P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v11i2

Aplikasi Semantic Search untuk Dokumen Arsip Berbasis DistilBERT dengan Antarmuka Web Interaktif

[1]Muhammad Rizqi Anugrah, [1]Oddy Virgantara Putra

[1]Program Studi Teknik Informatika, Universitas Darussalam Gontor, Ponorogo

E-mail: [1]muhammadrizqianugrah42@student.cs.unida.gontor.ac.id, [1]oddy@unida.gontor.ac.id

ABSTRAK

Pencarian dokumen arsip universitas membutuhkan mekanisme temu kembali informasi yang mampu menangani variasi istilah pada kueri dan metadata. Pencarian berbasis kata kunci memiliki keterbatasan ketika istilah pengguna tidak sama dengan istilah yang terdapat pada dokumen. Penelitian ini mengimplementasikan aplikasi pencarian semantik berbasis web menggunakan model *pre-trained* DistilBERT tanpa pelatihan ulang dan membandingkan kinerjanya dengan TF-IDF. Pemeringkatan DistilBERT menggunakan *cosine similarity* yang dilanjutkan dengan *lexical reranking* sehingga disebut DistilBERT Hybrid. Evaluasi dilakukan pada 703 dokumen dengan representasi *metadata-only* menggunakan benchmark final 50 kueri terstratifikasi, Top-5, dan *pooled relevance judgment* Top-20 dari kedua metode yang menghasilkan 1.459 pasangan penilaian. TF-IDF memperoleh *macro* Precision@5 sebesar 0,2520, Pooled Recall@5 sebesar 0,7453, dan F1@5 sebesar 0,3372, sedangkan DistilBERT Hybrid memperoleh masing-masing 0,2240, 0,7050, dan 0,3037; Pooled Recall@5 dan F1@5 dihitung pada 49 kueri karena satu kueri tidak memiliki dokumen relevan dalam *judged pool*. Hasil menunjukkan TF-IDF lebih tinggi pada benchmark ini, sehingga DistilBERT Hybrid belum memberikan peningkatan kinerja keseluruhan pada kondisi evaluasi *metadata-only*. Penelitian selanjutnya dapat mengevaluasi representasi isi dokumen yang lebih lengkap dan kueri dari penggunaan nyata.

Kata Kunci: *Pencarian semantik; DistilBERT; TF-IDF; arsip universitas; temu kembali informasi.*

ABSTRACT

Searching university archival documents requires an information retrieval mechanism that can handle variations between query terms and document metadata. Keyword-based search is limited when users employ terms that differ from those contained in the documents. This study implements a web-based semantic search application using a pre-trained DistilBERT model without retraining and compares its retrieval performance with TF-IDF. DistilBERT ranking uses cosine similarity followed by lexical reranking and is therefore referred to as DistilBERT Hybrid. The evaluation was conducted on 703 documents with metadata-only representations using a final stratified benchmark of 50 queries, Top-5 retrieval, and Top-20 pooled relevance judgment from both methods, resulting in 1,459 judged query-document pairs. TF-IDF achieved macro Precision@5 of 0.2520, Pooled Recall@5 of 0.7453, and F1@5 of 0.3372, while DistilBERT Hybrid achieved 0.2240, 0.7050, and 0.3037, respectively; Pooled Recall@5 and F1@5 were calculated over 49 queries because one query had no relevant document in the judged pool. The results show that TF-IDF performed better on this benchmark, while DistilBERT Hybrid did not improve overall retrieval performance under the metadata-only evaluation setting. Future work may evaluate richer document-content representations and queries derived from real usage.

P-ISSN : 2580-4316
E-ISSN : 2654-8054
https://doi.org/10.37817/ikraith-informatika.v11i2

Keywords: *Semantic search; DistilBERT; TF-IDF; university archives; information retrieval.*

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong perguruan tinggi untuk meningkatkan pengelolaan dokumen arsip dalam bentuk digital. Dokumen arsip merupakan sumber informasi penting yang mendukung kegiatan akademik dan administrasi universitas, seperti surat keputusan, pedoman akademik, laporan kegiatan, serta dokumen kelembagaan lainnya. Seiring bertambahnya jumlah dan jenis dokumen, proses pencarian informasi menjadi salah satu aspek penting dalam pengelolaan arsip digital. Sistem pengelolaan arsip berbasis web dapat membantu proses penyimpanan dan akses dokumen secara lebih terstruktur (Jian, 2022; Mewengkang et al., 2022). Namun, pengelolaan dokumen digital juga memerlukan mekanisme pencarian yang mampu menemukan informasi sesuai dengan kebutuhan pengguna (Jannah et al., 2022).

Salah satu keterbatasan dalam pencarian dokumen arsip adalah penggunaan metode pencocokan kata kunci (*keyword matching*) yang bergantung pada kesamaan istilah antara kueri pengguna dan teks dokumen. Pendekatan tersebut dapat mengalami kesulitan ketika pengguna menggunakan sinonim, variasi istilah, atau susunan kata yang berbeda dari dokumen yang dicari. Akibatnya, dokumen yang memiliki informasi relevan berpotensi tidak ditemukan apabila kata kunci yang digunakan tidak sesuai dengan istilah dalam dokumen. (Ladanavar et al., 2024) membahas pemanfaatan pencarian semantik berbasis BERT untuk meningkatkan pemahaman konteks kueri pengguna dan relevansi hasil pencarian.

Permasalahan tersebut juga menjadi perhatian dalam lingkungan Universitas Darussalam Gontor. Berdasarkan survei awal yang dilakukan terhadap 29 responden mahasiswa menggunakan kuesioner berskala Likert, sebanyak 72,4% responden menyatakan bahwa mereka perlu mencoba beberapa kata kunci berbeda untuk menemukan dokumen yang dibutuhkan. Selain itu, 62,1% responden menyatakan bahwa hasil pencarian sering kali tidak sesuai dengan kebutuhan mereka. Sementara itu, 89,7% responden menginginkan sistem pencarian yang mampu memahami makna kueri, bukan hanya mencocokkan kata kunci secara literal. Temuan survei awal ini menunjukkan adanya kebutuhan pengguna terhadap pendekatan pencarian dokumen yang mempertimbangkan konteks dan makna informasi.

Salah satu pendekatan yang dapat digunakan untuk menjawab kebutuhan tersebut adalah *semantic search*. Berbeda dengan pencarian berbasis kecocokan kata secara langsung, pencarian semantik memanfaatkan representasi numerik teks untuk mengukur kedekatan makna antara kueri dan dokumen. Pendekatan ini memungkinkan sistem mempertimbangkan hubungan kontekstual antarteks meskipun menggunakan istilah yang berbeda. (Wang et al., 2024) mengkaji berbagai penerapan model berbasis BERT dalam bidang *information retrieval* yang memanfaatkan representasi bahasa kontekstual untuk mendukung proses temu kembali informasi.

Dalam penelitian ini, model DistilBERT digunakan sebagai dasar pembentukan representasi semantik dokumen dan kueri. DistilBERT merupakan model berbasis arsitektur *transformer* yang dikembangkan melalui pendekatan *knowledge distillation* untuk menghasilkan model yang lebih ringkas dibandingkan BERT (Sanh et al., 2019). Karakteristik tersebut menjadikan DistilBERT relevan untuk dipertimbangkan dalam pengembangan aplikasi pemrosesan bahasa alami yang memperhatikan kebutuhan sumber daya komputasi (Silva Barbon & Akabane,

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v11i2

2022). Penelitian ini menggunakan model *pre-trained* DistilBERT tanpa melakukan proses pelatihan ulang (*training*).

Penelitian ini juga menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagai metode pembanding. TF-IDF merepresentasikan dokumen berdasarkan bobot istilah yang dihitung dari frekuensi kemunculan kata dalam dokumen dan koleksi dokumen. Berbeda dengan representasi kontekstual DistilBERT, TF-IDF mengandalkan informasi leksikal dalam proses pencarian. Penggunaan kedua pendekatan ini memungkinkan dilakukannya evaluasi relevansi hasil pencarian untuk mengetahui karakteristik dan perbedaan kinerja masing-masing metode pada dokumen arsip universitas. Evaluasi dirancang menggunakan metrik *Precision*, *Recall*, dan *F1-Score* berdasarkan acuan relevansi dokumen yang ditetapkan pada tahap evaluasi.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi aplikasi pencarian semantik dokumen arsip Universitas Darussalam Gontor berbasis web interaktif dengan memanfaatkan model *pre-trained* DistilBERT. Sistem dirancang untuk mendukung pencarian dokumen berdasarkan kemiripan semantik serta menyediakan antarmuka yang memudahkan pengguna dalam mengakses informasi arsip. Selain itu, penelitian ini bertujuan untuk mengevaluasi relevansi hasil pencarian DistilBERT Hybrid dengan TF-IDF sebagai metode pembanding. Kontribusinya adalah evaluasi DistilBERT Hybrid dan TF-IDF pada arsip universitas dengan representasi metadata dan ruang kandidat yang setara.

2. LANDASAN TEORI

2.1 Information Retrieval dan Semantic Search

Information Retrieval (IR) merupakan proses pencarian dan pengambilan informasi yang relevan dari sekumpulan dokumen berdasarkan kebutuhan pengguna. Sistem temu kembali informasi bertujuan untuk menemukan dokumen yang sesuai dengan kueri yang diberikan. Pendekatan pencarian konvensional umumnya mengandalkan kecocokan istilah antara kueri dan dokumen, sehingga memiliki keterbatasan ketika pengguna menggunakan sinonim atau variasi kata yang berbeda. (Wang et al., 2024) membahas perkembangan pendekatan *information retrieval* berbasis BERT yang memanfaatkan representasi bahasa kontekstual dalam proses pencarian informasi.

Semantic search merupakan pendekatan pencarian yang mempertimbangkan kesamaan makna antara kueri dan dokumen, bukan hanya kecocokan kata secara literal. Pendekatan ini memanfaatkan representasi teks dalam bentuk vektor untuk mengidentifikasi hubungan semantik antarteks. (Ladanavar et al., 2024) mengembangkan sistem pencarian semantik berbasis BERT dan Elasticsearch untuk meningkatkan pemahaman konteks kueri serta relevansi hasil pencarian. Dengan pendekatan tersebut, dokumen yang menggunakan istilah berbeda tetapi memiliki konteks serupa berpotensi ditemukan oleh sistem (Melo et al., 2023).

2.2 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan kata yang digunakan untuk merepresentasikan tingkat kepentingan suatu istilah dalam dokumen terhadap keseluruhan koleksi dokumen. *Term Frequency* (TF) mengukur frekuensi kemunculan suatu istilah dalam dokumen, sedangkan *Inverse Document Frequency* (IDF) memberikan bobot lebih tinggi pada istilah yang relatif jarang muncul dalam koleksi dokumen.

Dalam sistem temu kembali informasi, TF-IDF digunakan untuk membentuk representasi vektor dokumen dan kueri berdasarkan karakteristik leksikal. Tingkat kemiripan antara keduanya

P-ISSN : 2580-4316
E-ISSN : 2654-8054
https://doi.org/10.37817/ikraith-informatika.v11i2

kemudian dapat dihitung untuk menentukan urutan hasil pencarian. Meskipun efisien dalam merepresentasikan kecocokan istilah, TF-IDF memiliki keterbatasan dalam menangkap hubungan makna ketika kueri dan dokumen menggunakan kosakata berbeda (Wang et al., 2024). Dalam penelitian ini, TF-IDF digunakan sebagai metode pembanding untuk mengevaluasi pendekatan pencarian berbasis DistilBERT.

2.3 DistilBERT

DistilBERT merupakan versi terdistilasi dari BERT (*Bidirectional Encoder Representations from Transformers*) yang dikembangkan melalui teknik *knowledge distillation*. Teknik tersebut bertujuan menghasilkan model yang lebih ringkas dengan mentransfer kemampuan dari model BERT yang lebih besar (Karim et al., 2025). (Sanh et al., 2019) menjelaskan bahwa DistilBERT dirancang untuk mengurangi kompleksitas model sekaligus mempertahankan sebagian besar kemampuan representasi bahasa BERT. (Silva Barbon & Akabane, 2022) juga mengkaji penggunaan BERT dan DistilBERT dalam penerapan *transfer learning* untuk pemrosesan teks.

Penerapan model keluarga BERT pada tugas pemrosesan teks juga ditunjukkan dalam penelitian Buntoro et al. (2025a) yang menggunakan BioMedBERT dengan pendekatan *domain-adaptive fine-tuning* untuk klasifikasi teks medis. Penelitian lainnya oleh Buntoro et al. (2025b) menerapkan BioMedBERT bersama teknik penyeimbangan data dan *super-ensemble learning* untuk menangani klasifikasi abstrak medis yang tidak seimbang. Kedua penelitian tersebut menunjukkan contoh penerapan model berbasis BERT dalam tugas *Natural Language Processing* (NLP), khususnya klasifikasi teks. Berbeda dari kedua penelitian tersebut, penelitian ini memanfaatkan model DistilBERT yang telah dilatih sebelumnya (*pre-trained*) untuk membentuk representasi semantik dokumen dan kueri pada sistem pencarian informasi, tanpa melakukan pelatihan ulang.

2.4 Representasi Teks dan Text Embedding

Representasi teks merupakan proses mengubah informasi bahasa alami menjadi bentuk numerik agar dapat diproses secara komputasional. Salah satu pendekatannya adalah *text embedding*, yaitu representasi teks dalam ruang vektor yang memungkinkan hubungan antarteks dianalisis berdasarkan kedekatan representasinya. Worth (2023) membahas konsep *word embedding* dan ruang semantik yang menjadi dasar dalam pemodelan hubungan makna pada pemrosesan bahasa alami.

Representasi teks dapat bersifat statis maupun kontekstual. Pada representasi statis, suatu kata memiliki vektor yang relatif tetap, sedangkan representasi kontekstual mempertimbangkan hubungan kata dengan konteks penggunaannya. Model berbasis BERT dan DistilBERT menghasilkan representasi token yang dipengaruhi konteks teks sehingga dapat dimanfaatkan untuk membangun representasi semantik. Sebagai contoh penerapan *deep learning* pada pemrosesan teks bahasa Indonesia, Putra et al. (2022) mengembangkan iQurNet menggunakan *Convolutional Neural Network* (CNN) untuk klasifikasi teks terjemahan Al-Qur'an berbahasa Indonesia. Penelitian tersebut berfokus pada klasifikasi teks, bukan pembentukan *contextual embedding* DistilBERT maupun pencarian semantik.

2.5 Cosine Similarity

Cosine similarity merupakan metode pengukuran kemiripan antara dua vektor berdasarkan sudut yang terbentuk di antara keduanya. Pengukuran ini digunakan dalam sistem temu kembali informasi untuk menentukan tingkat kedekatan representasi vektor kueri dan dokumen. Secara matematis, *cosine similarity* dinyatakan sebagai berikut:

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v11i2

$$\cos(\theta) = (A \cdot B) / (\|A\| \times \|B\|) \quad (1)$$

Keterangan: A merupakan vektor kueri, B merupakan vektor dokumen, A · B menunjukkan hasil perkalian titik (*dot product*), sedangkan ||A|| dan ||B|| merupakan panjang masing-masing vektor. Nilai kemiripan yang lebih tinggi menunjukkan arah kedua vektor yang semakin dekat. Dalam pencarian semantik, pengukuran ini digunakan untuk mengurutkan dokumen berdasarkan kedekatan representasi semantiknya terhadap kueri pengguna (Aleksyuk et al., 2025).

3. METODOLOGI

3.1 Pendekatan Pengembangan Sistem dan Korpus Dokumen

Penelitian ini mengembangkan aplikasi pencarian dokumen arsip universitas berbasis web menggunakan pendekatan *Software Development Life Cycle* (SDLC). Sistem dikembangkan untuk mendukung pengelolaan arsip digital dan proses pencarian dokumen berdasarkan kemiripan semantik. Implementasi aplikasi menggunakan React, Vite, dan Tailwind CSS pada sisi *frontend*, FastAPI pada sisi *backend*, serta PostgreSQL dengan ekstensi pgvector untuk penyimpanan data dan representasi vektor.

Korpus yang digunakan pada evaluasi final terdiri atas 703 entri dokumen dengan metadata dan embedding. Sebanyak 697 dokumen memiliki status PUBLIC dan 6 dokumen berstatus PRIVATE. Berdasarkan sumber pengindeksannya, 702 dokumen berasal dari proses *metadata import*, sedangkan satu dokumen berasal dari mekanisme unggah PDF. Jumlah tersebut merupakan entri dokumen dalam sistem dan tidak menunjukkan bahwa seluruh dokumen memiliki berkas PDF fisik.

Untuk kepentingan evaluasi, seluruh 703 dokumen direpresentasikan menggunakan metadata yang sama, yaitu title, file_name, description, keywords, category, dan year. Evaluasi menggunakan cakupan akses Admin sehingga seluruh dokumen dalam korpus dapat menjadi kandidat pada kedua metode pembanding.

3.2 Arsitektur Sistem dan Persiapan Dokumen

Aplikasi menerapkan arsitektur client-server dengan FastAPI sebagai backend dan React sebagai antarmuka pengguna. Sistem menerapkan hak akses Admin dan Mahasiswa, dengan dokumen PRIVATE dibatasi untuk Admin. Persiapan dokumen dibedakan menjadi dua jalur. Dokumen yang berasal dari *metadata import* diolah menggunakan informasi metadata yang tersedia. Sementara itu, dokumen PDF baru yang diunggah melalui sistem dapat diproses menggunakan strategi Golden Pages, yaitu ekstraksi teks dari halaman pertama dan kedua PDF. Teks hasil ekstraksi dibersihkan dan digabungkan dengan metadata untuk membentuk masukan proses embedding.

Pada implementasi yang dievaluasi dalam penelitian ini, OCR tidak digunakan sebagai sumber representasi dokumen. Mekanisme unggah PDF menggunakan ekstraksi teks Golden Pages pada halaman 1–2, sedangkan evaluasi final menggunakan representasi *metadata-only*. Dengan demikian, informasi hasil ekstraksi PDF tidak digunakan sebagai representasi tambahan pada eksperimen perbandingan kedua metode.

3.3 Model DistilBERT dan Pembentukan Embedding

Model yang digunakan adalah cahya/distilbert-base-indonesian, yaitu model pre-trained yang dimanfaatkan tanpa proses pelatihan ulang. Teks masukan diproses menggunakan tokenizer Hugging Face yang sesuai dengan model, dengan batas panjang 512 token.

Model DistilBERT menghasilkan representasi kontekstual berdimensi 768 untuk setiap token. Representasi tersebut kemudian digabungkan melalui masked

P-ISSN : 2580-4316
E-ISSN : 2654-8054
https://doi.org/10.37817/ikraith-informatika.v11i2

mean pooling, yaitu perhitungan rata-rata representasi token dengan memperhatikan attention mask, sehingga diperoleh satu embedding berdimensi 768 untuk setiap dokumen. Proses yang sama digunakan untuk membentuk embedding kueri, tetapi embedding kueri tidak disimpan secara permanen.

Pada evaluasi final, embedding dokumen dibentuk kembali menggunakan representasi metadata-only dan disimpan sebagai artefak evaluasi terpisah dari embedding produksi. Pemeriksaan panjang teks menunjukkan maksimum 314 token pada korpus evaluator sehingga tidak terdapat dokumen evaluator yang melampaui batas masukan 512 token.

3.4 Semantic Retrieval dan Pemeringkatan

Embedding dokumen pada sistem disimpan menggunakan PostgreSQL dan pgvector dengan tipe vector(768). Pada proses pencarian, kueri pengguna diubah menjadi embedding berdimensi 768, kemudian dibandingkan dengan vektor dokumen menggunakan jarak kosinus. Pada implementasi pgvector, pengukuran ini menggunakan operator <=>, dengan skor semantik dihitung sebagai:

$$\text{semantic_score} = 1 - \text{cosine_distance} \quad (2)$$

Setelah menghasilkan kandidat berdasarkan skor semantik, sistem menerapkan *lexical reranking* untuk menentukan urutan akhir. Tahap ini memeriksa kecocokan istilah pada judul, nama berkas, kata kunci, deskripsi, serta perluasan kueri dan logika topik.

$$\text{ranking_score} = \text{semantic_score} + \text{lexical_boost} \quad (3)$$

Dengan demikian, alur pemeringkatan terdiri atas pembentukan embedding kueri menggunakan DistilBERT, perhitungan *semantic score* menggunakan cosine similarity, pemeringkatan ulang berdasarkan kecocokan leksikal dan topik, serta pengurutan berdasarkan *ranking score* untuk menghasilkan lima dokumen teratas.

Pada konfigurasi implementasi, komponen *lexical boost* memiliki batas maksimum 1,25. Nilai ini merupakan parameter implementasi sistem, bukan hasil penentuan parameter optimal yang berlaku secara umum. Pada antarmuka aplikasi, metode tetap ditampilkan sebagai DistilBERT. Namun, secara teknis pemeringkatan yang digunakan merupakan pencarian semantik berbasis DistilBERT yang dilanjutkan dengan *lexical reranking*, sehingga pada pembahasan eksperimen disebut DistilBERT Hybrid.

3.5 TF-IDF sebagai Metode Pembanding

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan sebagai pembanding berbasis leksikal. Pada evaluasi final, representasi dokumen TF-IDF dibentuk menggunakan field metadata yang sama dengan DistilBERT Hybrid, sedangkan kedua metode menerima teks kueri yang sama.

Pembentukan representasi TF-IDF menggunakan konfigurasi lowercase=True, rentang *n-gram* kata (1,2), min_df=1, dan max_df=0.95. Kueri diproses menggunakan vectorizer yang telah dibentuk dari korpus dokumen. Kemiripan antara kueri dan dokumen dihitung menggunakan cosine similarity, kemudian hasil pencarian diurutkan berdasarkan skor tertinggi.

Kedua metode dibandingkan menggunakan cakupan 703 dokumen yang sama dengan representasi *metadata-only*. Pemeriksaan kesetaraan ruang kandidat dilakukan sebelum pengujian akhir. Lima dokumen teratas dari setiap metode digunakan sebagai hasil pencarian yang dievaluasi.

3.6 Prosedur dan Metrik Evaluasi

Evaluasi final menggunakan benchmark 50 kueri yang dibekukan sebelum *final retrieval*. Kueri V01–V15 mempertahankan kueri Evaluation V1 secara persis, sedangkan V16–V50 merupakan 35 kueri baru yang disusun berdasarkan korpus aktual. Set tersebut

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v11i2

diperlakukan sebagai *Expanded VALIDATION* atau *final evaluation benchmark*, bukan *unseen TEST*. Lima strata desain digunakan dengan masing-masing 10 kueri, yaitu Direct/Lexical, Synonym/Paraphrase, Abbreviation/Domain, Natural Language/Semantic Gap, dan Compound/Topical. Set TEST project tetap disegel dan tidak digunakan.

Acuan relevansi dibangun melalui pendekatan *pooled relevance judgment*. Untuk setiap kueri, Top-20 kedua metode digabung dan dideduplikasi; metode asal, peringkat, dan skor disembunyikan saat penilaian *blind*. Sebanyak 452 judgment V1 yang kompatibel dipertahankan, kemudian 1.007 judgment baru ditambahkan untuk V16–V50. Ground truth final mencakup 1.459 pasangan kueri–dokumen, terdiri atas 117 relevan dan 1.342 tidak relevan, tanpa pasangan berstatus review. Ground truth dibekukan sebelum *final run*; setelah itu tidak dilakukan perubahan label, kueri, pool, representasi, ranking, maupun *lexical cap*.

Kinerja sistem diukur menggunakan Precision@5, Pooled Recall@5, dan F1@5. Precision@5 menyatakan proporsi dokumen relevan pada lima hasil teratas, sedangkan Pooled Recall@5 menyatakan proporsi dokumen relevan dalam lima hasil teratas dibandingkan seluruh dokumen relevan yang terdapat pada *judged pool*. F1@5 merupakan rata-rata harmonik dari kedua nilai tersebut pada masing-masing kueri (Rozaq et al., 2025).

$$\text{Precision@5} = \text{relevan pada Top-5} / 5 \quad (4)$$

$$\text{Pooled Recall@5} = \text{relevan pada Top-5} / \text{relevan dalam pool} \quad (5)$$

$$\text{F1@5} = 2 \times (\text{Precision@5} \times \text{Pooled Recall@5}) / (\text{Precision@5} + \text{Pooled Recall@5}) \quad (6)$$

Karena *relevance judgment* dibangun dari gabungan Top-20 kedua metode dan bukan penilaian menyeluruh terhadap seluruh 703 dokumen, nilai recall dilaporkan sebagai Pooled Recall@5. Satu kueri, V32, tidak memiliki dokumen relevan dalam *judged pool*; untuk kueri ini Precision@5 = 0,0000, sedangkan Pooled Recall@5 dan F1@5 dinyatakan N/A, bukan 0.

Nilai akhir setiap metrik dihitung menggunakan *macro average* per kueri. Precision@5 dihitung pada 50 kueri, sedangkan Pooled Recall@5 dan F1@5 dihitung pada 49 kueri karena V32 dikecualikan dari kedua metrik tersebut. Macro F1@5 merupakan rata-rata F1 per kueri, bukan hasil perhitungan dari macro Precision dan macro Pooled Recall. *Candidate parity* diverifikasi PASS 50/50 dan tidak dilakukan *post-result tuning*.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Implementasi Sistem

Aplikasi pencarian semantik dokumen arsip universitas berhasil diimplementasikan sebagai sistem berbasis web dengan dua peran, yaitu Admin dan Mahasiswa. Sistem menyediakan autentikasi, pencarian dan penyaringan dokumen berdasarkan metadata, detail dokumen, pengunduhan berkas yang tersedia, serta pengelolaan dokumen oleh Admin berupa penambahan, perubahan metadata, dan penghapusan. Dokumen PRIVATE hanya dapat diakses oleh Admin.

Halaman evaluasi Admin membandingkan TF-IDF dan DistilBERT menggunakan kueri yang sama dengan masing-masing lima hasil teratas. Metrik ditampilkan hanya untuk kueri yang memiliki acuan relevansi, sedangkan kueri bebas tetap dapat digunakan untuk membandingkan hasil pencarian.

Sebelum evaluasi final, kedua metode menggunakan ruang kandidat yang sama, yaitu 703 dokumen dengan representasi metadata-only. Kesetaraan ruang kandidat tidak berarti identitas maupun urutan hasil pencarian kedua metode sama.

4.2 Hasil Evaluasi TF-IDF dan DistilBERT Hybrid

Evaluasi final dilakukan terhadap 50 kueri pada *Expanded VALIDATION* menggunakan Precision@5, Pooled Recall@5, dan F1@5. Nilai rata-rata

P-ISSN : 2580-4316
E-ISSN : 2654-8054
https://doi.org/10.37817/ikraith-informatika.v11i2

makro kedua metode disajikan pada Tabel 1.

Tabel 1. Perbandingan hasil makro TF-IDF dan DistilBERT Hybrid

Metrik	TF-IDF	Hybrid
Precision@5	0,2520	0,2240
Pooled Recall@5	0,7453	0,7050
F1@5	0,3372	0,3037

Berdasarkan Tabel 1, TF-IDF memperoleh nilai rata-rata makro lebih tinggi pada ketiga metrik. Selisih Precision@5 sebesar 0,0280, Pooled Recall@5 sebesar 0,0403, dan F1@5 sebesar 0,0335. Pada benchmark final ini, hasil tersebut menunjukkan bahwa DistilBERT Hybrid belum menghasilkan peningkatan kinerja keseluruhan dibandingkan TF-IDF.

Perhitungan Precision@5 menggunakan 50 kueri, sedangkan macro Pooled Recall@5 dan macro F1@5 menggunakan 49 kueri karena V32 tidak memiliki dokumen relevan dalam *judged pool*. Kondisi tersebut dipertahankan apa adanya tanpa mengganti kueri, memaksa label relevan, memperluas pool, atau melakukan tuning ranking setelah hasil diketahui.

Benchmark 50 kueri dirancang seimbang pada lima strata, masing-masing 10 kueri, untuk mencakup pencarian Direct/Lexical, Synonym/Paraphrase, Abbreviation/Domain, Natural Language/Semantic Gap, dan Compound/Topical. Pembagian ini merupakan strata desain evaluasi, bukan bukti bahwa satu kategori tertentu selalu lebih sulit atau lebih mudah bagi salah satu metode.

Kesetaraan ruang kandidat terpenuhi untuk seluruh 50 kueri: kedua metode menggunakan 703 ID dokumen yang sama, representasi *metadata-only* yang sama, kueri yang sama, ground truth yang sama, dan Top-K yang sama. Dengan demikian, perbedaan nilai pada Tabel 1 merefleksikan hasil kedua metode pada kondisi evaluasi yang dikontrol tersebut.

4.3 Pembahasan

Pada benchmark final yang lebih luas, TF-IDF memperoleh macro Precision@5, Pooled Recall@5, dan F1@5 yang lebih tinggi daripada DistilBERT Hybrid. Hasil ini menunjukkan bahwa penambahan representasi semantik melalui DistilBERT Hybrid belum memberikan peningkatan kinerja keseluruhan pada konfigurasi evaluasi *metadata-only* yang digunakan.

Salah satu penjelasan yang mungkin adalah karakteristik metadata evaluator yang terdiri atas *title*, *file_name*, *description*, *keywords*, *category*, dan *year*. Field-field tersebut dapat mengandung istilah yang memiliki overlap langsung dengan kueri, sehingga sinyal leksikal TF-IDF tetap kuat. Namun, penelitian ini tidak melakukan *ablation study* terhadap setiap field metadata, sehingga pengaruh spesifik masing-masing atribut tidak dapat dipastikan.

DistilBERT Hybrid menggunakan model *pre-trained* tanpa *fine-tuning* untuk tugas retrieval/domain dan kemudian menerapkan *lexical reranking*. Pada representasi metadata yang relatif ringkas, pendekatan tersebut belum tentu memberikan keuntungan tambahan yang cukup untuk melampaui baseline leksikal. Pernyataan ini merupakan interpretasi atas kondisi eksperimen, bukan bukti kausal, karena penelitian tidak melakukan eksperimen terpisah terhadap *fine-tuning*, *semantic-only retrieval*, atau masing-masing komponen reranking.

Hasil ini tidak menunjukkan bahwa TF-IDF selalu lebih unggul atau bahwa DistilBERT gagal. Kesimpulan dibatasi pada korpus 703 dokumen, benchmark 50 kueri, representasi *metadata-only*, *pooled relevance judgment*, Top-K = 5, dan konfigurasi sistem yang digunakan. Pooled Recall@5 juga tidak dapat ditafsirkan sebagai *exhaustive recall* terhadap seluruh korpus.

4.4 Keterbatasan Evaluasi

Acuan relevansi dibangun menggunakan *pooled relevance judgment* dari Top-20 TF-IDF ∪ Top-20

DistilBERT Hybrid untuk tiap kueri, bukan penilaian menyeluruh terhadap seluruh 703 dokumen. Dokumen relevan di luar *judged pool* mungkin tidak tercakup, sehingga Pooled Recall@5 hanya menggambarkan cakupan terhadap dokumen relevan yang telah dinilai. Selain itu, V32 tidak memiliki dokumen relevan dalam *judged pool* sehingga recall dan F1 untuk kueri tersebut tidak didefinisikan.

Benchmark final terdiri atas 50 kueri yang disusun secara *stratified corpus-supported*, bukan sampel acak dari lalu lintas pengguna nyata dan bukan *unseen held-out TEST*; V01–V15 juga telah digunakan pada Evaluation V1. Evaluasi menggunakan representasi *metadata-only* dan tidak menguji isi penuh seluruh PDF maupun signifikansi statistik. Kategori kueri berfungsi sebagai strata desain, bukan sebagai bukti perbedaan tingkat kesulitan antarkategori.

5. KESIMPULAN

Penelitian ini berhasil mengimplementasikan aplikasi pencarian semantik dokumen arsip universitas berbasis web menggunakan model *pre-trained* DistilBERT tanpa proses pelatihan ulang. Sistem memanfaatkan *embedding* berdimensi 768, *cosine similarity*, dan *lexical reranking* untuk menghasilkan pemeringkatan dokumen berdasarkan kueri pengguna.

Evaluasi dilakukan pada 703 dokumen dengan representasi *metadata-only* menggunakan benchmark final 50 kueri dan *pooled relevance judgment* sebanyak 1.459 pasangan. TF-IDF memperoleh macro Precision@5 sebesar 0,2520, macro Pooled Recall@5 sebesar 0,7453, dan macro F1@5 sebesar 0,3372; DistilBERT Hybrid memperoleh masing-masing 0,2240, 0,7050, dan 0,3037. Precision dihitung pada 50 kueri, sedangkan Pooled Recall dan F1 pada 49 kueri karena satu kueri tidak memiliki dokumen relevan dalam *judged pool*.

Pada kondisi eksperimen ini, TF-IDF menghasilkan nilai makro yang lebih tinggi daripada DistilBERT Hybrid. Temuan tersebut tidak dapat digeneralisasikan sebagai keunggulan universal TF-IDF karena berlaku pada korpus, query set, *pooling protocol*, representasi *metadata-only*, dan konfigurasi yang digunakan. Penelitian selanjutnya dapat menguji representasi isi dokumen yang lebih lengkap serta kueri yang berasal dari penggunaan nyata untuk menilai kemampuan pencarian semantik pada kondisi yang lebih beragam.

6. UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Universitas Darussalam Gontor dan Program Studi Teknik Informatika atas dukungan selama pelaksanaan penelitian ini. Penulis juga menyampaikan terima kasih kepada dosen pembimbing atas arahan dan bimbingan yang diberikan selama proses penelitian. Secara khusus, penulis mengucapkan terima kasih kepada kedua orang tua yang telah memberikan dukungan moral, doa, serta pembiayaan selama proses pendidikan dan pelaksanaan penelitian ini.

DAFTAR PUSTAKA

Aleksyuk, M. V., Shaburov, A. A., Tereshkin, D. A., & Ushankova, M. Y. (2025). Development and Optimization of a Semantic Search System for Scanned Legal Documents. *2025 XXVIII International Conference on Soft Computing and Measurements (SCM)*, 136–139. https://doi.org/10.1109/SCM66446.2025.11060319

Buntoro, G. A., Putra, O. V., & Purnomo, M. H. (2025a). Domain-Adaptive Fine-Tuning of BioMedBERT for Medical Text Classification. *Engineering, Technology and Applied Science Research*, *15*(6), 28523–28529. https://doi.org/10.48084/etasr.13026

Buntoro, G. A., Putra, O. V., & Purnomo, M. H. (2025b). Hybrid Balancing and Super-Ensemble Learning With BiomedBERT for Imbalanced Medical Abstract Classification. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *14*(3), 515–528.

https://doi.org/10.23887/janapati.v14i3.103811

Jannah, R., Rizkyana, F. W., & Budiantoro, R. A. (2022). Audit of A Web-Based Electronic Documents and Record Management System (WEDRMS): Oversight Efforts to Improve Administration in Higher Educational Institutions. *BISECER (Business Economic Entrepreneurship)*, *5*(2), 53–60. https://doi.org/10.61689/bisecer.v5i2.427

Jian, A. (2022). *The Construction of Digital Management Information System of University Documents and Archives Based on Distributed Architecture*. (Nmdme 2022), 155–160. https://doi.org/10.5220/0011908100003613

Karim, A. A. J., Asad, K. H. M., & Alam, M. G. R. (2025). Larger models yield better results? Streamlined severity classification of ADHD-related concerns using BERT-based knowledge distillation. *PLOS ONE*, *20*(2), e0315829. https://doi.org/10.1371/journal.pone.0315829

Ladanavar, S. M., Kamble, R., Goudar, R. H., Kaliwal, R. B., Rathod, V., Deshpande, S. L., G M, D., & Kulkarni, A. (2024). Enhancing User Query Comprehension and Contextual Relevance with a Semantic Search Engine using BERT and ElasticSearch. *EAI Endorsed Transactions on Internet of Things*, *10*. https://doi.org/10.4108/eetiot.6993

Melo, R., Santos, P. A., & Dias, J. (2023). *A Semantic Search System for the Supremo Tribunal de Justiça* (pp. 142–154). https://doi.org/10.1007/978-3-031-49011-8_12

Mewengkang, A., Sengkey, M. M., Lengkong, J. S. J., & Rotty, V. N. J. (2022). Design and Implementation of Web-Based Archive Management Information System. *International Journal of Information Technology and Education*, *1*(4), 73–81. https://doi.org/10.62711/ijite.v1i4.80

Putra, O. V., Elfianita, F. D., Harmini, T., Trisnani, A., & Nugroho, A. (2022). iQurNet: A Deep Convolutional Neural Network for Text Classification on the Indonesian Holy Quran Translation. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 86–91. https://doi.org/10.1109/ISMODE53584.2022.9743132

Rozaq, A., Fitri, R., Akbar, M. S., Sumari, M. H., Khafiyya, N., Wardiah, I., & Sholihin, F. (2025). Legal Literacy in Indonesia: Leveraging Semantic-Based AI and NLP for Enhanced Civil Law Access. *E3S Web of Conferences*, *622*, 03002. https://doi.org/10.1051/e3sconf/202562203002

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, *22*(21), 8184. https://doi.org/10.3390/s22218184

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys*, *56*(7), 1–33. https://doi.org/10.1145/3648471

Worth, P. J. (2023). Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, *13*(01), 1–21. https://doi.org/10.4236/ijis.2023.131001