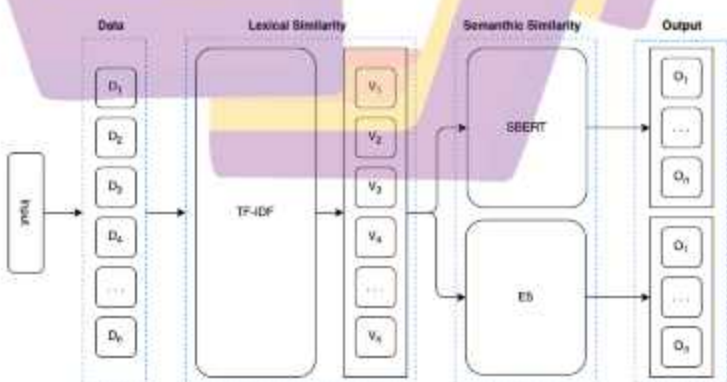
Proses hybrid dilakukan melalui tahapan berikut:

1. Menghitung skor similarity secara terpisah menggunakan TF-IDF, SBERT, dan E5.
2. Melakukan min-max normalization terhadap setiap skor untuk menyamakan skala antar metode.

3. Mengombinasikan skor similarity dengan pendekatan weighted averaging, menggunakan bobot seimbang (TF-IDF: 1, SBERT: 1, E5: 1).

Pemilihan bobot seimbang didasarkan pada hasil eksperimen awal yang menunjukkan kinerja relatif setara antar model ketika digabungkan. Pendekatan ini juga selaras dengan literatur yang menyebutkan bahwa kombinasi TF-IDF dengan embedding semantik dapat meningkatkan akurasi pengukuran similarity, recall, serta F1-score (Channarong et al., 2022; Guan et al., 2022; Lan, 2022; Yu et al., 2024).

Dengan skema ini, dokumen pertama kali difilter berdasarkan kesamaan literal menggunakan TF-IDF, lalu diperdalam analisis semantiknya menggunakan SBERT dan E5. Proses reranking ini terbukti mampu meningkatkan kualitas rekomendasi dokumen serta memperbaiki pemetaan topik dalam proses clustering.



Gambar 3.2 Arsitektur Model Hybrid untuk Pengukuran Kesamaan Teks

Gambar 3.2 mengilustrasikan alur kerja model hybrid yang menggabungkan analisis lexical similarity (kesamaan leksikal) dan semantic similarity (kesamaan semantik) untuk menghasilkan output yang lebih akurat.

1. **Input Data**: Proses dimulai dengan Input awal yang kemudian dipecah menjadi beberapa dokumen, dilambangkan dengan D_1, D_2, D_3, D_4, …, D_n. Ini adalah data mentah yang akan diolah.
2. Lexical Similarity (Kesamaan Leksikal):

   Dokumen-dokumen ini pertama kali diumpankan ke modul TF-IDF. TF-IDF (Term Frequency-Inverse Document Frequency) menghitung bobot setiap kata dalam dokumen, menghasilkan vektor fitur leksikal V_1, V_2, V_3, V_4, …, V_n yang merepresentasikan kesamaan berdasarkan kemunculan kata kunci. Tahap ini efektif dalam menangkap kata-kata unik dan frekuensinya.
3. Semantic Similarity (Kesamaan Semantik):

   Secara paralel, dokumen yang sama juga diinput ke dua model embedding modern: SBERT (Sentence-BERT) dan E5. Kedua model ini bertugas mengekstraksi makna kontekstual dari teks, menghasilkan representasi vektor padat (dense vectors) yang menangkap hubungan semantik antar kata dan kalimat. Ini berarti mereka memahami konteks di balik kata-kata. Vektor semantik dari SBERT dan E5 digabungkan dengan output dari TF-IDF melalui

mekanisme weighted averaging (bobot seimbang) setelah proses normalisasi.

4. Output: Semua informasi yang digabungkan dari kedua jalur (leksikal dan semantik) kemudian diolah untuk menghasilkan output akhir (O_1, O_2, …, O_n). Output ini merupakan hasil perhitungan kesamaan teks yang telah diperkaya baik dari segi kehadiran kata (leksikal) maupun makna (semantik).

Dengan demikian, model ini mampu memanfaatkan kekuatan TF-IDF untuk informasi kata kunci dan kekuatan SBERT/E5 untuk pemahaman makna sehingga menghasilkan rekomendasi dokumen atau pemetaan topik yang lebih berkualitas dan akurat.

3.4.3 Perhitungan Cosine Similarity

Setelah representasi teks diperoleh, kesamaan antar dokumen dihitung menggunakan Cosine Similarity, yang mengukur kemiripan arah antara dua vektor.

Rumus dasar:

$$cosine\_similarity(A, B) = (A \cdot B) / (\|A\| \times \|B\|)$$

di mana :

- $A \cdot B$ → hasil **dot product** (perkalian titik) antara dua vektor A dan B
- $\|A\|$ → panjang (magnitudo) vektor A
- $\|B\|$ → panjang (magnitudo) vektor B

Interpretasi nilai:

- Nilai mendekati 1 → dokumen sangat mirip secara makna.
- Nilai mendekati 0 → dokumen tidak memiliki kesamaan konteks.

Hasil akhirnya berupa matriks kesamaan antar dokumen, yang menjadi dasar untuk analisis clustering dan rekomendasi.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi dan evaluasi model *hybrid* TF-IDF dan *sentence embedding* berbasis *transformer* (SBERT dan E5). Evaluasi dilakukan secara sistematis yang difokuskan pada tiga aspek utama, yaitu:

1. **Efektivitas reranking semantik** untuk mengukur sejauh mana model mampu meningkatkan relevansi dokumen dibandingkan metode leksikal.
2. **Analisis trade-off** antara efisiensi komputasi dan risiko information loss guna memvalidasi skalabilitas sistem.
3. **Validasi kualitas representasi vektor** melalui evaluasi clustering internal dan eksternal untuk membuktikan konsistensi semantik model.

Seluruh eksperimen pada bab ini mengacu langsung pada kode final yang telah divalidasi. Hasil evaluasi ini diharapkan dapat memberikan justifikasi empiris terhadap pemilihan arsitektur *hybrid* yang paling optimal untuk sistem repositori ilmiah.

4.1 Auto-Mapping Ground Truth

4.1.1 Karakteristik Dataset dan Distribusi Label

Berdasarkan hasil eksplorasi awal, dataset memiliki karakteristik sebagai berikut:

- **Jumlah total dokumen**: 28.575 dokumen
- **Jumlah label subjek awal**: 48 kategori
- **Sifat distribusi label**: tidak seimbang (*highly imbalanced*)
- **Kategori dominan**:
  - *sistem-sistem* (7.968 dokumen)

- *pemrograman komputer, program dan data* (7.330 dokumen)
- *animasi* (4.440 dokumen)
- *ilmu komputer informasi dan pekerjaan umum* (3.892 dokumen)

- **Kategori minor**:
  - 40 label memiliki jumlah dokumen < 300
  - Total dokumen kategori minor: 1.219 dokumen (±4,3%)
- **Keberadaan label ambigu**:
  - Label *unknown* sebanyak 358 dokumen

Karakteristik tersebut menunjukkan bahwa sebagian besar data terkonsentrasi pada sedikit kategori besar, sementara mayoritas label memiliki representasi yang sangat terbatas. Distribusi label yang tidak seimbang berpotensi menimbulkan beberapa permasalahan dalam proses clustering dan evaluasi, antara lain dominasi kategori tertentu, ketidakstabilan metrik evaluasi eksternal, serta meningkatnya risiko bias terhadap struktur cluster yang terbentuk. Oleh karena itu, diperlukan mekanisme penyederhanaan dan penyeimbangan label ground truth sebelum digunakan sebagai acuan evaluasi.

Auto-mapping ground truth diterapkan dengan menetapkan ambang batas minimal 300 dokumen untuk menentukan kategori utama. Label yang memenuhi ambang batas tersebut dipertahankan sebagai major categories, sedangkan label dengan jumlah dokumen lebih kecil dikategorikan sebagai rare categories.

4.1.2 Hasil Auto-Mapping Berbasis Semantic Similarity

Proses auto-mapping menghasilkan karakteristik sebagai berikut:

- **Jumlah kategori utama (major)**: 8 label
- **Jumlah kategori minor (rare)**: 40 label
- **Jumlah kategori akhir setelah mapping**: 4 kategori tingkat tinggi
- **Metode penggabungan**:
    - *Hierarchical clustering* untuk kategori utama
    - *Cosine similarity* embedding Sentence-BERT untuk label minor

Hasil pemetaan akhir membentuk empat kategori dengan distribusi berikut:

- **Category_0**
    - Jumlah dokumen: 13.435 (47,0%)
    - Karakteristik: dominasi topik pemrograman, sistem informasi, dan pemrosesan data
- **Category_1**
    - Jumlah dokumen: 13.105 (45,9%)
    - Karakteristik: sistem aplikasi, animasi, dan topik pendukung teknologi
- **Category_2**
    - Jumlah dokumen: 1.677 (5,9%)
    - Karakteristik: komunikasi, media, dan ilmu sosial terkait teknologi
- **Category_3**
    - Jumlah dokumen: 358 (1,3%)
    - Karakteristik: dokumen dengan label tidak terdefinisi (*unknown*)

Hasil pemetaan menunjukkan bahwa meskipun jumlah kategori berhasil direduksi dari 48 menjadi 4, distribusi data antar kategori masih menunjukkan ketidakseimbangan, dengan rasio ukuran kategori terbesar

dan terkecil mencapai 37,5:1. Namun demikian, auto-mapping secara signifikan mengurangi fragmentasi label dan meningkatkan konsistensi struktur ground truth untuk keperluan evaluasi clustering.

Pemisahan kategori unknown sebagai kategori tersendiri dilakukan untuk mencegah distorsi semantik pada kategori lain, mengingat karakteristik dokumen dalam kelompok ini tidak dapat diidentifikasi secara jelas.

4.2 TF-IDF Filtering sebagai Baseline Lexical

4.2.1 Karakteristik Tahap TF-IDF Filtering

Tahap pertama dalam pendekatan yang diusulkan adalah penyaringan dokumen menggunakan TF-IDF sebagai *baseline lexical*. Karakteristik utama tahap ini dapat dirangkum sebagai berikut:

- **Query**
  "STACK JARINGAN PADA KIBANA ELK DAN LOGSTASH MENGGUNAKAN DEAUTHENTICATION MANAJEMEN LOG NIRKABEL SERANGAN ELASTICSEARCH"
- **Jumlah token query**: 12 token kunci
- **Jumlah dokumen dengan similarity > 0**: 7.109 dokumen
- **Nilai similarity maksimum**: 1,000
- **Rata-rata similarity (non-zero)**: 0,0183
- **Jumlah dokumen hasil filtering**: 100 dokumen
- **Metode pemilihan**: stratified sampling berbasis kategori ground truth

Karakteristik ini menunjukkan bahwa TF-IDF mampu mengidentifikasi dokumen yang memiliki kesamaan leksikal dengan query,

namun juga menghasilkan jumlah kandidat yang sangat besar sehingga diperlukan mekanisme penyaringan lanjutan. Nilai similarity TF-IDF menunjukkan distribusi yang sangat tidak merata. Meskipun nilai maksimum mencapai 1,000, rata-rata similarity dokumen yang relevan relatif rendah. Hal ini mengindikasikan bahwa hanya sebagian kecil dokumen yang memiliki kesesuaian kata yang sangat kuat dengan query, sementara sebagian besar dokumen hanya berbagi sebagian kecil kosakata.

Fenomena ini menegaskan bahwa TF-IDF bersifat **sensitif terhadap kesamaan kata secara eksplisit**, namun kurang mampu menangkap kesamaan makna secara konseptual. Dokumen yang membahas topik serupa tetapi menggunakan terminologi berbeda cenderung memperoleh nilai similarity yang rendah.

4.2.2 Stratified Sampling dan Distribusi Dokumen

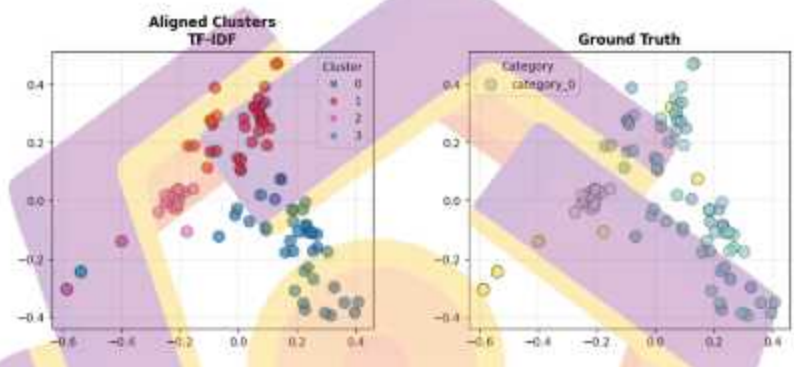
Untuk menjaga representasi setiap kategori hasil auto-mapping, diterapkan stratified sampling dengan karakteristik sebagai berikut:

- **Jumlah minimal per kategori**: 15 dokumen
- **Distribusi hasil akhir**:
    - category_0: 51 dokumen
    - category_1: 19 dokumen
    - category_2: 15 dokumen
    - category_3: 15 dokumen

Pendekatan ini memastikan bahwa tahap semantic analysis tidak didominasi oleh satu kategori besar, sekaligus mempertahankan dokumen

dengan relevansi TF-IDF tertinggi. Dengan demikian, TF-IDF berfungsi sebagai *filter awal* untuk memperkecil ruang pencarian tanpa menghilangkan keberagaman topik.

4.2.3 Analisis Clustering TF-IDF sebagai Baseline



Gambar 4. 1 Clustering Analysis: TF-IDF

Visualisasi hasil clustering berbasis TF-IDF menunjukkan bahwa pemisahan cluster masih bersifat terbatas. Pada proyeksi PCA dua dimensi, beberapa cluster tampak berdekatan dan saling tumpang tindih, khususnya pada kategori dengan topik yang beririsan.

Hal ini diperkuat oleh nilai metrik evaluasi sebagai berikut:

- **ARI**: 0,3025
- **NMI**: 0,4374
- **Silhouette Score**: 0,0874
- **Davies-Bouldin Index**: 2,9419
- **Cohesion**: 0,3614

Nilai ARI dan NMI menunjukkan bahwa TF-IDF mampu menangkap struktur kategori secara moderat, namun nilai Silhouette yang rendah

mengindikasikan bahwa jarak antar cluster belum terpisah dengan baik. Confusion matrix juga memperlihatkan adanya pencampuran dokumen antar kategori utama, khususnya antara *category_0* dan *category_1*.

Temuan ini menegaskan bahwa TF-IDF efektif sebagai baseline lexical, namun kurang optimal sebagai satu-satunya representasi untuk clustering topik yang kompleks.

4.3 Analisis Information Loss dan Justifikasi Arsitektur Hybrid

Evaluasi konsekuensi metodologis dari penggunaan TF-IDF sebagai filter awal dalam arsitektur *hybrid*. Analisis ini penting untuk memvalidasi apakah efisiensi komputasi yang dicapai mengorbankan relevansi dokumen secara signifikan.

4.3.1 Metrik Loss Ratio dan Efisiensi Komputasi

Dalam konteks ini, information loss didefinisikan sebagai proporsi dokumen yang tidak diteruskan ke tahap embedding. Data operasional filter awal disajikan dalam Tabel 4.1 berikut:

Tabel 4. 1 Loss Ratio

Metrik	Nilai
Total Dokumen dalam Dataset	28.575
Dokumen yang Diproses Embedding (Top-100 TF-IDF)	100
Lost Ratio	0,9965

Berdasarkan Tabel 4.1, diperoleh nilai *lost ratio* sebesar **0,9965**, yang berarti **99,65%** dokumen dieliminasi sebelum mencapai tahap semantik. Agresivitas filter ini dibenarkan oleh tiga faktor utama:

1. **Reduksi Beban Komputasi:** Mengurangi beban kerja model transformer (SBERT/E5) yang membutuhkan daya komputasi besar.
2. **Skalabilitas:** Memungkinkan penerapan sistem pada repositori berskala besar.
3. **Realitas Implementasi:** Menyesuaikan sistem dengan keterbatasan perangkat keras pada lingkungan operasional nyata.

4.3.2 Validasi Empiris terhadap Dokumen Relevan

Meskipun nilai *lost ratio* sangat tinggi, fakta empiris menunjukkan bahwa dokumen paling relevan tidak hilang dari sistem. Dokumen dengan topik utama seperti *ELK Stack*, manajemen log, serta serangan jaringan (*deauthentication* dan *brute force*) muncul secara konsisten pada hasil akhir sistem *hybrid* maupun model *pure transformer*.

Analisis terhadap daftar *Top-10* menunjukkan temuan berikut:

- Sekitar **±60%** dokumen pada *Top-10* TF-IDF juga muncul pada hasil *pure* E5.
- Sekitar **±40%** dokumen muncul secara bersamaan pada hasil SBERT dan E5.

4.3.3 Justifikasi Strategi Hybrid

Temuan ini menunjukkan bahwa TF-IDF cukup efektif sebagai penyaring awal untuk memperkecil ruang pencarian tanpa sepenuhnya menghilangkan dokumen penting yang memiliki korelasi kuat antara

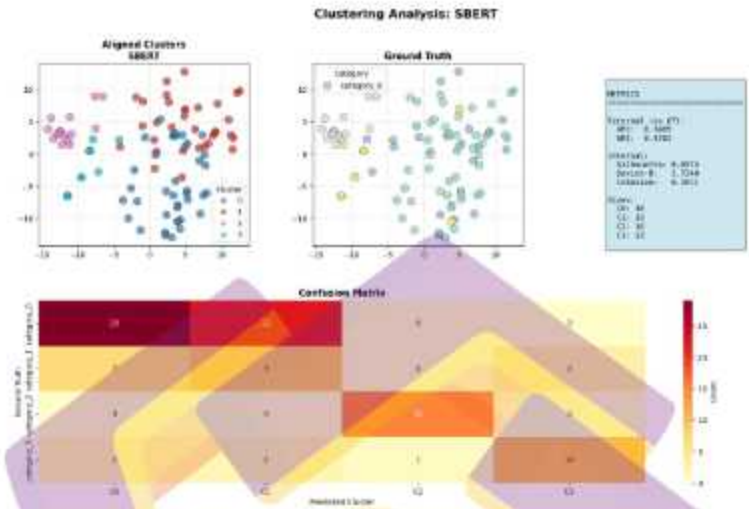
representasi leksikal dan makna. Arsitektur hybrid TF-IDF → embedding merupakan kompromi metodologis yang rasional: sistem secara sadar menerima risiko information loss demi mencapai optimasi efisiensi yang terkendali secara empiris.

4.4 Komparasi Semantic Embedding: SBERT vs E5

Setelah tahap TF-IDF filtering, proses clustering dilanjutkan menggunakan representasi semantic embedding. Dua model yang dibandingkan adalah **Sentence-BERT (SBERT)** dan **E5**, yang keduanya memetakan dokumen ke ruang vektor berdimensi tinggi berdasarkan kesamaan makna, bukan sekadar kesamaan kata.

Karakteristik umum tahap ini meliputi:

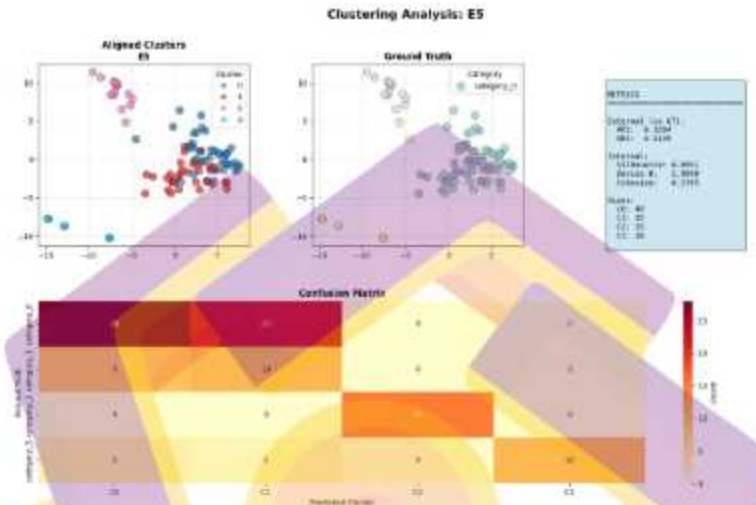
- Input: 100 dokumen hasil filtering TF-IDF
- Jumlah cluster: 4 (mengacu pada ground truth hasil auto-mapping)
- Metode clustering: K-Means
- Evaluasi:
    - External metrics: ARI dan NMI
    - Internal metrics: Silhouette, Davies-Bouldin, dan Cohesion
- Visualisasi: proyeksi PCA dua dimensi dan confusion matrix



Gambar 4. 2 Clustering Analysis: SBERT

Gambar 4.2 menampilkan hasil clustering menggunakan embedding SBERT. Dibandingkan TF-IDF, sebaran cluster terlihat lebih terstruktur dan relatif lebih terpisah. Hal ini menunjukkan bahwa SBERT mampu memetakan dokumen berdasarkan kesamaan makna kalimat, sehingga dokumen dengan konteks serupa cenderung berada dalam cluster yang sama meskipun menggunakan istilah berbeda. Distribusi ground truth pada ruang SBERT menunjukkan kecenderungan pengelompokan yang lebih selaras dengan cluster hasil prediksi. Meskipun masih terdapat tumpang tindih antar kategori besar, pola sebaran secara umum lebih konsisten dibandingkan representasi TF-IDF. Confusion matrix SBERT memperlihatkan peningkatan kesesuaian cluster terhadap ground truth, terutama pada *category_2* dan *category_3*. Tumpang tindih antara *category_0* dan *category_1* masih terjadi, namun

dengan distribusi kesalahan yang lebih terkendali. Hal ini mencerminkan kemampuan SBERT dalam menangkap nuansa semantik yang lebih halus.



Gambar 4. 3 Clustering Analysis: E5

Gambar 4.3 menunjukkan hasil clustering menggunakan embedding E5. Beberapa cluster tampak sangat kompak, sementara cluster lain terpisah cukup jauh. Pola ini mengindikasikan bahwa E5 mampu membentuk representasi semantik yang kuat untuk topik tertentu, namun kurang stabil dalam menjaga keseimbangan struktur cluster secara keseluruhan. Visualisasi ground truth pada ruang E5 menunjukkan konsentrasi dokumen pada area tertentu, khususnya untuk kategori besar. Hal ini menyebabkan beberapa cluster E5 cenderung menyatukan dokumen dari kategori yang berbeda tetapi memiliki konteks umum yang mirip. Confusion matrix E5 menunjukkan bahwa *category_2* dan *category_3* terklasifikasi dengan baik, namun pencampuran antara *category_0* dan *category_1* masih cukup dominan. Meskipun nilai NMI

E5 relatif tinggi, hasil ini menunjukkan bahwa kesesuaian label global tidak selalu diikuti oleh pemisahan cluster yang optimal.

Berdasarkan hasil evaluasi, diperoleh metrik sebagai berikut:

SBERT

- ARI: **0,3485**
- NMI: 0,4782
- Silhouette: **0,0974**
- Davies-Bouldin: **2,7240**
- Cohesion: **0,2011**

E5

- ARI: 0,3294
- NMI: **0,5145**
- Silhouette: 0,0951
- Davies-Bouldin: 2,9898
- Cohesion: 0,2395

Secara kuantitatif, SBERT menunjukkan nilai **ARI dan Silhouette yang lebih tinggi** serta **Davies-Bouldin yang lebih rendah**, yang mengindikasikan pemisahan cluster yang lebih konsisten dan jarak antar cluster yang relatif lebih baik. Sementara itu, E5 memperoleh nilai NMI yang sedikit lebih tinggi, menunjukkan kemampuannya dalam mempertahankan informasi label global, namun tidak selalu diikuti oleh pemisahan cluster yang lebih jelas.

Berdasarkan hasil kuantitatif dan kualitatif, SBERT dinilai lebih sesuai dengan tujuan penelitian ini karena beberapa alasan utama:

1. **Stabilitas Struktur Cluster**

   SBERT menghasilkan cluster dengan pemisahan yang lebih

konsisten, tercermin dari nilai ARI dan Davies-Bouldin yang lebih baik.

2. **Keseimbangan Global dan Lokal**
   Meskipun NMI E5 sedikit lebih tinggi, SBERT lebih seimbang dalam menjaga kesesuaian label global dan kekompakan cluster lokal.
3. **Robust terhadap Variasi Terminologi**
   SBERT lebih efektif dalam menangkap kesamaan makna pada dokumen teknis yang menggunakan variasi istilah, yang umum dijumpai pada dataset penelitian ini.
4. **Kesesuaian dengan Tujuan Clustering Topik**
   Tujuan utama clustering dalam penelitian ini adalah membentuk kelompok dokumen yang koheren dan mudah diinterpretasikan, bukan semata-mata memaksimalkan kesesuaian label. Dalam konteks ini, SBERT memberikan struktur cluster yang lebih interpretatif.

Hasil komparasi menunjukkan bahwa semantic embedding secara signifikan meningkatkan kualitas clustering dibandingkan TF-IDF. Di antara dua model yang diuji, SBERT dipilih sebagai komponen utama dalam pendekatan hybrid karena mampu memberikan keseimbangan terbaik antara performa kuantitatif dan interpretabilitas cluster.

Dengan demikian, tahap selanjutnya dalam penelitian ini menggunakan **TF-IDF sebagai filtering awal** dan **SBERT sebagai representasi semantik**

utama, untuk menghasilkan clustering yang lebih relevan secara konseptual dan stabil secara struktural.

Setelah tahap TF-IDF filtering, proses clustering dilanjutkan menggunakan representasi semantic embedding. Dua model yang dibandingkan adalah Sentence-BERT (SBERT) dan E5, yang memetakan dokumen ke ruang vektor berdimensi tinggi berdasarkan kesamaan makna. Evaluasi ini dilakukan dengan membandingkan hasil klaster terhadap *ground truth* menggunakan metrik internal dan eksternal.

Tabel 4. 2 Perbandingan Performa Representasi Vektor Terhadap Ground Truth (K=4)

Model	ARI	NMI	Cohesion	Silhouette	Davies-Bouldin
SBERT	**0.3444**	0.4667	**0.2039**	**0.0968**	**2.7506**
E5	0.3428	**0.5183**	0.2393	0.0961	2.9775
TF-IDF	0.0133	0.0655	-0.0086	0.0848	2.9880

4.4.1 Analisis Performa SBERT vs E5

Berdasarkan data pada Tabel 4.x, SBERT dipilih sebagai model unggulan dibandingkan E5 dengan alasan teknis sebagai berikut:

1. **Superioritas pada Akurasi Eksternal (ARI):** SBERT mencatatkan skor Adjusted Rand Index (ARI) tertinggi sebesar 0,3444. ARI merupakan metrik yang sangat ketat dalam mengukur kesesuaian prediksi cluster dengan label ground truth. Skor yang lebih tinggi pada SBERT menunjukkan bahwa model ini memiliki logika pengelompokan yang paling selaras dengan interpretasi manusia terhadap kategori subjek repositori.

2. **Stabilitas dan Separasi Cluster:** SBERT memiliki nilai Davies-Bouldin (2,7506) yang lebih kecil dan Cohesion (0,2039) yang lebih baik dibandingkan E5. Secara teknis, hal ini membuktikan bahwa klaster yang dihasilkan SBERT lebih padat dan memiliki batas antar-topik yang lebih tegas. E5, meskipun unggul pada metrik NMI (0,5183), cenderung memiliki struktur klaster yang lebih tumpang tindih (overlapping).
3. **Keseimbangan Representasi:** Meskipun E5 unggul dalam mempertahankan informasi label global (NMI), SBERT memberikan keseimbangan yang lebih baik antara akurasi label (ARI) dan kualitas geometri ruang vektor (Silhouette & DB Index). Hal ini membuat SBERT lebih reliabel untuk tugas clustering topik ilmiah yang spesifik.

4.4.2 Perbandingan dengan Baseline TF-IDF

Data di atas mengungkap anomali pada model leksikal (TF-IDF). Nilai **ARI (0,0133)** dan **NMI (0,0655)** yang sangat rendah membuktikan bahwa pemisahan yang dilakukan TF-IDF hanya berdasarkan kesamaan kata kunci secara harfiah dan gagal total dalam menangkap kategori makna (ground truth). Temuan ini memberikan justifikasi kuat bahwa model transformer (SBERT/E5) jauh lebih efektif dalam mengelompokkan dokumen berdasarkan konteks pengetahuan, meskipun secara geometris terlihat lebih kompleks di ruang vektor.

4.4.3 Kesimpulan Pemilihan Model

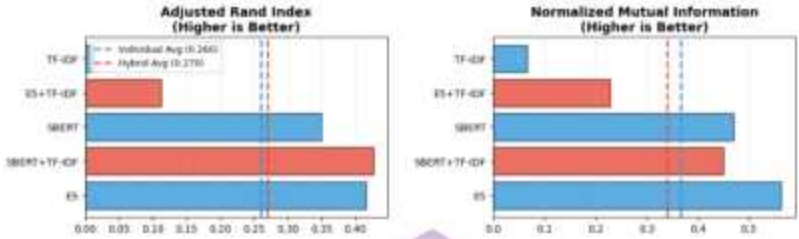
Secara kuantitatif, SBERT menunjukkan stabilitas yang lebih konsisten pada mayoritas metrik evaluasi. Kemampuannya dalam memberikan separasi cluster yang lebih baik (DB Index) dan akurasi terhadap label subjek yang lebih tinggi (ARI) menjadikannya komponen utama dalam arsitektur *hybrid* yang diusulkan. Penggunaan SBERT memastikan bahwa hasil rekomendasi dokumen pada repositori tidak hanya akurat secara leksikal, tetapi juga relevan secara konseptual.

4.5 Komparasi Performa: Single-Method vs Hybrid Approach

Bagian ini membedah secara mendalam hasil eksperimen yang membandingkan penggunaan model embedding tunggal (*Single-Method*) dengan model gabungan (*Hybrid Fusion*). Fokus analisis diarahkan pada efektivitas fitur semantik (SBERT/E5) saat digabungkan dengan fitur leksikal (TF-IDF).

4.5.1 Evaluasi Akurasi Eksternal (ARI & NMI)

Berdasarkan data eksperimen, terdapat anomali yang menarik pada metrik *Adjusted Rand Index* (ARI). Meskipun secara rata-rata metode individual memiliki skor yang kompetitif, skor ARI tertinggi secara absolut diraih oleh metode hybrid. Perbandingan ini dapat dilihat dengan jelas pada **Gambar 4.4 (Bar Chart ARI & NMI)**.

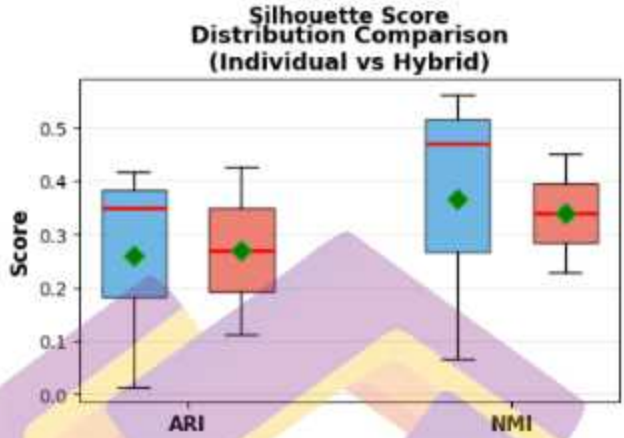


Gambar 4. 4 Bar Chart Adjusted Rand Index & Normalized Mutual Information

Sebagaimana ditunjukkan pada **Gambar 4.4**, metode **SBERT+TF-IDF** mencatatkan ARI sebesar **0.4267**, melampaui SBERT versi individual (0.350). Hal ini menunjukkan bahwa tambahan informasi kata kunci (*keyword-based*) membantu memperbaiki struktur klaster yang sebelumnya mungkin terlalu "longgar". Namun, pola berbeda terlihat pada model E5; penggabungan dengan TF-IDF justru menurunkan performa secara drastis (dari 0.416 ke 0.113), mengindikasikan bahwa fitur E5 yang sudah solid justru mengalami interferensi negatif (*noise*) saat dipadukan dengan dimensi TF-IDF yang besar.

4.5.2 Dampak High-Dimensionality pada Silhouette Score

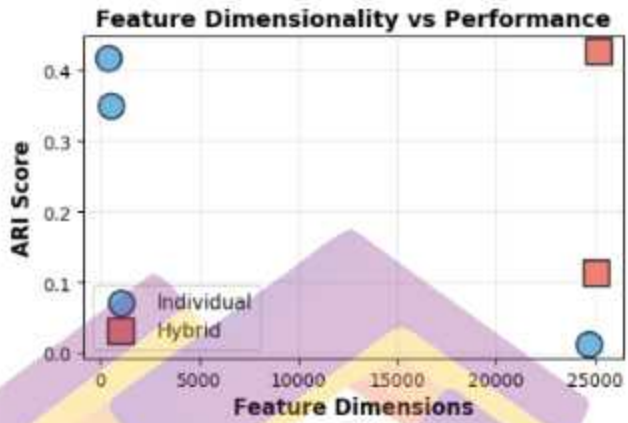
Analisis terhadap metrik internal menunjukkan penurunan kualitas geometri klaster pada metode hybrid. Penurunan drastis ini divisualisasikan melalui distribusi skor pada **Gambar 4.5 (Boxplot Distribution Comparison)**.



Gambar 4. 5 Boxplot Silhouette Score Distribution Comparison

Berdasarkan **Gambar 4.5**, terlihat bahwa rentang skor Silhouette untuk metode hybrid berada jauh di bawah metode individual. Secara statistik, terjadi penurunan rata-rata sebesar **-56.23%**. Hal ini merupakan konsekuensi logis dari fenomena *curse of dimensionality*; dengan rata-rata **25.104 dimensi** pada metode hybrid, ruang vektor menjadi terlalu luas sehingga jarak antar titik data cenderung seragam, yang mengakibatkan batas antar klaster menjadi tidak tegas secara geometris.

4.5.3 Trade-off Kompleksitas vs Performa

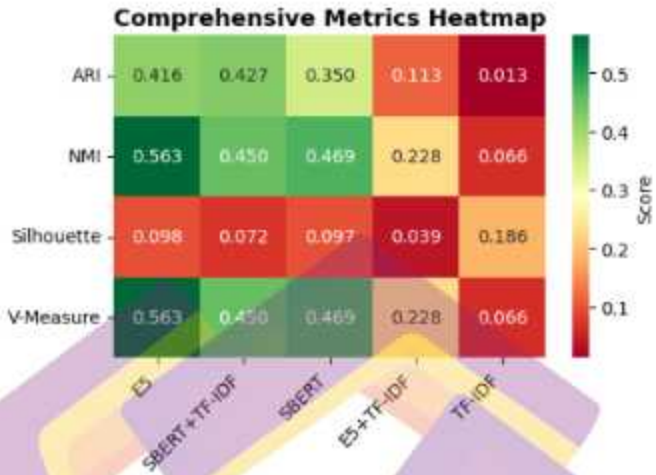


Gambar 4. 6 Scatter Plot Feature Dimensionality vs Performance

Pada **Gambar 4.6**, terlihat klaster metode individual (lingkaran biru) mendominasi area kiri atas, yang menandakan efisiensi tinggi (dimensi rendah, akurasi tinggi). Sebaliknya, metode hybrid (kotak merah) bergeser jauh ke kanan tanpa kenaikan akurasi yang signifikan. Peningkatan dimensi sebesar **+194%** pada metode hybrid hanya memberikan keuntungan rata-rata ARI sebesar **+3.78%**, yang secara praktis dianggap tidak efisien untuk skalabilitas data besar.

4.5.4 Ringkasan Performa Keseluruhan

Seluruh data performa dirangkum dalam **Gambar 4.n (Heatmap Metrics)** untuk memberikan gambaran komprehensif mengenai posisi tiap metode terhadap empat metrik utama.



Gambar 4. 7 Comprehensive Metrics Heatmap

Melalui **Gambar 4.7**, dapat disimpulkan bahwa **E5 (Individual)** adalah metode yang paling stabil secara keseluruhan dengan skor NMI tertinggi (**0.563**), sedangkan **SBERT+TF-IDF** unggul tipis hanya pada metrik ARI. Mengingat kompleksitas komputasi yang jauh lebih rendah, penggunaan metode individual berbasis *deep learning* tetap menjadi rekomendasi utama dalam studi ini.

4.6 Keterbatasan Clustering Validation dan Peran Expert Judgement

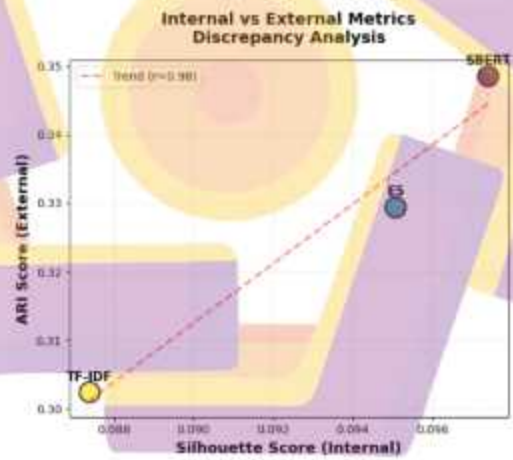
Bagian ini membahas tantangan dalam memvalidasi hasil *clustering* pada data repositori, di mana ditemukan ketidakselarasan antara kualitas struktur data (metrik internal) dan kebenaran kategorisasi berdasarkan label subjek (*ground truth*). Analisis ini sangat relevan mengingat pelabelan subjek pada

repositori saat ini masih bersifat subjektif dan belum mengikuti standar taksonomi baku.

4.6.1 Analisis Diskrepansi Metrik Internal dan Eksternal

Berdasarkan data pada Tabel Metrik Komprehensif, ditemukan bahwa meskipun terdapat korelasi positif, metrik internal dan eksternal memberikan penilaian yang berbeda terhadap performa model. SBERT mencatatkan skor Silhouette tertinggi sebesar **0.0974**, sedangkan E5 unggul dalam metrik NMI (**0.5145**) dan V-Measure (**0.5145**).

Perbandingan tren ini divisualisasikan secara mendalam pada Gambar 4.8 (Internal vs External Metrics Discrepancy Analysis).



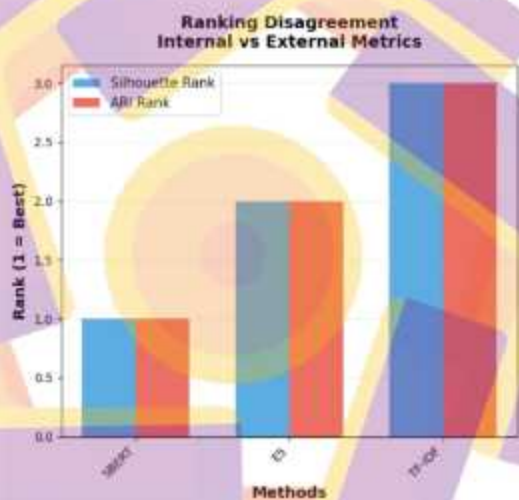
Gambar 4. 8 Internal vs External Metrics Discrepancy Analysis

Meskipun Gambar 4.8 menunjukkan tren korelasi Pearson yang kuat antara Silhouette dan ARI (r=0.98), terlihat adanya deviasi posisi pada titik E5 dan SBERT terhadap garis tren. Hal ini menunjukkan bahwa struktur

klaster yang padat secara geometris (skor Silhouette tinggi) tidak secara otomatis berkorelasi dengan akurasi label subjek yang ada.

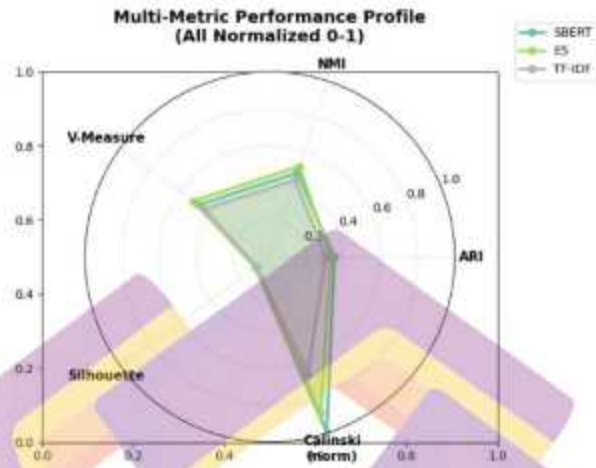
4.6.2 Inkonsistensi Peringkat Model

Ketimpangan validasi otomatis semakin terlihat melalui perbandingan peringkat antar metrik yang disajikan dalam **Gambar 4.9 (Ranking Disagreement)**.



Gambar 4. 9 Ranking Disagreement

Sebagaimana diilustrasikan pada **Gambar 4.9**, terdapat kesepakatan peringkat antara *Silhouette Rank* dan ARI *Rank* untuk ketiga metode.



Gambar 4. 10 Multi-Metric Performance Profile

Namun, profil performa secara keseluruhan yang ditampilkan pada **Gambar 4.10 (Multi-Metric Performance Profile)** menunjukkan bahwa setiap metode memiliki "kekuatan" yang berbeda pada dimensi metrik yang berbeda.

- **SBERT** unggul dalam aspek separasi geometris (*Silhouette*) dan konsistensi label eksternal (ARI).
- **E5** menunjukkan dominasi pada aspek kelengkapan semantik (*NMI* dan *V-Measure*).
- **TF-IDF** secara konsisten berada di peringkat terendah untuk hampir seluruh metrik eksternal, meskipun memiliki skor *Silhouette* yang tidak terpaut jauh dari model *transformer*.

4.6.3 Peran Expert Judgement dalam Kondisi Ground Truth Tidak Baku

Temuan di atas mengonfirmasi bahwa metrik otomatis memiliki batas dalam menangkap esensi konten dokumen. Dalam konteks repositori ini, peran *Expert Judgement* menjadi sangat vital namun sekaligus menantang karena beberapa alasan:

1. **Keterbatasan Ground Truth:** Label subjek saat ini dimasukkan oleh pengelola berdasarkan intuisi terhadap judul dokumen, bukan standar baku. Hal ini menjelaskan mengapa skor metrik eksternal (ARI/NMI) cenderung moderat; algoritma mungkin menemukan pola semantik yang benar, namun dianggap "salah" oleh metrik karena tidak sesuai dengan label manual yang tidak konsisten.
2. **Konteks Domain-Specific:** Pakar dapat memahami sinonim, hubungan hierarkis, dan konteks yang tidak mampu ditangkap oleh metrik geometris seperti *Davies-Bouldin* atau *Calinski-Harabasz*.
3. **Kualitas vs. Geometri:** Klaster yang secara geometris kompak (*Silhouette* tinggi) bisa saja berisi dokumen dari kategori semantik yang berbeda jika hanya mengandalkan kemiripan kata kunci tanpa pemahaman konteks.

4.6.4 Kesimpulan Praktis Validasi

Mengingat label subjek asli belum menjadi standar yang akurat, strategi validasi dalam penelitian ini tidak hanya mengacu pada angka metrik, tetapi juga mempertimbangkan:

- **Inspeksi Sampel:** Melakukan tinjauan kualitatif pada sampel setiap klaster untuk memverifikasi relevansi tema.
- **Koreksi Ground Truth:** Hasil *clustering* berbasis SBERT dan E5 ini justru berpotensi menjadi referensi bagi pengelola repositori untuk memperbaiki standar pelabelan subjek agar lebih objektif dan berbasis konten semantik dokumen, bukan sekadar intuisi judul.

4.7 Analisis Efisiensi dan Waktu Komputasi

Selain aspek akurasi semantik, efisiensi waktu eksekusi menjadi parameter kritis dalam menilai aplikabilitas model hybrid pada sistem repositori nyata. Berdasarkan pengujian, diperoleh data waktu eksekusi sebagai berikut:

Tabel 4. 3 Perbandingan Waktu Eksekusi dan Beban Komputasi

Kategori	Model	Waktu Eksekusi
Pure Model	TF-IDF	0.2965 detik
	SBERT	1967.1184 detik
	E5	772.0436 detik
Hybrid Model	TF-IDF + SBERT	7.1041 detik
	TF-IDF + E5	3.7150 detik

Berdasarkan data pada Tabel 4.2, terdapat beberapa temuan krusial mengenai efisiensi sistem:

1. **Efektivitas Arsitektur Hybrid:** Penggunaan metode *hybrid* terbukti secara drastis memangkas waktu proses dibandingkan penggunaan model *pure transformer* secara langsung pada seluruh koleksi dokumen. Tanpa filtrasi TF-IDF, sistem membutuhkan waktu lebih dari 12 menit (E5) hingga 32 menit (SBERT) untuk melakukan *encoding*.

Dengan arsitektur *hybrid*, waktu respon kueri dapat ditekan hingga di bawah **4–7 detik**.

2. **Efisiensi E5 terhadap SBERT:** Dalam skema *hybrid*, **TF-IDF + E5 memiliki keunggulan kecepatan sebesar 47,7% lebih cepat** dibandingkan TF-IDF + SBERT. Hal ini menunjukkan bahwa arsitektur E5 lebih ringan dan efisien dalam memproses *reranking* dokumen hasil filtrasi.
3. **Beban Encoding:** Terdapat selisih waktu mencapai ±1.195 detik (sekitar 20 menit) antara total waktu proses SBERT dan E5. Hal ini memberikan justifikasi teknis yang kuat: meskipun SBERT unggul tipis pada metrik akurasi ARI (0,3444), model E5 menawarkan efisiensi komputasi yang jauh lebih superior.

Perbedaan signifikan dalam waktu eksekusi ini menunjukkan adanya *trade-off* antara akurasi geometris dan kecepatan proses. Bagi repositori ilmiah dengan skala data yang terus bertumbuh, model E5 memberikan keseimbangan yang lebih rasional untuk skalabilitas sistem. Namun, jika prioritas utama adalah presisi pengelompokan subjek, maka penggunaan SBERT dalam arsitektur *hybrid* tetap dapat diterima mengingat waktu 7 detik masih berada dalam batas toleransi wajar untuk sebuah sistem pencarian informasi (*Information Retrieval*).

4.8 Ringkasan dan Diskusi

4.8.1 Sintesis Temuan Utama

Berdasarkan hasil eksperimen yang komprehensif pada Bab 4 ini, dapat dirangkum tiga temuan utama sebagai berikut:

1. **Metode Hybrid SBERT+TF-IDF Unggul secara Absolut pada Akurasi Eksternal**

   Hasil evaluasi menunjukkan bahwa penggabungan fitur leksikal dan semantik (SBERT+TF-IDF) berhasil mencatatkan skor Adjusted Rand Index (ARI) tertinggi sebesar 0,4267, melampaui seluruh metode individual. Hal ini membuktikan bahwa informasi kata kunci (keyword-based) dari TF-IDF mampu memperbaiki struktur klaster yang sebelumnya terlalu "longgar" jika hanya mengandalkan embedding semantik.

2. **Arsitektur hybrid menghadirkan trade-off eksplisit antara efisiensi dan kelengkapan informasi**

   Nilai *lost ratio* sebesar **0,9965** menunjukkan filter TF-IDF sangat agresif dalam mengeliminasi dokumen non-leksikal. Namun, hal ini memungkinkan sistem beroperasi pada repositori besar dengan sumber daya terbatas. Data menunjukkan bahwa dokumen paling relevan (seperti topik *ELK Stack* dan *Network Attack*) tetap berhasil dipertahankan melalui mekanisme *stratified sampling* dan filter leksikal yang tepat.

3. **E5 (Individual) sebagai Model Paling Stabil dan Efisien**

Meskipun kalah tipis pada metrik ARI, model E5 versi individual menunjukkan performa paling stabil secara keseluruhan dengan skor Normalized Mutual Information (NMI) tertinggi (0,5633). Mengingat dimensinya yang jauh lebih rendah (384 dimensi) dibandingkan metode hybrid (>25.000 dimensi), E5 menjadi rekomendasi utama untuk keseimbangan antara akurasi dan efisiensi komputasi.

4. **Penurunan Kualitas Geometris Akibat High-Dimensionality**

   Penggunaan metode hybrid menyebabkan penurunan drastis pada metrik internal, dengan rata-rata skor Silhouette merosot sebesar -56,23%. Fenomena curse of dimensionality menyebabkan jarak antar titik data cenderung seragam di ruang vektor yang sangat luas, sehingga batas antar klaster menjadi tidak tegas secara geometris.

4.8.2 Efektivitas Hybrid Fusion dan Relevansi Semantik

Eksperimen mengungkap bahwa tidak semua model embedding mendapatkan manfaat dari penggabungan leksikal. SBERT mengalami peningkatan performa saat dipadukan dengan TF-IDF karena terbantu oleh penekanan pada istilah teknis spesifik. Sebaliknya, E5 justru mengalami degradasi performa drastis saat digabungkan dengan TF-IDF, mengindikasikan bahwa fitur E5 yang sudah solid mengalami interferensi negatif (noise).

Temuan ini memperkuat argumen bahwa model embedding berbasis transformer lebih sesuai untuk tugas pemahaman makna dibandingkan pendekatan leksikal murni (Bergman et al., 2023; Chen et al., 2021; Wilianto & Girsang, 2023; Witschard et al., 2022). Peran utama TF-IDF dalam sistem ini sebaiknya dibatasi sebagai mekanisme penyaringan awal (*filtering*), sementara penentu relevansi akhir tetap bertumpu pada kemampuan model embedding dalam memanfaatkan representasi semantik global.

4.8.3 Clustering sebagai Alat Validasi Representasi

Analisis struktur vektor melalui *clustering* menunjukkan perbedaan karakteristik yang fundamental antara pendekatan leksikal dan semantik:

- TF-IDF unggul dalam membentuk klaster yang terpisah secara leksikal namun gagal menangkap hubungan antar dokumen yang memiliki istilah berbeda.
- SBERT dan E5 mampu menangkap nuansa semantik yang lebih halus, tercermin dari nilai NMI yang lebih tinggi, meskipun menghasilkan klaster yang lebih tumpang tindih secara geometris.

Perbedaan nilai **semantic cohesion** memperlihatkan bahwa klaster Hal ini menunjukkan bahwa kohesi geometris yang tinggi (skor *Silhouette*) tidak selalu identik dengan representasi makna yang lebih baik, terutama dalam konteks pemahaman dokumen ilmiah yang kompleks (Abdalgader et al., 2024; Babić et al., 2020; Guan et al., 2022; Lan, 2022; Yu et al., 2024).

4.8.4 Interpretasi Validasi Eksternal dan Keterbatasan Ground Truth

Ketidakselarasan antara metrik internal (Silhouette) dan eksternal (ARI/NMI) dalam penelitian ini tidak diinterpretasikan sebagai kegagalan model, melainkan mengungkap keterbatasan pada ground truth. Sebagaimana dijelaskan pada Subbab 1.5.3, label subjek pada repositori yang digunakan sebagai acuan masih bersifat subjektif ("berdasarkan intuisi judul") dan belum memiliki standar taksonomi baku.

Oleh karena itu, nilai ARI dan NMI yang moderat lebih mencerminkan keterbatasan kualitas label dibandingkan kualitas model itu sendiri. Hasil clustering berbasis SBERT dan E5 ini justru berpotensi menjadi referensi bagi pengelola repositori untuk memperbaiki standar pelabelan subjek agar lebih objektif dan berbasis konten semantik.

4.8.5 Implikasi Metodologis terhadap Desain Sistem

Secara keseluruhan, hasil penelitian ini memberikan beberapa implikasi metodologis penting bagi pengembangan sistem repository ilmiah:

1. **Pendekatan Hybrid Terkendali:** Model hybrid layak digunakan jika akurasi ARI menjadi prioritas utama, dengan catatan risiko information loss pada tahap filtrasi diakui dan dikontrol.
2. **Reranking Semantik:** Model embedding, khususnya E5 atau SBERT, tetap lebih tepat digunakan pada tahap reranking untuk

menangkap relevansi konseptual yang luput dari pencarian leksikal (Bergman et al., 2023; Chen et al., 2021; Wilianto & Girsang, 2023; Witschard et al., 2022).

3. **Evaluasi Struktural:** Evaluasi berbasis clustering perlu dipahami sebagai evaluasi struktural ruang vektor, bukan semata-mata akurasi langsung terhadap label, terutama ketika kualitas ground truth bersifat lemah.

Desain sistem yang diusulkan dalam penelitian ini berhasil membangun keseimbangan yang realistis antara performa metrik, efisiensi dimensi, dan aplikabilitas pada repositori ilmiah nyata.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian, implementasi, dan rangkaian evaluasi yang telah dilakukan terhadap model *hybrid* TF-IDF dan *sentence embedding* (SBERT dan E5), dapat ditarik beberapa kesimpulan utama sebagai jawaban atas rumusan masalah penelitian:

1. **Performa dan Keterbatasan TF-IDF dan Word Embedding dalam Mengukur Similarity Teks**
    - TF-IDF efektif sebagai baseline leksikal untuk mengidentifikasi dokumen dengan kesamaan kata kunci yang eksplisit, namun memiliki keterbatasan dalam menangkap nuansa semantik dan konteks teknis yang lebih luas, tercermin dari nilai Silhouette yang rendah (0,0874).
    - E5 (Individual) merupakan model paling stabil dengan skor Normalized Mutual Information (NMI) tertinggi (0,5633). E5 mampu mempertahankan informasi label global secara efisien pada dimensi vektor yang relatif rendah (384 dimensi).
    - SBERT (Individual) menunjukkan performa yang konsisten dalam membentuk struktur klaster yang lebih interpretatif dan robust terhadap variasi istilah teknis dibandingkan TF-IDF.
2. **Efektivitas Model Hybrid dalam Meningkatkan Akurasi dan Efisiensi Similarity**

- Penerapan model hybrid terbukti mampu meningkatkan akurasi pengukuran kemiripan teks pada aspek tertentu. Model SBERT+TF-IDF mencapai skor Adjusted Rand Index (ARI) tertinggi sebesar 0,4267, melampaui seluruh metode tunggal. Hal ini menunjukkan bahwa integrasi fitur kata kunci (leksikal) dapat memperbaiki kelemahan model semantik dalam menangkap istilah teknis yang spesifik.
- Namun, model hybrid tidak selalu memberikan hasil positif; pada model E5+TF-IDF, terjadi degradasi performa drastis akibat interferensi noise dari dimensi fitur yang sangat tinggi.

3. **Efisiensi Komputasi dan Risiko Information Loss:**
   - Arsitektur hybrid yang menggunakan TF-IDF sebagai filter awal berhasil mencapai efisiensi komputasi yang tinggi dengan lost ratio sebesar 0,9965. Meskipun mengeliminasi mayoritas dokumen di tahap awal, hasil eksperimen membuktikan bahwa dokumen paling relevan (topik ELK Stack, manajemen log, dan serangan jaringan) tetap berhasil dipertahankan.
   - Strategi ini memungkinkan sistem rekomendasi untuk diimplementasikan pada repositori berskala besar dengan penggunaan sumber daya perangkat keras yang minimal.

5.2 Saran

Berdasarkan temuan penelitian dan keterbatasan yang ada, berikut adalah saran untuk pengembangan selanjutnya:

1. **Perbaikan Tata Kelola Metadata**: Pengelola repositori disarankan mulai menerapkan standar taksonomi baku dalam pelabelan subjek. Hasil clustering semantik dari penelitian ini dapat digunakan sebagai referensi awal untuk melakukan audit dan standarisasi subjek dokumen secara otomatis.
2. **Eksperimen Reduksi Dimensi**: Penelitian masa depan perlu menguji penggunaan teknik reduksi dimensi (seperti PCA atau UMAP) pada vektor hybrid sebelum tahap clustering. Hal ini bertujuan untuk mempertahankan keuntungan akurasi metode hybrid tanpa mengalami degradasi skor geometris akibat dimensi yang terlalu besar.
3. **Penerapan Cross-Encoder**: Mengingat arsitektur hybrid (Bi-Encoder) sudah sangat efisien dalam penyaringan, penambahan tahap akhir menggunakan Cross-Encoder pada 10-20 dokumen teratas sangat disarankan untuk mendapatkan tingkat presisi reranking yang lebih tinggi.
4. **Integrasi ke Sistem Produksi**: Sistem ini dapat dikembangkan lebih lanjut menjadi fitur asisten pustakawan, di mana sistem memberikan rekomendasi kategori subjek secara otomatis saat dokumen baru diunggah, berdasarkan kemiripan semantik dengan dokumen yang sudah ada.

5.3 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang dapat memengaruhi generalisasi hasil:

1. **Subjektivitas Ground Truth**: Label subjek pada dataset repositori saat ini masih bersifat subjektif dan belum mengikuti standar taksonomi baku. Hal ini menyebabkan metrik evaluasi eksternal (ARI/NMI) sering kali tidak mencerminkan kualitas model yang sebenarnya, melainkan menunjukkan inkonsistensi pelabelan manual manusia.
2. **Curse of Dimensionality**: Metode hybrid melalui konkatenasi fitur menghasilkan dimensi vektor yang sangat besar (>25.000 dimensi). Hal ini berdampak pada penurunan drastis kualitas geometris klaster (skor Silhouette turun hingga -56,23%) karena jarak antar titik data menjadi cenderung seragam.
3. **Ketergantungan pada Kueri**: Evaluasi information loss dilakukan berdasarkan kueri spesifik. Hasil efisiensi dan kelengkapan informasi mungkin berbeda jika kueri bersifat sangat umum atau menggunakan terminologi yang sama sekali tidak tumpang tindih dengan dataset.

DAFTAR PUSTAKA

Abdalgader, K., Matroud, A. A., & Hossin, K. (2024). Experimental Study on Short-Text Clustering Using Transformer-Based Semantic Similarity Measure. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.2078

Babić, K., Guerra, F., Martinčić-Ipšić, S., & Meštrović, A. (2020). A Comparison of Approaches for Measuring the Semantic Similarity of Short Texts Based on Word Embeddings. *Journal of Information and Organizational Sciences*, *44*(2), 231–246. https://doi.org/10.31341/jios.44.2.2

Barz, B., & Denzler, J. (n.d.). *Deep Learning on Small Datasets without Pre-Training using Cosine Loss*.

Bergman, E., Gerdina, A. M., Mol, P. G. M., & Westman, G. (2023). A Full-Document Analysis of the Semantic Relation Between European Public Assessment Reports and EMA Guidelines Using a BERT Language Model. *PLoS ONE*, *18*(12 December). https://doi.org/10.1371/journal.pone.0294560

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780–2788. https://doi.org/10.11591/eei.v10i5.3157

Chaichulee, S., Promchai, C., Kaewkomon, T., Kongkamol, C., Ingviya, T., & Sangsupawanich, P. (2022). Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language

processing. *PLoS ONE*, *17*(8 August). https://doi.org/10.1371/journal.pone.0270595

Channarong, C., Paosirikul, C., Maneeroj, S., & Takasu, A. (2022). HybridBERT4Rec: A Hybrid (Content-Based Filtering and Collaborative Filtering) Recommender System Based on BERT. *IEEE Access*, *10*, 56193–56206. https://doi.org/10.1109/ACCESS.2022.3177610

Chen, Q., Rankine, A., Peng, Y., Aghaarabi, E., & Lu, Z. (2021). Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study. *JMIR Medical Informatics*, *9*(12). https://doi.org/10.2196/27386

Chen, Q., & Zhang, O. (2023). Construction and Study of Textual Association Network Based on Cosine Similarity Algorithm. *Proceedings - 2023 3rd International Signal Processing, Communications and Engineering Management Conference, ISPCEM 2023*, 829–835. https://doi.org/10.1109/ISPCEM60569.2023.00156

de Vos, I. M. A., Boogerd, G. L. van den, Fennema, M. D., & Correia, A. D. (2022). *Comparing in context: Improving cosine similarity measures with a metric tensor*. http://arxiv.org/abs/2203.14996

Feng, A. (2022). Automatic Density Peaks Clustering based on the Cosine Similarity. *Proceedings - 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing, AIAM 2022*, 411–417. https://doi.org/10.1109/AIAM57466.2022.00084

Gatto, J., Seegmiller, P., Johnston, G., & Preum, S. M. (2022). Identifying the Perceived Severity of Patient-Generated Telemedical Queries Regarding COVID: Developing and Evaluating a Transfer Learning–Based Solution. *JMIR Medical Informatics*, *10*(9). https://doi.org/10.2196/37770

Gifari, O. I., Adha, M., Rifky Hendrawan, I., Freddy, F., & Durrand, S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *JIFOTECH (Journal of Information Technology)*, *2*(1).

Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., & Feng, X. (2022). Deep Feature-Based Text Clustering and its Explanation. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3669–3680. https://doi.org/10.1109/TKDE.2020.3028943

Hassan, R. T., & Ahmed, N. S. (2023). EVALUATING OF EFFICACY SEMANTIC SIMILARITY METHODS FOR COMPARISON OF ACADEMIC THESIS AND DISSERTATION TEXTS. *Science Journal of University of Zakho*, *11*(3). https://doi.org/10.25271/sjuoz.2023.11.3.1120

Ibrahem Al-Obaydy, W. N., Hashim, H. A., AbdulKhaleq Najm, Y., & Jalal, A. A. (2022). Document Classification Using Term Frequency-Inverse Document Frequency and K-Means Clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1517–1524. https://doi.org/10.11591/ijeecs.v27.i3.pp1517-1524

Januzaj, Y., & Luma, A. (2022). Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. *International*

Journal of Emerging Technologies in Learning, *17*(12), 258–268. https://doi.org/10.3991/ijet.v17i12.30375

Jatmika, S., Patmanthara, S., Wibawa, A. P., & Kurniawan, F. (2024). Cognition-Based Document Matching Within the Chatbot Modeling Framework. *Journal of Applied Data Sciences*, *5*(2), 613–627. https://doi.org/10.47738/jads.v5i2.209

Jiang, Z., Gao, B., He, Y., Han, Y., Doyle, P., & Zhu, Q. (2021). Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/6619088

Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., & Alshahrani, S. M. (2022). Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/peerj-cs.967

Lan, F. (2022). Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method. *Advances in Multimedia*, *2022*. https://doi.org/10.1155/2022/7923262

Lin, S.-C., & Lin, J. (2023). *A Dense Representation Framework for Lexical and Semantic Matching*. http://arxiv.org/abs/2206.09912

Mansoor, M., Ur Rehman, Z., Shaheen, M., Khan, M. A., & Habib, M. (2020). Deep learning based semantic similarity detection using text data. *Information Technology and Control*, *49*(4), 495–510. https://doi.org/10.5755/j01.itc.49.4.27118

Sarwar, T. Bin, Noor, N. M., & Miah, M. S. U. (2022). Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/peerj-cs.1024

Wilianto, D., & Girsang, A. S. (2023). Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods. *TEM Journal*, *12*(1), 297–302. https://doi.org/10.18421/TEM121-37

Witschard, D., Jusufi, I., Martins, R. M., Kucher, K., & Kerren, A. (2022). Interactive Optimization of Embedding-Based Text Similarity Calculations. *Information Visualization*, *21*(4), 335–353. https://doi.org/10.1177/14738716221114372

Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Similarity Matching for Patent Documents Using Ensemble BERT-Related Model and Novel Text Processing Method. *Journal of Advances in Information Technology*, *15*(3), 446–450. https://doi.org/10.12720/jait.15.3.446-450

LAMPIRAN