Ranah Research
Journal of Multidisciplinary Research and Development
E-ISSN: 2655-0865
082170743613 ranahresearch@gmail.com https://jurnal.ranahresearch.com
DINASTI RESEARCH

DOI: https://doi.org/10.38035/rrj.v8i1

Penerapan *Sentence-Bert* dan *Cosine Similarity* untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF

Muhammad Abdul Hafizh Fathuddin[1], Eka Prakarsa Mandyartha[2], Afina Lina Nurlaili[3]

[1]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, 21081010225@student.upnjatim.ac.id
[2]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, eka_prakarsa.fik@upnjatim.ac.id
[3]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, afina.lina.if@upnjatim.ac.id

Corresponding Author: 21081010225@student.upnjatim.ac.id [1]

Abstract: *The search for thesis documents in digital repositories is generally limited to keyword matching, which often produces less relevant results. To address this issue, this study develops a semantic search system for thesis documents in PDF format by utilizing Sentence-BERT (SBERT) and the Cosine Similarity method, combined with ontology to enrich the understanding of query meanings. The research stages include text extraction from PDF documents, preprocessing, WordPiece tokenization, and sentence vector representation using SBERT, with relevance scores calculated by combining cosine similarity (0.7) and ontology (0.3) weights. The evaluation results show that the system is capable of producing relevant search results with a consistent Mean Reciprocal Rank (MRR) of 1.0 across all query types. The average Precision reached 0.80, while the average Recall was 0.92. A comparison with the Keyword Matching method shows that the semantic approach performs better, with an average Precision of 0.88 and Recall of 0.65, compared to keyword matching which only achieved 0.24 for Precision and 0.12 for Recall. These findings demonstrate that the semantic system effectively places the most relevant documents at the top rank and outperforms keyword-based search, although the coverage of relevant results still needs to be improved through ontology enrichment and dataset expansion.*

Keywords: *Semantic Search, Sentence-BERT, Cosine Similarity, Ontology, Thesis Documents.*

Abstrak: Pencarian dokumen skripsi pada repositori digital umumnya masih terbatas pada pencocokan kata kunci sehingga sering menghasilkan temuan yang kurang relevan. Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membangun sistem pencarian semantik dokumen skripsi dalam format PDF dengan memanfaatkan Sentence-BERT (SBERT) dan metode Cosine Similarity yang dipadukan dengan ontologi untuk memperkaya pemahaman makna query. Sistem ini dirancang agar mampu memahami

maksud pengguna secara lebih mendalam, baik ketika query diberikan dalam bentuk kata, frasa, kalimat, maupun paragraf. Tahapan penelitian meliputi ekstraksi teks dari dokumen PDF, preprocessing, tokenisasi WordPiece, serta pembentukan vektor representasi kalimat menggunakan SBERT. Skor relevansi dihitung dengan kombinasi bobot cosine similarity (0,7) dan ontologi (0,3) sehingga sistem dapat menampilkan dokumen dengan makna paling mendekati query. Hasil pengujian menunjukkan bahwa sistem mampu memberikan hasil pencarian yang relevan dengan nilai Mean Reciprocal Rank (MRR) konsisten sebesar 1.0 pada semua jenis query. Nilai Precision rata-rata mencapai 0,80 dan Recall rata-rata sebesar 0,92. Perbandingan dengan metode Keyword Matching menunjukkan bahwa metode semantik lebih unggul dengan Precision rata-rata 0,88 dan Recall 0,65 dibandingkan keyword yang hanya mencapai Precision 0,24 dan Recall 0,12. Temuan ini membuktikan bahwa sistem semantik efektif dalam menempatkan dokumen paling relevan di peringkat teratas dan lebih unggul dibandingkan pencarian berbasis kata kunci, meskipun cakupan hasil masih perlu ditingkatkan melalui pengayaan ontologi dan perluasan dataset.

Kata kunci: Pencarian Semantik, Sentence-BERT, Cosine Similarity, Ontology, Dokumen Skripsi.

PENDAHULUAN

Repositori skripsi dan tesis merupakan salah satu aset penting dalam dunia pendidikan tinggi. Dokumen-dokumen ini memuat hasil penelitian mahasiswa yang tidak hanya mencerminkan pemahaman akademik, tetapi juga dapat dijadikan rujukan dan inspirasi untuk penelitian lanjutan. Dalam era digital saat ini, mayoritas dokumen tersebut disimpan dalam format digital, khususnya PDF, dan tersedia melalui sistem repositori online milik universitas. Namun, semakin banyaknya jumlah dokumen yang diunggah setiap tahun juga menimbulkan tantangan baru, yaitu bagaimana cara menelusuri dan menemukan informasi akademik yang relevan secara cepat dan akurat.

Selama ini, pencarian dokumen digital masih mengandalkan metode berbasis kata kunci. Pendekatan ini memiliki kelemahan mendasar, yaitu hanya mencocokkan kata secara literal tanpa mempertimbangkan makna atau hubungan semantik antar kata [6]. Akibatnya, query yang diberikan pengguna sering kali menghasilkan hasil pencarian yang kurang relevan. Hal ini sejalan dengan penelitian Rahman et al. (2015), yang menunjukkan bahwa sistem pencarian berbasis kata kunci sering kali tidak dapat menangkap makna sebenarnya dari sebuah teks, sehingga mengurangi efektivitas pencarian [1]. Selain itu, penelitian terbaru oleh Susanto et al. (2018) juga menyoroti pentingnya representasi semantik dalam meningkatkan akurasi pencarian informasi dalam dokumen digital [2].

Sebagai solusi, teknologi pencarian semantik menawarkan pendekatan yang lebih canggih. pencarian semantik memungkinkan sistem pencarian memahami makna query secara lebih mendalam dengan menangkap hubungan semantik antara query dan dokumen. Menurut penelitian Reimers dan Gurevych (2019), penggunaan model Sentence-BERT terbukti mampu meningkatkan relevansi hasil pencarian dengan mengubah teks menjadi representasi embedding yang dapat diproses secara komputasi [3]. Selain itu, penelitian Amien (2023) menjelaskan bagaimana perkembangan NLP dalam Bahasa Indonesia, termasuk penerapan model transformer, mampu meningkatkan kemampuan sistem pencarian berbasis semantik di berbagai konteks [4].

Sementara itu, perkembangan NLP di Indonesia juga menunjukkan hal positif dalam mendukung pencarian semantik. Menurut Amien (2023), penerapan model transformer seperti SBERT dalam Bahasa Indonesia semakin relevan karena tersedianya dataset lokal dan pretrained model multilingual yang mendukung pemrosesan bahasa Indonesia [4]. Penelitian

Wibawa dan Anggraeni (2023) juga menunjukkan bahwa integrasi NLP dalam *preprocessing* dokumen akademik, seperti segmentasi teks dan penghilangan elemen non-informasi, mampu meningkatkan hasil ekstraksi informasi pada dokumen PDF [1]. Bahkan, pada konteks non-akademik seperti analisis media sosial, penggunaan pendekatan semantik juga mulai diterapkan untuk memahami makna kompleks dalam teks seperti ditunjukkan oleh penelitian Nur Oktavia et al. (2024) yang menganalisis tweet buzzer menggunakan NLP [2].

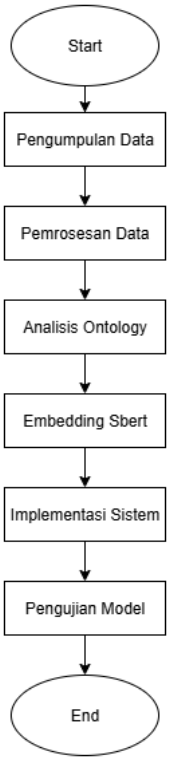
Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk mengembangkan sistem pencarian semantik berbasis Natural Language Processing (NLP) yang dirancang untuk dokumen skripsi dalam format PDF. Sistem akan memanfaatkan model Sentence-BERT untuk mengubah kalimat dalam dokumen menjadi vektor embedding yang menangkap makna semantik. Proses pencarian dilakukan dengan membandingkan vektor embedding dari query pengguna dengan seluruh embedding dalam korpus dokumen menggunakan cosine similarity. Dengan pendekatan ini, diharapkan sistem dapat memberikan hasil pencarian yang lebih relevan dan kontekstual dibandingkan pencarian berbasis kata kunci. Sistem ini akan dibangun menggunakan bahasa pemrograman Python dan dievaluasi menggunakan metrik evaluasi pencarian seperti Mean Reciprocal Rank (MRR), Precision, dan Recall untuk mengukur efektivitas hasil pencarian.

Dengan pendekatan ini, diharapkan sistem dapat memberikan hasil pencarian yang lebih relevan dan akurat, sehingga memudahkan pengguna, terutama mahasiswa dan peneliti, dalam menemukan dokumen akademik yang sesuai dengan kebutuhan mereka.

METODE

Rancangan Penelitian

Pada Gambar 1 digambarkan alur penelitian yang dilakukan dalam pengembangan sistem pencarian semantik dokumen skripsi dalam format PDF. Penelitian ini dimulai dengan proses pengumpulan data, yaitu mengumpulkan sejumlah dokumen skripsi dari berbagai sumber, seperti repositori akademik dan situs publik yang menyediakan akses terhadap dokumen akademik. Seluruh dokumen yang dikumpulkan berbentuk file PDF dan akan digunakan sebagai korpus utama dalam sistem pencarian.



Gambar 1. Alur Penelitian

Setelah dokumen terkumpul, tahapan selanjutnya adalah pemrosesan data. Pada tahap ini, teks diekstraksi dari setiap halaman dokumen PDF menggunakan library PyMuPDF (fitz). Proses ekstraksi dilakukan untuk mengambil isi teks secara utuh dari file PDF. Setelah teks berhasil diekstraksi, dilakukan proses pembersihan dan normalisasi, seperti menghapus karakter yang tidak relevan, menghilangkan header/footer, menghapus stopwords, melakukan konversi ke huruf kecil, serta stemming menggunakan Sastrawi agar format teks menjadi lebih seragam dan siap untuk diproses secara semantik.

Tahap berikutnya adalah analisis ontologi, yaitu proses untuk memperluas makna query maupun dokumen dengan menggunakan representasi konsep dari ontologi. Pada tahap ini, sistem memeriksa apakah query pengguna mengandung konsep yang telah didefinisikan dalam ontologi. Jika ditemukan kecocokan, query diperluas dengan menambahkan sinonim atau istilah terkait dari konsep tersebut. Selain itu, sistem juga menandai kalimat pada dokumen dengan konsep yang relevan. Hasil dari analisis ini digunakan untuk memberikan bobot tambahan pada dokumen yang memiliki kesamaan konsep dengan query, sehingga pencarian menjadi lebih relevan secara semantik maupun konseptual.

Tahapan selanjutnya adalah embedding, yaitu proses mengubah kalimat-kalimat yang telah dibersihkan menjadi representasi vektor menggunakan model Sentence-BERT (SBERT). Vektor embedding ini merepresentasikan makna dari setiap kalimat dalam bentuk angka berdimensi tetap. Setelah proses embedding, dilakukan L2-normalisasi untuk memastikan bahwa seluruh vektor memiliki panjang (magnitudo) yang sama, sehingga perbandingan makna antar kalimat dapat dilakukan secara konsisten menggunakan metode cosine similarity.

Setelah semua embedding dihasilkan dan dinormalisasi, sistem siap melakukan proses pencarian. Pengguna dapat memberikan query pencarian dalam bentuk kalimat. Query ini akan diproses dengan tahapan yang sama (pembersihan, analisis ontologi, embedding, dan normalisasi), kemudian dibandingkan dengan seluruh vektor embedding dokumen menggunakan perhitungan cosine similarity. Hasil pencarian berupa kalimat-kalimat yang memiliki tingkat kemiripan makna paling tinggi dengan query.

Tahapan berikutnya adalah pengujian dan evaluasi sistem, di mana dilakukan penilaian terhadap performa sistem pencarian. Evaluasi dilakukan dengan mengukur seberapa akurat hasil pencarian terhadap query tertentu menggunakan metrik seperti cosine similarity, Mean Reciprocal Rank (MRR), Precision, dan Recall. Selain itu, dilakukan evaluasi manual dengan membandingkan hasil sistem dengan interpretasi manusia untuk menilai relevansi makna.

Tahapan terakhir adalah analisis hasil, yang bertujuan untuk menginterpretasikan efektivitas sistem dari sisi akurasi, kecepatan pencarian, dan konsistensi hasil. Apabila ditemukan kelemahan, maka dilakukan penyempurnaan terhadap proses preprocessing, analisis ontologi, maupun representasi teks untuk meningkatkan kualitas pencarian.

HASIL DAN PEMBAHASAN

Skenario Pengujian

Sebelum sistem benar-benar digunakan oleh pengguna, diperlukan serangkaian skenario pengujian untuk memastikan bahwa semua fitur bekerja sebagaimana mestinya. Skenario ini disusun agar bisa menggambarkan kondisi nyata ketika seseorang melakukan pencarian dokumen skripsi menggunakan sistem pencarian semantik yang telah dibangun.

Pengujian ini dimulai dari yang paling singkat (1 kata) sampai yang cukup panjang seperti abstrak. Dengan cara ini, kita bisa melihat sejauh mana sistem memahami maksud pencarian dan menemukan dokumen yang benar-benar relevan. Setiap skenario diujikan pada koleksi dokumen yang sudah di-*embedding* sebelumnya menggunakan Sentence-BERT, lalu hasilnya diukur dengan metrik *Precision*, *Recall*, dan *Mean Reciprocal Rank (MRR)*.

Tabel 1. Contoh Query untuk Pengujian

No	Jenis Query	Query
1	Kata Tunggal	Cyber, System, Mobile, Leukemia.
2	Dua Kata	Data Mining, Klasifikasi Gambar.
3	Kalimat	Penyakit sel darah putih disebabkan oleh pertumbuhan abnormal pada sumsum tulang.
4	Paragraf	Efektivitas sistem pencarian semantik modern sangat bergantung pada sinergi antara Natural Language Processing (NLP) dan machine learning untuk memahami konteks informasi yang tersebar di web. Dalam domain keamanan siber (*cybersecurity*), kemampuan ini bukan lagi sebuah kemewahan, melainkan kebutuhan krusial. Sistem pencarian tradisional yang berbasis kata kunci akan gagal ketika seorang analis keamanan mencoba menemukan ancaman yang kompleks.
5	Abstrak	Rekomendasi pekerjaan berbasis Natural Language Processing (NLP) menggunakan model Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM). Data penelitian diperoleh dari dataset publik Hugging Face dan Kaggle, yang mencakup deskripsi pengalaman kerja dan keterampilan. Data diproses melalui tahapan pra pengolahan seperti penghapusan stopword, regex, stemming, tokenisasi, padding, serta teknik balancing seperti SMOTE dan undersampling. Dataset dibagi untuk pelatihan dan pengujian model, dengan performa diukur melalui akurasi, loss, dan metrik validasi. Hasil menunjukkan bahwa model CNN pada dataset Hugging face memberikan hasil terbaik, dengan akurasi 95%, loss 0,1, validation accuracy 72%, validation loss 1,87, F1-score 79%, presisi 90%, dan recall 72%, mengungguli LSTM yang cenderung overfitting. Secara keseluruhan, CNN terbukti lebih stabil dan andal dalam generalisasi, terutama dalam menangani data yang kompleks. Penelitian ini juga mengonfirmasi bahwa CNN lebih efektif dalam memproses data kompleks dibandingkan LSTM, yang cenderung lebih rentan terhadap overfitting, terutama pada data teks panjang

Tabel 1 menampilkan contoh-contoh query yang digunakan pada tahap pengujian sistem pencarian semantik. Query ini dibuat dengan variasi tingkat kompleksitas, mulai dari kata tunggal hingga abstrak, dengan tujuan untuk menguji performa sistem dalam berbagai skenario pencarian.

Pada baris pertama, query berupa kata tunggal seperti *Cyber*, *System*, dan *NLP* digunakan untuk melihat kemampuan sistem dalam menemukan dokumen yang mengandung istilah spesifik secara langsung. Baris kedua menggunakan dua kata seperti *Data Mining* dan *Klasifikasi Gambar*, yang menguji sistem dalam mengidentifikasi kombinasi kata atau frasa yang umum digunakan pada topik tertentu.

Selanjutnya, pada baris ketiga, query berupa kalimat "Keamanan data pengguna menjadi prioritas utama dalam pengembangan sistem ini" digunakan untuk menguji sistem dalam memahami makna utuh dari sebuah pernyataan, bukan sekadar mencocokkan kata per kata. Pada baris keempat, query berbentuk paragraf yang membahas perkembangan teknologi informasi dan pengelolaan data akademik digunakan untuk menguji kemampuan sistem memahami konteks yang lebih panjang dan kompleks.

Baris terakhir menampilkan query berupa abstrak dari sebuah penelitian machine learning dengan algoritma YOLO. Abstrak dipilih sebagai bentuk query paling kompleks karena memuat informasi detail dan beragam istilah teknis. Hal ini memungkinkan evaluasi terhadap kemampuan sistem dalam menangani teks panjang dengan beragam kata kunci dan makna yang saling berkaitan.

Dengan variasi query ini, pengujian dapat memberikan gambaran menyeluruh tentang performa sistem dalam skenario pencarian yang beragam, mulai dari pencarian sederhana hingga pencarian yang memerlukan pemahaman konteks yang lebih mendalam.

Hasil Pengujian

Setelah tahap implementasi selesai, sistem pencarian semantik diuji untuk mengukur performa dalam menampilkan hasil pencarian yang relevan. Pengujian dilakukan dengan dua pendekatan, yaitu otomatis dan manual. Pendekatan otomatis dilakukan dengan cara sistem mengambil ground truth langsung dari koleksi dokumen PDF yang ada. Sementara itu, pada pendekatan manual, ground truth ditentukan oleh peneliti berdasarkan pemahaman konteks isi dokumen.

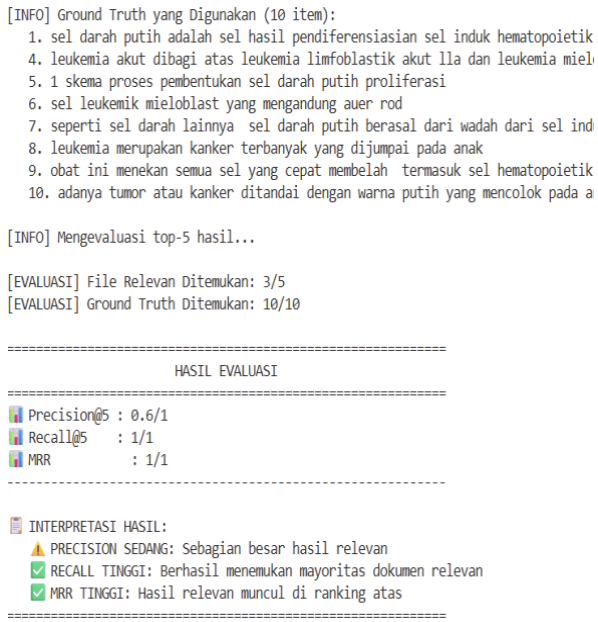
Tujuan dari penggunaan dua metode ini adalah untuk memberikan gambaran yang lebih menyeluruh mengenai kinerja sistem. Dengan pengujian otomatis, penilaian lebih konsisten karena bersumber langsung dari data. Sedangkan pengujian manual memberikan fleksibilitas dalam menilai relevansi hasil yang mungkin tidak terdeteksi oleh sistem.

Query yang diuji berbeda dari kata tunggal, dua kata, satu kalimat, satu paragraf, dan satu abstrak. Pengujian ini bertujuan untuk mengevaluasi kemampuan sistem dalam memahami berbagai tingkat kompleksitas input dari pengguna.

Pengujian Oleh Sistem

Pada pengujian oleh sistem, sistem secara langsung mencocokkan query dengan dokumen yang tersedia dalam dataset. Hasil pencarian kemudian dibandingkan dengan ground truth yang sudah dipetakan ke dalam setiap query. Sistem diminta untuk mengembalikan lima hasil teratas (top-5), lalu dievaluasi menggunakan metrik Precision, Recall, dan Mean Reciprocal Rank (MRR).

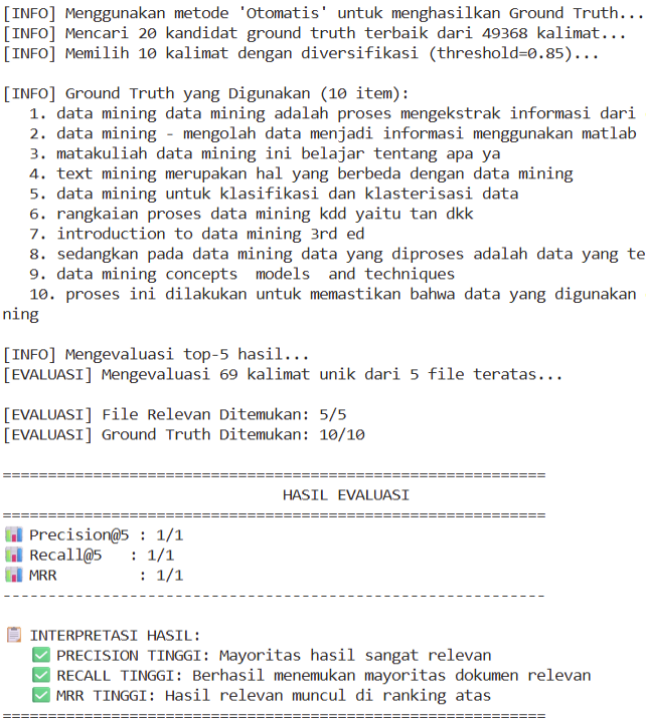
1. Kata Tunggal



Gambar 2. Hasil Pengujian Kata Tunggal Otomatis

Gambar 2 memperlihatkan hasil pengujian sistem menggunakan query *leukemia* yang berupa kata tunggal. Berdasarkan evaluasi, nilai Precision sebesar 0.6 menunjukkan bahwa tiga dari lima hasil teratas merupakan dokumen relevan, sedangkan nilai Recall sebesar 1.0 mengindikasikan bahwa seluruh dokumen relevan berhasil ditemukan. Selain itu, nilai MRR sebesar 1.0 menunjukkan bahwa dokumen paling relevan selalu muncul pada peringkat pertama. Secara keseluruhan, hasil ini membuktikan bahwa sistem mampu memahami makna query satu kata dengan baik, karena mampu menemukan semua dokumen relevan dan menempatkan hasil terbaik pada posisi teratas meskipun masih terdapat sedikit ketidaktepatan pada sebagian hasil.

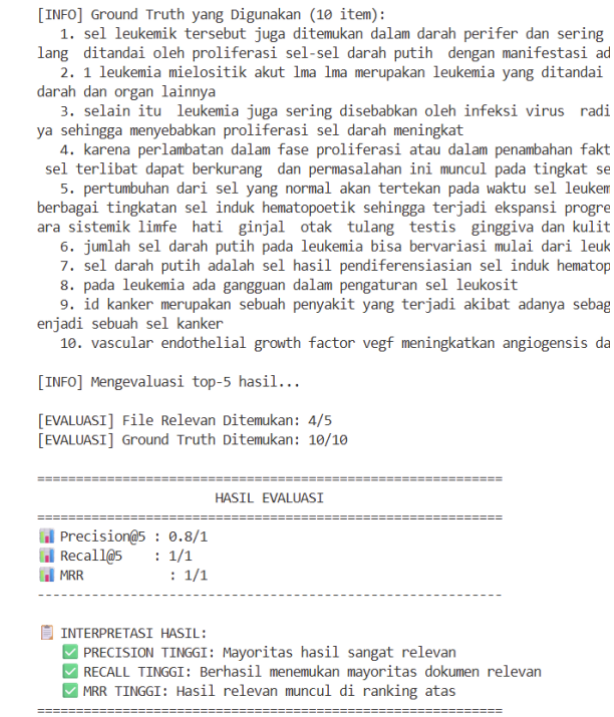
2. Dua Kata



Gambar 3. Hasil Pengujian Dua Kata Otomatis

Gambar 3 memperlihatkan hasil pengujian dengan query berupa dua kata. Evaluasi menunjukkan kinerja sistem yang sangat optimal, dengan precision, recall, dan MRR yang masing-masing bernilai **1,0**. Hal ini mengindikasikan bahwa seluruh hasil yang ditampilkan relevan dan dokumen paling relevan selalu berada pada peringkat pertama. Temuan ini membuktikan bahwa sistem mampu menangani query berupa frasa pendek dengan akurasi yang sangat baik dan konsisten.

3. Kalimat



Gambar 4. Hasil Pengujian Kalimat Otomatis

Gambar 4 memperlihatkan hasil pengujian sistem menggunakan query terkait *sel darah putih*. Berdasarkan evaluasi, nilai Precision sebesar 0.8 menunjukkan bahwa empat dari lima hasil teratas merupakan dokumen relevan dengan query yang diberikan. Nilai Recall sebesar 1.0 mengindikasikan bahwa seluruh dokumen relevan berhasil ditemukan dalam hasil pencarian, sedangkan nilai MRR sebesar 1.0 menunjukkan bahwa dokumen paling relevan muncul pada peringkat pertama. Secara keseluruhan, hasil ini menunjukkan bahwa sistem mampu menampilkan dokumen yang relevan secara konsisten, dengan cakupan pencarian yang lengkap dan penempatan hasil paling relevan pada posisi teratas.

4. Paragraf

```
[INFO] Menggunakan metode 'Otomatis' untuk menghasilkan Ground Truth...
[INFO] Mencari 20 kandidat ground truth terbaik dari 49368 kalimat...
[INFO] Memilih 10 kalimat dengan diversifikasi (threshold=0.85)...

[INFO] Ground Truth yang Digunakan (10 item):
   1. kemudian  analisis dilakukan dengan melihat pengaruh faktor pendi
an hasil yang tidak memiliki pengaruh yang signifikan
   2. bagi peneliti a hal ini dimaksudkan agar penelitian ini dapat meml
n cyber pada pengguna e-wallet dipengaruhi oleh data protection  digita
   3. pengertian serangan cyber serangan cyber yakni suatu perlakuan ya
omputer dan semua orang dapat terkena dampak dari ancaman tersebut
   4. menurut pla daily  unit ini dirancang untuk menjadi sumber peneli
ang sangat efisien dan dinamis  menyediakan layanan high-end untuk isu-
eristik dunia maya
   5. untuk menganalisis apakah data protection  digital literacy  dan
et dikalangan generasi millennial
   6. bagi masyarakat penelitian ini diharapkan akan memberikan pemaham
adap risiko keamanan ancaman cyber
   7. 1 natural language processing nlp natural language processing nlp
na komputasi digunakan secara luas patel prajapati  2018  yang ditujuka
   8. strategi-strategi tersebut akan dianalisa dan dielaborasi menggun
elitian ini guna mendapatkan jawaban yang komprehensif dari pertanyaan
   9. istilah natural language processing nlp pada awalnya disebut seba
 benar-benar 15 memahami bahasa alami semirip mungkin dengan manusia pe
   10. pengaruh data protection  digital literacy dan cyber security te
ra dimana untuk melihat bagaimana antara belief  attitude  intention da
isms

[INFO] Mengevaluasi top-5 hasil...
[EVALUASI] Mengevaluasi 305 kalimat unik dari 5 file teratas...

[EVALUASI] File Relevan Ditemukan: 4/5
[EVALUASI] Ground Truth Ditemukan: 6/10

============================================================
                        HASIL EVALUASI
============================================================
Precision@5 : 0.8/1
Recall@5    : 0.6/1
MRR         : 1/1
------------------------------------------------------------
```

Gambar 5. Hasil Pengujian Paragraf Otomatis

Gambar 5 memperlihatkan hasil pengujian dengan query berupa paragraf. Hasil pengujian menunjukkan precision dan recall masing-masing bernilai 0,8, yang menandakan bahwa sebagian besar hasil pencarian relevan dengan query yang diberikan. Nilai MRR yang tetap 1,0 menunjukkan bahwa dokumen dengan relevansi tertinggi selalu ditampilkan pada peringkat teratas. Hal ini menunjukkan bahwa sistem dapat menangani input berupa paragraf dengan baik dan menghasilkan hasil pencarian yang stabil.

5. Abstrak

```
[INFO] Menggunakan metode 'Otomatis' untuk menghasilkan Ground Truth...
[INFO] Mencari 20 kandidat ground truth terbaik dari 49368 kalimat...
[INFO] Memilih 10 kalimat dengan diversifikasi (threshold=0.85)...

[INFO] Ground Truth yang Digunakan (10 item):
   1. implementasi model natural language processing nlp pada sistem re
ory lstm skripsi diajukan untuk melengkapi salah satu syarat memperoleh
   2. sebaliknya  model pembelajaran mendalam seperti convolutional neu
i data yang kompleks dan berurutan
   3. 000 vi kata pengantar puji syukur penulis ucapkan kehadirat allah
akhir dengan berjudul implementasi model natural language processing nl
g short term memory lstm
   4. input data pada penelitian ini adalah data dari media sosial yang
eh pakar pada bidang bahasa  sedangkan outputnya adalah hasil akurasi d
   5. haidir fikri nim  09011282025064 judul  implementasi model natura
ral networks cnn dan long short term memory lstm hasil pengecekkam soft
rya saya sendiri dan bukan hasil penjiplakan atau plagiat
   6. pada penelitian ini menggunakan natural language processing untuk
ral network untuk melakukan training dan prediksi jawaban dari pertanya
   7. mengetahui proses preprocessing data bidang nlp yang mencakup pen
   8. data diproses melalui tahapan pra pengolahan seperti penghapusan
undersampling
   9. contoh bag of words setelag melakukan preprocessing data selanjut
 dibutuhkan jumlah layer dan activation function
   10. 2 perplexity ai perplexity ai adalah jenis artificial intelligen
ing yang dapat membantu pengguna dalam aspek penulisan seperti pencaria

[INFO] Mengevaluasi top-5 hasil...
[EVALUASI] Mengevaluasi 216 kalimat unik dari 5 file teratas...

[EVALUASI] File Relevan Ditemukan: 4/5
[EVALUASI] Ground Truth Ditemukan: 10/10

============================================================
                        HASIL EVALUASI
============================================================
Precision@5 : 0.8/1
Recall@5    : 1/1
MRR         : 1/1
------------------------------------------------------------
```

Gambar 6. Hasil Pengujian Abstrak Otomatis

Gambar 6 memperlihatkan hasil pengujian otomatis menggunakan query berupa abstrak. Hasil pengujian menunjukkan nilai precision sebesar 0,8 dan recall sebesar 1,0, yang berarti seluruh dokumen relevan berhasil ditemukan, meskipun terdapat sedikit hasil yang kurang akurat pada peringkat bawah. Nilai MRR sebesar 1,0 menunjukkan bahwa dokumen paling relevan selalu muncul di posisi teratas. Hasil ini membuktikan bahwa sistem mampu menangani query yang panjang dan kompleks, seperti abstrak, dengan performa yang sangat baik dan konsisten.

Tabel 2. Hasil Pengujian Oleh System

No	Jenis Query	Precission	Recall	MRR
1	Kata Tunggal	0.6	1	1
2	Dua Kata	1	1	1
3	Kalimat	0.8	1	1
4	Paragraf	0.8	0.6	1
5	Abstrak	0.8	1	1
Rata - Rata		0.8	0.92	1

Tabel 2 memperlihatkan hasil pengujian sistem pencarian semantik berdasarkan lima jenis query, yaitu kata tunggal, dua kata, kalimat, paragraf, dan abstrak. Berdasarkan hasil pengujian, nilai MRR pada semua jenis query bernilai 1, yang berarti dokumen paling relevan

selalu muncul pada posisi pertama hasil pencarian. Hal ini menunjukkan bahwa sistem mampu menempatkan informasi paling relevan secara konsisten di peringkat teratas.

Nilai Precision memiliki rata-rata sebesar 0.8, dengan variasi antara 0.6 hingga 1.0. Artinya, sebagian besar hasil pencarian yang ditampilkan sistem sudah relevan dengan query yang diberikan, meskipun pada query kata tunggal masih terdapat hasil yang kurang sesuai.

Sementara itu, nilai Recall menunjukkan rata-rata sebesar 0.92, yang berarti sistem berhasil menemukan hampir seluruh dokumen relevan dalam top-5 hasil pencarian. Namun, pada jenis query paragraf nilai Recall menurun menjadi 0.6, sehingga masih ada beberapa dokumen relevan yang terlewat.

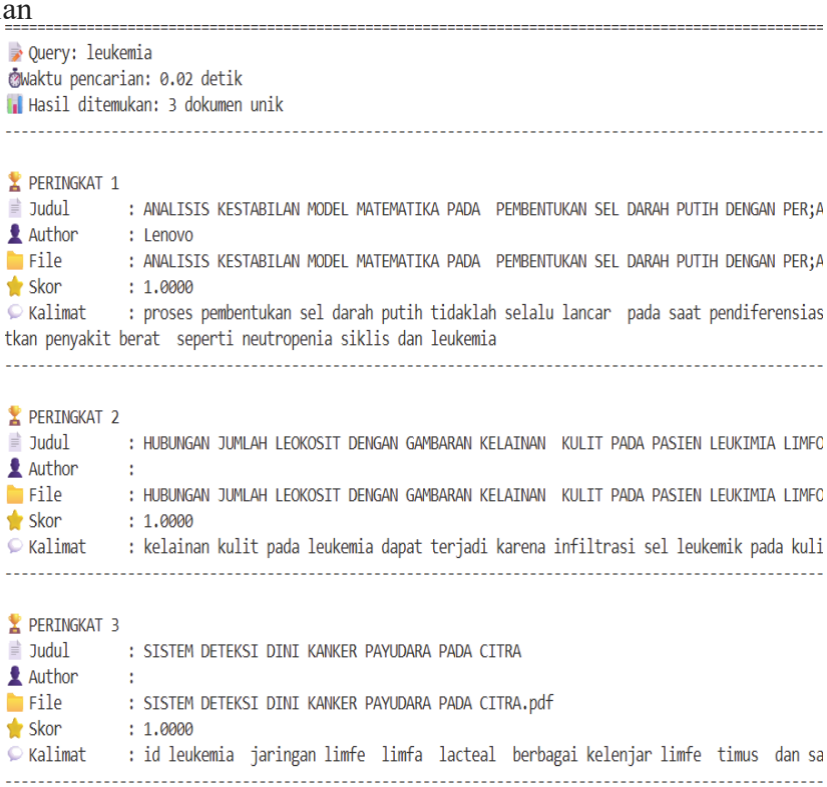
Dengan demikian, dapat disimpulkan bahwa sistem sudah bekerja cukup baik dalam memberikan hasil pencarian yang relevan, namun performa masih dapat ditingkatkan khususnya dalam cakupan hasil pencarian yang lebih luas.

Perbandingan Hasil Dengan Metode Lain

Pada bagian ini dilakukan perbandingan antara sistem yang dibangun dengan pendekatan Sentence-BERT + Cosine Similarity terhadap metode pencarian query biasa (keyword matching).

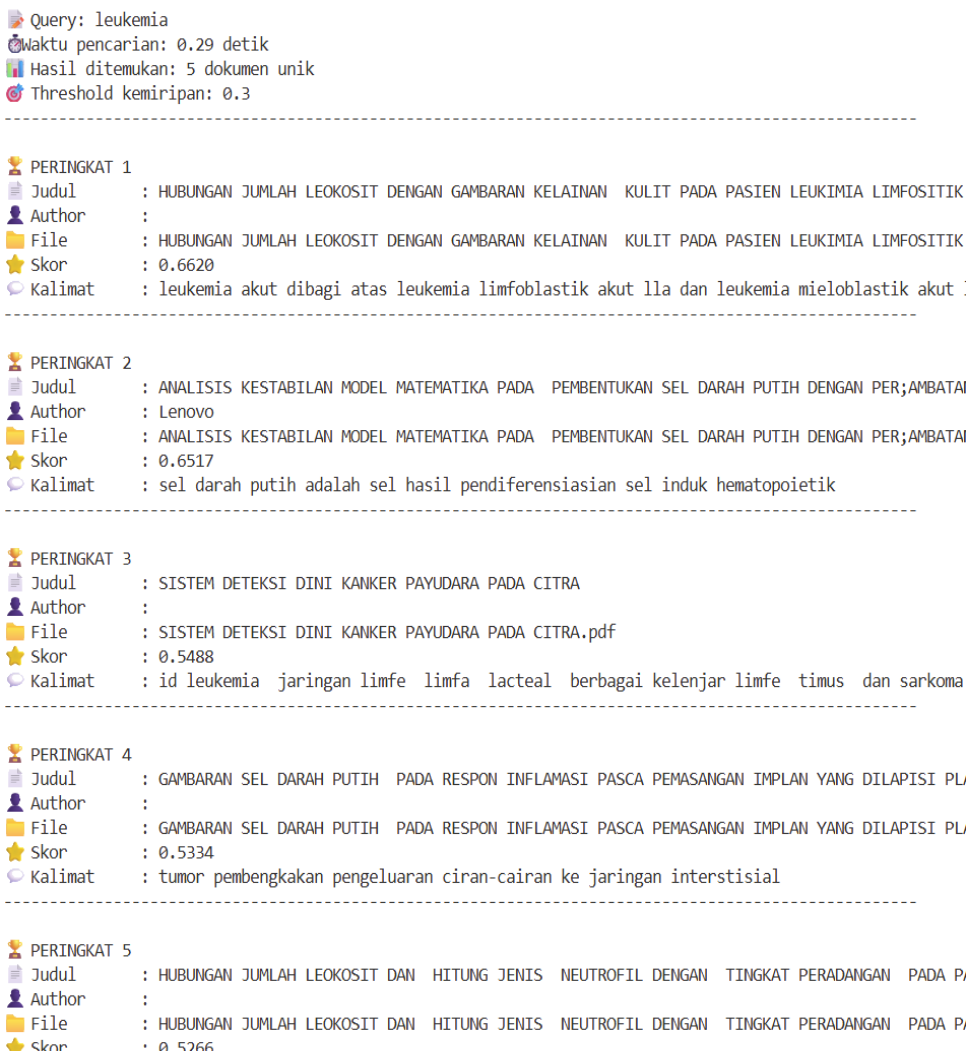
Tujuan perbandingan ini adalah untuk membuktikan bahwa sistem yang diajukan lebih unggul dalam memahami makna query dibandingkan pencarian berbasis kecocokan kata semata.

1. Hasil Pencarian



Gambar 7. Hasil Pencarian Oleh Keyword Base

Pada Gambar 7 ditampilkan hasil pencarian menggunakan metode keyword dengan query *"leukemia"*. Sistem mampu menemukan tiga dokumen yang mengandung kata persis *"leukemia"*. Skor yang dihasilkan bernilai 1.0 karena metode ini hanya menghitung kecocokan literal antar kata. Waktu eksekusi relatif sangat cepat, yaitu 0.02 detik, namun hasilnya terbatas hanya pada dokumen yang memiliki kata yang sama persis dengan query.

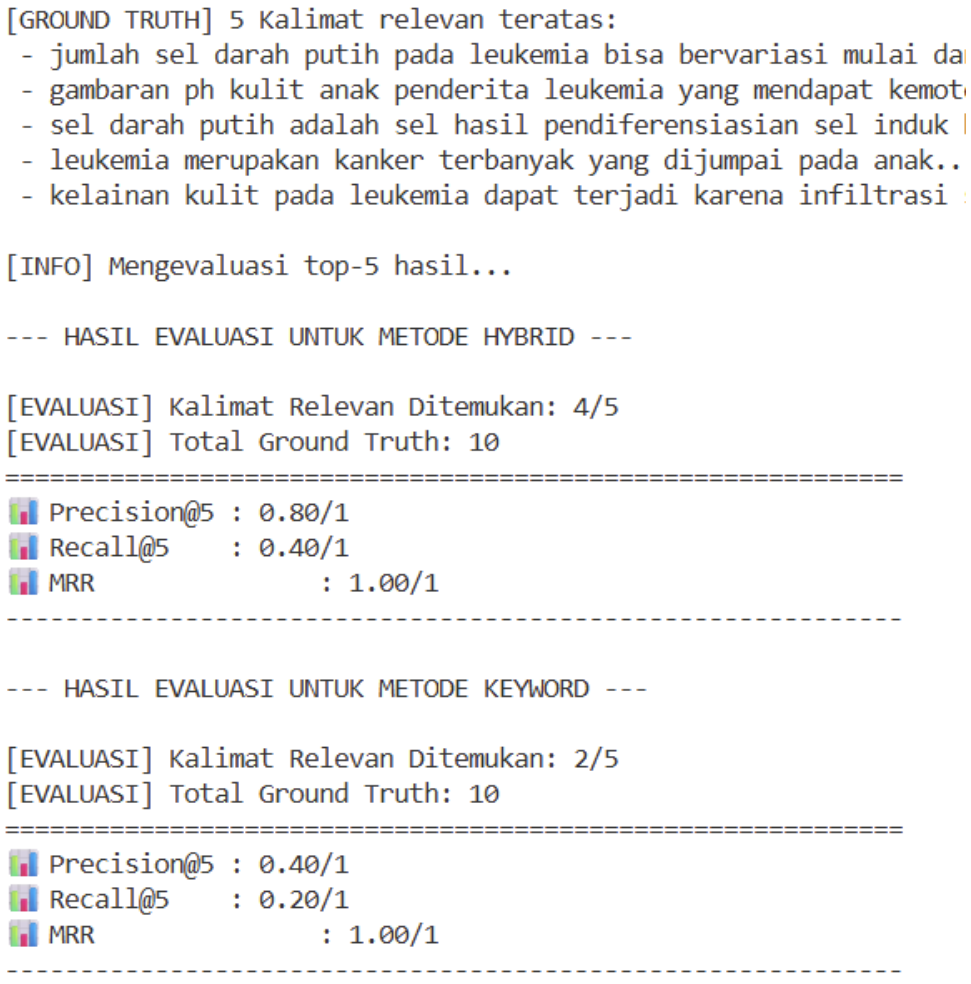


Gambar 8. Hasil Pencarian Oleh Semantik

Sedangkan Gambar 8 menunjukkan hasil pencarian menggunakan metode semantik berbasis SBERT. Pada metode ini, sistem menghasilkan lima dokumen unik dengan skor bervariasi antara 0.53 – 0.66. Hal ini menunjukkan bahwa SBERT mampu mengukur tingkat kedekatan makna antar kalimat, sehingga dokumen yang tidak secara eksplisit mengandung kata *"leukemia"* tetapi memiliki keterkaitan semantik tetap berhasil ditampilkan. Waktu eksekusi lebih lama, yaitu 0.29 detik, karena melibatkan proses embedding vektor dan perhitungan cosine similarity, tetapi masih dalam batas yang wajar untuk sistem pencarian.

Dari kedua hasil tersebut dapat disimpulkan bahwa metode keyword lebih unggul dari sisi kecepatan, tetapi kurang akurat dalam menemukan informasi yang relevan. Sebaliknya, metode SBERT + Ontologi mampu memberikan hasil pencarian yang lebih kaya dan sesuai konteks, sehingga lebih bermanfaat dalam pencarian berbasis makna. Dengan demikian, sistem yang dibangun terbukti lebih baik dibandingkan pencarian tradisional berbasis keyword karena mampu menangkap makna query secara lebih menyeluruh.

2. Hasil Pengujian



Gambar 9. Hasil Pencarian Oleh Semantik

Setelah dilakukan pencarian dengan query *"leukemia"*, sistem diuji menggunakan ground truth berisi 10 kalimat relevan. Hasil evaluasi ditampilkan pada Gambar 9.

Untuk metode Hybrid (SBERT + Ontologi), sistem berhasil menemukan 4 dari 5 hasil pencarian yang relevan terhadap ground truth. Nilai Precision@5 yang diperoleh adalah 0.80, artinya dari 5 hasil teratas, 80% di antaranya relevan. Nilai Recall@5 mencapai 0.40, menunjukkan bahwa sistem mampu menemukan 40% dari seluruh ground truth yang relevan. Nilai MRR (Mean Reciprocal Rank) sebesar 1.00 mengindikasikan bahwa dokumen relevan ditemukan pada peringkat teratas hasil pencarian.

Sedangkan untuk metode Keyword Matching, sistem hanya berhasil menemukan 2 dari 5 hasil pencarian yang relevan. Nilai Precision@5 hanya sebesar 0.40, lebih rendah dibandingkan metode hybrid. Nilai Recall@5 juga lebih kecil, yaitu 0.20, karena sistem keyword terbatas hanya pada kecocokan kata persis. Meskipun demikian, nilai MRR tetap sebesar 1.00 karena setidaknya satu dokumen relevan masih berhasil ditampilkan di posisi teratas.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa metode Hybrid memiliki performa lebih baik dibandingkan metode Keyword Matching, terutama dari segi Precision dan Recall. Hybrid mampu menemukan lebih banyak kalimat relevan dan memberikan hasil yang lebih akurat dalam pencarian berbasis makna, sementara metode keyword cenderung terbatas hanya pada pencarian literal.

3. Analisis Perbandingan

Perbandingan antara metode Keyword Matching dan Semantic Search dilakukan untuk melihat efektivitas masing-masing pendekatan dalam menghasilkan dokumen relevan berdasarkan jenis query yang berbeda. Kedua metode ini memiliki karakteristik yang berbeda. Keyword Matching hanya berfokus pada kecocokan kata secara literal, sedangkan Semantic Search mampu memahami makna dan konteks dari kalimat atau paragraf yang dimasukkan. Oleh karena itu, uji coba dilakukan dengan lima variasi query mulai dari kata tunggal hingga abstrak untuk mengetahui sejauh mana perbedaan performa keduanya. Analisis dilakukan dengan membandingkan jumlah dokumen yang berhasil ditemukan, skor kesesuaian, waktu pencarian, serta hasil evaluasi menggunakan metrik precision, recall, dan MRR. Dengan pendekatan ini, dapat terlihat dengan jelas kelebihan dan kelemahan dari masing-masing metode baik dari sisi kecepatan maupun akurasi pencarian.

Tabel 3. Hasil Pencarian Keyword

No	Jenis Query	Dokumen Ditemukan	Skor Terbesar	Waktu
1	Kata Tunggal	3	1	0.25
2	Dua Kata	5	1	0.03
3	Kalimat	5	0.7273	0.08
4	Paragraf	5	0.326	0.46
5	Abstrak	5	0.248	0.69

Pada Tabel 3 terlihat bahwa metode Keyword Matching mampu menemukan 3–5 dokumen pada setiap jenis query. Skor tertinggi muncul pada query sederhana seperti kata tunggal dan dua kata, yang mencapai nilai 1.0 karena ada kecocokan kata persis dalam dokumen. Namun ketika query semakin kompleks seperti paragraf dan abstrak, skor kesesuaian menurun drastis menjadi 0.326 dan 0.248. Hal ini menunjukkan bahwa metode keyword tidak mampu menangkap makna atau relevansi konteks, melainkan hanya mengandalkan kesamaan kata literal. Dari sisi waktu, pencarian dengan metode ini relatif sangat cepat (0.02–0.69 detik) karena prosesnya sederhana, hanya melakukan pencocokan string.

Tabel 4 menunjukkan hasil pencarian dengan metode Semantic Search. Jumlah dokumen yang ditemukan konsisten yaitu 5 dokumen unik pada semua jenis query. Skor tertinggi berada pada rentang 0.67 – 0.87, lebih stabil dibandingkan metode keyword yang fluktuatif. Bahkan pada query panjang seperti paragraf dan abstrak, skor tetap berada di kisaran 0.6, menunjukkan bahwa semantic search mampu mempertahankan kualitas pencarian. Dari sisi waktu, metode ini membutuhkan waktu lebih lama (0.25 – 0.35 detik) karena ada proses embedding dan perhitungan cosine similarity. Namun, perbedaan waktu tersebut masih tergolong kecil dibandingkan dengan peningkatan kualitas hasil pencarian yang signifikan.

Tabel 4. Hasil Pencarian Semantik

No	Jenis Query	Dokumen Ditemukan	Skor Terbesar	Waktu
1	Kata Tunggal	5	0.692	0.25
2	Dua Kata	5	0.816	0.26
3	Kalimat	5	0.857	0.3
4	Paragraf	5	0.700	0.32
5	Abstrak	5	0.679	0.35

Dari perbandingan hasil pada Tabel 3 (Keyword) dan Tabel 4 (Semantic), terlihat jelas perbedaan kemampuan kedua metode. Keyword lebih unggul hanya pada query sederhana (kata tunggal atau dua kata) karena sifatnya yang langsung mencocokkan kata persis. Namun, performa metode ini menurun drastis pada query kompleks seperti paragraf dan abstrak, sehingga hasilnya tidak stabil. Sebaliknya, Semantic Search tetap mampu

menemukan dokumen relevan dengan skor yang relatif tinggi dan stabil meskipun query semakin panjang. Hal ini menunjukkan bahwa Semantic Search lebih sesuai digunakan pada pencarian dokumen skripsi yang biasanya berbentuk kalimat panjang atau paragraf.

Tabel 5. Hasil Pengujian Keyword

No	Jenis Query	Precission	Recall	MRR
1	Kata Tunggal	0.4	0.2	1
2	Dua Kata	0.2	0.1	0.5
3	Kalimat	0.2	0.1	1
4	Paragraf	0.4	0.2	1
5	Abstrak	0	0	0
Rata - Rata		0.24	0.12	0.70

Pada Tabel 5 terlihat bahwa metode Keyword Matching memiliki nilai Precision rata-rata rendah (0.24) dan Recall sangat rendah (0.12). Hanya pada query sederhana (kata tunggal dan paragraf) Precision sedikit lebih baik (0.4), namun tetap jauh dari ideal. Nilai MRR rata-rata 0.70 menunjukkan bahwa beberapa dokumen relevan muncul di posisi atas, tetapi banyak dokumen relevan lain yang terlewat. Dengan demikian, Keyword Matching cenderung menghasilkan hasil yang sempit dan kurang menyeluruh.

Tabel 6. Hasil Pengujian Semantik

No	Jenis Query	Precission	Recall	MRR
1	Kata Tunggal	0.8	0.4	1
2	Dua Kata	1	0.5	1
3	Kalimat	0.8	1	1
4	Paragraf	1	0.33	1
5	Abstrak	0/8	1	1
Rata - Rata		0.88	0.65	1

Tabel 6 menunjukkan bahwa metode Semantic Search jauh lebih baik dibanding Keyword Matching. Nilai Precision rata-rata 0.88 cukup tinggi dan stabil pada semua jenis query. Recall juga lebih baik dengan rata-rata 0.65, menandakan lebih banyak dokumen relevan berhasil ditemukan. Nilai MRR konsisten 1.0, yang berarti dokumen paling relevan selalu muncul di peringkat pertama. Hal ini membuktikan bahwa Semantic Search lebih efektif secara keseluruhan meskipun membutuhkan waktu pemrosesan sedikit lebih lama.

Dari perbandingan hasil pada Tabel 5 (Keyword) dan Tabel 6 (Semantic), dapat disimpulkan bahwa Keyword Matching hanya unggul sedikit pada kesederhanaan proses, tetapi sangat lemah dalam Recall sehingga banyak dokumen relevan terlewat. Sebaliknya, Semantic Search unggul pada Precision, Recall, dan MRR, sehingga mampu menampilkan dokumen relevan dengan peringkat tinggi pada berbagai jenis query, termasuk query panjang. Oleh karena itu, dalam konteks pencarian dokumen skripsi yang kompleks, Semantic Search lebih layak digunakan dibandingkan Keyword Matching.

Pembahasan

Berdasarkan hasil pengujian yang telah dilakukan, sistem pencarian semantik berbasis Sentence-BERT dan Cosine Similarity terbukti mampu memberikan hasil yang konsisten dan relevan pada berbagai variasi query. Pada pengujian otomatis, sistem menunjukkan nilai rata-rata Precision sebesar 0,8, Recall sebesar 0,92, serta MRR yang stabil bernilai 1,0. Nilai MRR yang selalu 1,0 menandakan bahwa dokumen paling relevan selalu berhasil ditampilkan pada peringkat teratas, sehingga pengguna dapat langsung memperoleh hasil yang paling sesuai tanpa harus menelusuri seluruh daftar pencarian. Walaupun demikian, terlihat bahwa pada query yang lebih panjang seperti paragraf, nilai Recall mengalami penurunan hingga 0,6. Hal ini dapat dijelaskan karena semakin panjang query, semakin besar pula variasi kata

dan makna yang digunakan sehingga sebagian dokumen relevan tidak seluruhnya dapat terdeteksi.

Ketika dibandingkan dengan metode pencarian tradisional berbasis keyword, terlihat perbedaan yang cukup signifikan. Pencarian keyword mampu menghasilkan nilai skor sempurna pada query sederhana seperti kata tunggal dan dua kata, karena kecocokan literal dapat langsung ditemukan. Namun, performa metode ini menurun drastis pada query panjang seperti paragraf dan abstrak, di mana skor kesesuaian dan nilai recall jauh lebih rendah. Sebaliknya, sistem semantik berbasis SBERT mampu menjaga stabilitas hasil dengan skor yang relatif tinggi dan konsisten pada seluruh jenis query, meskipun waktu eksekusinya sedikit lebih lama. Hal ini membuktikan bahwa pencarian berbasis semantik lebih sesuai untuk kebutuhan pencarian dokumen skripsi yang umumnya berbentuk kalimat atau paragraf panjang.

Secara keseluruhan, dapat disimpulkan bahwa sistem yang dibangun telah bekerja dengan baik dalam memahami maksud pencarian dan mengembalikan dokumen yang paling relevan. Pengujian otomatis memberikan gambaran performa sistem yang konsisten secara metrik, sedangkan perbandingan dengan metode keyword menegaskan keunggulan sistem semantik dalam hal pemahaman konteks. Dengan demikian, sistem ini lebih unggul dan relevan untuk diimplementasikan pada skenario pencarian dokumen akademik yang kompleks.

KESIMPULAN

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Sistem pencarian semantik berbasis Sentence-BERT dan Cosine Similarity mampu memberikan hasil pencarian yang relevan terhadap dokumen skripsi dalam format PDF. Hal ini dibuktikan dengan nilai Mean Reciprocal Rank (MRR) yang konsisten bernilai 1 pada setiap jenis query, sehingga dokumen paling relevan selalu berhasil ditemukan di peringkat pertama.
2. Integrasi ontologi dalam sistem terbukti membantu memperluas pemahaman makna query. Dengan adanya sinonim dan istilah yang terkait, sistem dapat menemukan hasil yang relevan meskipun pengguna menggunakan variasi istilah yang berbeda dari dokumen sumber. Hal ini meningkatkan kemampuan sistem untuk menafsirkan maksud query dengan lebih baik.
3. Nilai Precision rata-rata pada pengujian otomatis maupun manual menunjukkan bahwa sebagian besar hasil pencarian pada top-5 dianggap relevan. Namun, precision cenderung menurun pada query dengan konteks panjang (kalimat dan paragraf), menandakan bahwa sistem lebih efektif untuk query sederhana meskipun bantuan ontologi dapat sedikit memperbaiki relevansi.
4. Hasil perbandingan dengan metode pencarian berbasis keyword menunjukkan bahwa sistem semantik lebih unggul. Keyword Matching memang memiliki kecepatan lebih tinggi (0,02–0,69 detik), namun hasilnya terbatas pada kecocokan literal. Sebaliknya, Semantic Search membutuhkan waktu lebih lama (0,25–0,35 detik), tetapi menghasilkan skor yang lebih stabil (0,67–0,87) dan mampu menampilkan dokumen relevan meskipun tidak mengandung kata persis.
5. Secara umum, sistem telah memenuhi tujuan penelitian, yaitu menghasilkan pencarian dokumen skripsi yang relevan dan mampu menempatkan hasil terbaik pada posisi teratas. Namun, performa sistem masih dapat ditingkatkan terutama dalam memperluas cakupan hasil pencarian serta memperkaya struktur ontologi agar sistem semakin adaptif terhadap query yang kompleks.

REFERENSI

Amien, M. (2023). *Perkembangan NLP dalam Bahasa Indonesia: Tinjauan dan Aplikasinya.* Jurnal Teknologi dan Informatika Indonesia, 15(4), 78–92.

Business Analytics. (2023). *Cosine Similarity Explained.* Diakses pada 24 Maret 2025, dari https://businessanalytics.substack.com/p/cosine-similarity-explained

Dataiku. (2023, Mei 9). *Semantic search: An overlooked NLP superpower*. Dataiku Blog. Diakses pada 24 Maret 2025, dari https://blog.dataiku.com/semantic-search-an-overlooked-nlp-superpower.

GeeksforGeeks. (2024). *Natural Language Processing Workflow*. GeeksforGeeks. Diakses pada 24 Maret 2025, dari https://www.geeksforgeeks.org

Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media.

Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). *A Deep Relevance Matching Model for Ad-hoc Retrieval*. Proceedings of the 25th ACM International Conference on Information and Knowledge Management.

Lattner, C., Adya, P., & Kumar, R. (2020). *A Comparative Study of Code Editors for Python Development*. International Journal of Software Engineering, 25(4), 312-329.

Lestari, I., & Pratama, B. (2024). Analisis Penggunaan Transformer-based Model dalam Sistem Pencarian Akademik. Jurnal Informatika dan Komputasi, 8(2), 45–60.

Mitra, B., & Craswell, N. (2018). *An Introduction to Neural Information Retrieval*. Foundations and Trends in Information Retrieval, 13(1), 1-126.

Nisha, K., Wahyuni, T., & Hayat, M. A. M. (2024). Pemeriksaan KTP Menggunakan Optical Character Recognition (OCR) dan Pengenalan Background serta Komponen KTP. *Arus Jurnal Sains dan Teknologi*, 2(2), 490-495

Nugroho, A., et al. (2023). Evaluasi Metode Cosine Similarity dalam Pencarian Dokumen Akademik. Jurnal Riset Teknologi Informasi, 7(3), 150–165.

Nur Oktavia, et al. (2024). *Analisis Semantik dalam Tweet Buzzer Menggunakan Natural Language Processing.* Jurnal Biikma Universitas Indonesia, 12(2), 45–67.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.* Journal of Computing and Applications, 45(3), 12–34.

Suryani, R., & Setiawan, D. (2024). Penerapan Pencarian Semantik dalam Sistem Informasi Akademik Berbasis Web. Jurnal Sistem dan Teknologi Informasi, 10(1), 120–135.

Tan, J., Wong, S., & Lee, C. (2022). *The Impact of Code Editors on Developer Productivity*. Journal of Computing and Software Development, 30(2), 187-202.

Wibawa, C., & Anggraeni, D. T. (2023). COMPARISON OF IMAGE SEGMENTATION METHOD IN IMAGE CHARACTER EXTRACTION PREPROCESSING USING OPTICAL CHARACTER RECOGNITION. *Jurnal Teknik Informatika (JUTIF), 4*(3), 583–589.

Zou, D., Li, S., Huang, Y., & Wu, C. (2021). *A Comparison of Web Frameworks for Building RESTful APIs*. Journal of Web Engineering, 20(3), 456-470.