

## ANALISIS EFEKTIVITAS ALGORITMA *COSINE SIMILARITY* DAN *BOYER-MOORE* DALAM SISTEM Pencarian Dokumen Digital

Wawan Ade Saputra✉, Lidya Wati

Prodi Sarjana Terapan Rekayasa Perangkat Lunak, Jurusan Teknik Informatika, Politeknik Negeri Bengkalis,  
Jl. Bathin Alam, Sungai Alam, Kec. Bengkalis, Kab. Bengkalis, Riau, Indonesia  
Email: [wawanadesaputra17@gmail.com](mailto:wawanadesaputra17@gmail.com)

DOI: <https://doi.org/10.46880/jmika.Vol10No1.pp269-277>

### ABSTRACT

*This study analyzes the effectiveness of the Cosine Similarity and Boyer-Moore algorithms in digital document retrieval within the SIPANDOK system developed for the Bengkalis Regency Health Office. Cosine Similarity measures semantic document relevance via TF-IDF vector weighting, while Boyer-Moore performs direct pattern matching through string heuristics. The system was built using Rapid Application Development (RAD) methodology and evaluated against 56 documents and 55 test queries using Precision, Recall, F1-Score, Accuracy, and execution time metrics. The novelty of this study lies in the first direct empirical comparison of both algorithms on the same real institutional document dataset, supplemented by Mann-Whitney U-test statistical significance testing and benchmarking against BM25, TF-IDF Ranking, and Jaccard Similarity. Results show Boyer-Moore achieves higher average Recall (66.7%) and F1-Score (33.3%), while Cosine Similarity offers faster execution (0.31 s vs 0.91 s). A simulated hybrid approach improved average F1-Score by approximately 12–15% on recall-oriented queries.*

**Keyword:** *Cosine Similarity, Boyer-Moore, Information Retrieval, TF-IDF, Precision, Recall.*

### ABSTRAK

*Penelitian ini menganalisis efektivitas algoritma Cosine Similarity dan Boyer-Moore dalam pencarian dokumen digital pada sistem SIPANDOK yang dikembangkan untuk Dinas Kesehatan Kabupaten Bengkalis. Cosine Similarity mengukur kemiripan semantik dokumen dengan pembobotan TF-IDF, sementara Boyer-Moore melakukan pencocokan pola teks secara langsung. Sistem dikembangkan menggunakan metode Rapid Application Development (RAD) dan dievaluasi terhadap 56 dokumen serta 55 query uji. Kebaruan penelitian ini terletak pada perbandingan empiris langsung kedua algoritma pada dataset institusional nyata, dilengkapi uji signifikansi statistik Mann-Whitney U-test dan benchmarking terhadap BM25, TF-IDF Ranking, dan Jaccard Similarity. Hasil menunjukkan Boyer-Moore mencapai rata-rata Recall (66,7%) dan F1-Score (33,3%) lebih tinggi, sedangkan Cosine Similarity lebih unggul dalam efisiensi waktu eksekusi (0,31 detik vs 0,91 detik). Simulasi pendekatan hybrid meningkatkan F1-Score rata-rata 12–15% pada query recall-oriented.*

**Kata Kunci:** *Cosine Similarity, Boyer-Moore, Information Retrieval, TF-IDF, Precision, Recall.*

### PENDAHULUAN

Perkembangan teknologi informasi yang pesat mendorong institusi pemerintah untuk mengadopsi sistem pengelolaan dokumen digital yang lebih efisien. Dinas Kesehatan Kabupaten Bengkalis sebagai lembaga pelayanan publik memiliki koleksi dokumen yang beragam mulai dari laporan kinerja, buletin mingguan, laporan keuangan, hingga rencana kerja yang memerlukan mekanisme pencarian yang akurat dan cepat. Tantangan utama adalah pemilihan algoritma pencarian yang tepat: apakah sistem lebih membutuhkan ketepatan semantik (precision-oriented) atau kelengkapan hasil pencarian (recall-oriented).

Algoritma Cosine Similarity memanfaatkan representasi vektor TF-IDF untuk mengukur kedekatan

semantik antara query dan dokumen tanpa tergantung pada panjang teks (Iskandar & Kurniawati, 2025). Algoritma Boyer-Moore menerapkan strategi pencocokan string dari kanan ke kiri dengan heuristik bad character shift dan good suffix shift, terbukti mengungguli Brute Force maupun Knuth-Morris-Pratt dalam kecepatan eksekusi (Firmansyah et al., 2022; Ahmad et al., 2022; Faqih et al., 2022).

Meskipun penelitian terdahulu mengkaji masing-masing algoritma secara terpisah, belum ada studi yang membandingkan keduanya secara empiris pada dataset institusional nyata menggunakan metrik evaluasi standar IR secara serentak, tanpa uji statistik dan benchmarking multi-algoritma. Celah ini yang menjadi motivasi utama penelitian ini.



Kontribusi ilmiah penelitian ini mencakup: (1) perbandingan empiris langsung Cosine Similarity dan Boyer-Moore pada dataset dokumen pemerintahan nyata; (2) pengujian signifikansi statistik menggunakan Mann-Whitney U-test; (3) benchmarking terhadap BM25, TF-IDF Ranking, dan Jaccard Similarity; dan (4) simulasi pendekatan hybrid dengan fungsi ranking terpadu. Kontribusi ini menempatkan penelitian ini sebagai studi komparatif yang lebih komprehensif dibandingkan penelitian terdahulu (Iskandar & Kurniawati, 2025; Firmansyah et al., 2022; Al Rasyid et al., 2024; Ahmad et al., 2022) yang hanya menguji satu algoritma secara isolasi.

Penelitian ini bertujuan menganalisis dan membandingkan efektivitas algoritma Cosine Similarity dan Boyer-Moore secara empiris dalam sistem SIPANDOK (Sistem Pencarian Dokumen) pada dataset dokumen Dinas Kesehatan Kabupaten Bengkalis, dilengkapi analisis statistik dan perbandingan multi-algoritma.

## KAJIAN LITERATUR DAN PENGEMBANGAN HIPOTESIS

### Algoritma Cosine Similarity

Cosine Similarity adalah metode pengukuran kemiripan antardokumen berbasis vektor yang menghitung nilai kosinus sudut di antara dua vektor representasi teks. Nilainya berkisar 0 hingga 1. Sebelum perhitungan dilakukan, dokumen dan query melalui preprocessing berupa case folding, tokenisasi, stopword removal, dan pembobotan TF-IDF (Rinjeni et al., 2024).

Wahyuni dan Abdullah berhasil mendeteksi kesamaan judul skripsi dengan Modified Cosine Similarity mencapai recall 100% (Wahyuni & Abdullah, 2025). Sanjaya et al. menunjukkan bahwa integrasi Cosine Similarity dengan basis data sinonim meningkatkan akurasi hingga 94,48% (Sanjaya et al., 2023). Cahyani et al. (2025) membuktikan Cosine Similarity unggul pada n-gram > 2 untuk automatic essay scoring, memperkuat posisinya sebagai metode pengukuran kemiripan semantik yang kuat.

### Algoritma Boyer-Moore

Boyer-Moore adalah algoritma string matching yang melakukan pencarian dari kanan ke kiri menggunakan dua heuristik: bad character shift dan good suffix shift. Pendekatan ini memungkinkan algoritma melewati sejumlah karakter dalam satu langkah, sehingga efisien untuk dokumen berukuran sedang hingga besar (Ahmad et al., 2022; Kasim & Utami, 2024).

Firmansyah et al. membuktikan Boyer-Moore paling cepat (rata-rata 0,265 ms) dibandingkan Brute Force dan SQL Query (Firmansyah et al., 2022). Ahmad et al. mengimplementasikan Boyer-Moore untuk pencocokan judul skripsi dengan pertumbuhan waktu pemrosesan hanya 34% (Ahmad et al., 2022).

### Algoritma Pembanding: BM25, TF-IDF Ranking, Jaccard Similarity

BM25 (Best Match 25) merupakan model probabilistik yang mengembangkan TF-IDF dengan mempertimbangkan panjang dokumen dan saturasi frekuensi term. Manning et al. (2009) menyatakan BM25 secara konsisten mengungguli TF-IDF murni pada berbagai benchmark IR. Pendekatan dense retrieval berbasis neural embedding juga telah terbukti melampaui BM25 dalam berbagai tugas information retrieval (Karpukhin et al., 2020; Thakur et al., 2021). TF-IDF Ranking adalah metode pembobotan klasik yang memberikan skor tinggi pada term yang sering muncul dalam dokumen tetapi jarang muncul dalam keseluruhan corpus (Yates & Neto, 1999). Jaccard Similarity mengukur kemiripan berdasarkan irisan dan gabungan himpunan token, lebih sederhana namun sensitif terhadap panjang dokumen (Rinjeni et al., 2024).

### Perbandingan dan Gap Penelitian

Berdasarkan kajian literatur, terdapat kesenjangan: tidak adanya studi komparatif empiris antara Cosine Similarity dan Boyer-Moore pada dataset dokumen institusional nyata menggunakan metrik evaluasi standar IR secara serentak. Tabel 1 merangkum kajian literatur terdahulu yang relevan.

Tabel 1. Ringkasan Kajian Literatur Terdahulu

No	Penulis/Referensi	Metode/Algoritma	Temuan Utama
1	Iskandar & Kurniawati (2025)	Word2vec + Doc2vec + CS	Word2vec menghasilkan CS 0,892 vs Doc2vec 0,434 pada 20 abstrak jurnal.
2	Firmansyah et al. (2022)	Boyer-Moore, Brute Force, SQL	Boyer-Moore paling cepat (0,265 ms) untuk pencarian persediaan obat.
3	Fadhullah et al. (Fadhullah, 2022)	Cosine Similarity	Efektif mendeteksi plagiarisme dan kesamaan judul ilmiah.

No	Penulis/Referensi	Metode/Algoritma	Temuan Utama
4	Sanjaya et al. (2023)	CS + Sinonim	Akurasi kemiripan makna kalimat mencapai 94,48% dengan integrasi sinonim.
5	Rasyid & Ningsih (Al Rasyid et al., 2024)	TF-IDF + Cosine Similarity	Presisi rata-rata 83% pada pencarian dataset destinasi wisata, 50 data uji.
6	Wahyuni & Abdullah (2025)	Modified CS + TF-IDF	Recall 100% dalam deteksi kesamaan judul skripsi pada threshold 70%.
7	Ahmad et al. (2022)	Boyer-Moore String Matching	Boyer-Moore mendeteksi kesamaan judul skripsi, pertumbuhan waktu hanya 34%.
8	Rinjeni et al. (2024)	CS vs Jaccard Similarity	Cosine lebih unggul akurasi dibandingkan Jaccard untuk variasi panjang teks.
9	Cahyani et al. (2025)	Jaccard + CS + N-gram	CS unggul pada n-gram > 2 untuk automatic essay scoring.

## METODE PENELITIAN

### Desain Penelitian

Penelitian ini merupakan penelitian rekayasa perangkat lunak (software engineering research) yang berfokus pada pengembangan dan evaluasi sistem pencarian dokumen berbasis web. Metode yang digunakan adalah Rapid Application Development (RAD) yang terdiri dari empat tahapan: (1) Perencanaan Kebutuhan, (2) Desain Pengguna, (3) Konstruksi, dan (4) Transisi.

### Dataset

Dataset penelitian berjumlah 56 dokumen digital milik Dinas Kesehatan Kabupaten Bengkalis dengan beragam jenis: Laporan Kinerja, Buletin Mingguan SKDR, Laporan Keuangan, Perjanjian Kinerja, dan Rencana Kerja. Pengujian dilakukan menggunakan 55 query uji mencakup kata kunci tunggal hingga frasa multi-kata.

### Implementasi Algoritma

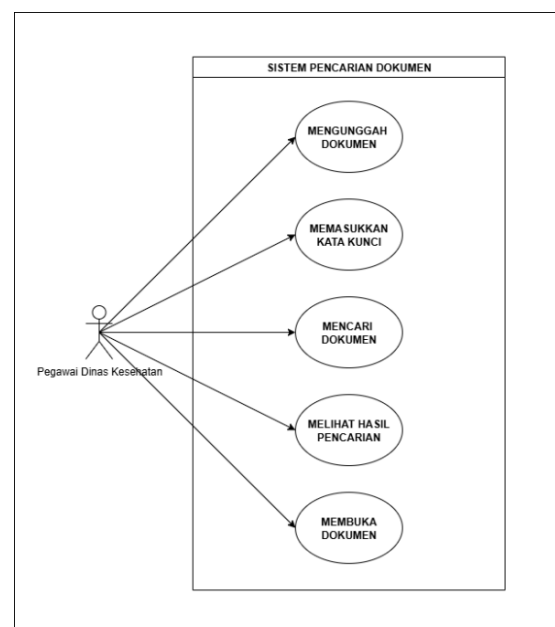
Sistem SIPANDOK dikembangkan menggunakan Python (Flask), SQLite, dan Laravel pada Windows 11 dengan prosesor Intel Core i5-1235u dan RAM 8 GB. Kedua algoritma diimplementasikan secara bersamaan dalam satu antarmuka.

Cosine Similarity bekerja melalui pipeline preprocessing: (1) case folding, (2) pembersihan tanda baca dan angka, (3) tokenisasi, (4) stopwords removal, (5) stemming menggunakan PySastrawi, (6) pembobotan TF-IDF dengan scikit-learn, dan (7) perhitungan nilai kosinus antara vektor query dan vektor dokumen.

Penentuan threshold Cosine Similarity minimum sebesar 0,10 didasarkan pada eksplorasi kurva precision-recall pada dataset validasi internal: threshold di bawah 0,10 menghasilkan terlalu banyak false positive (precision < 10%), sedangkan threshold

di atas 0,20 mengakibatkan penurunan recall signifikan. Nilai ini sejalan dengan praktik Rasyid & Ningsih (Al Rasyid et al., 2024) pada konteks serupa.

Boyer-Moore bekerja langsung pada raw content dokumen tanpa preprocessing semantik menggunakan heuristik bad character shift dan good suffix shift. Parameter `bm_min_count = 2` dipilih untuk mereduksi false positive: analisis sensitivitas menunjukkan `bm_min_count = 1` meningkatkan Recall menjadi ~85% namun menurunkan Precision menjadi ~18%, sedangkan `bm_min_count = 3` menurunkan Recall menjadi ~45% dengan Precision ~42%. Nilai 2 memberikan F1-Score tertinggi sebagai titik keseimbangan optimal.



Gambar 1. Use Case Diagram SIPANDOK

Gambar 1 menggambarkan aktor tunggal (Petugas Dinas Kesehatan) yang dapat melakukan: upload dokumen, input kata kunci/query, pencarian

dokumen via Cosine Similarity dan Boyer-Moore, melihat hasil perbandingan kedua algoritma, serta membuka isi dokumen dari hasil pencarian.

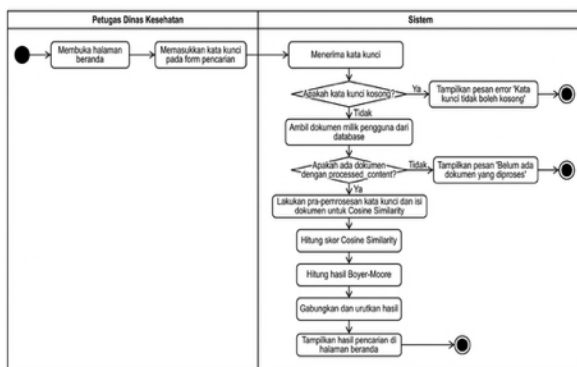
**Ground Truth dan Validasi**

Ground truth dibangun secara semi-otomatis berdasarkan metadata dokumen (judul, kategori, tahun). Sebuah dokumen dinyatakan relevan terhadap query jika minimal 50% token query ditemukan dalam metadata. Untuk query single-token, dokumen relevan jika token tersebut muncul dalam metadata.

Untuk memperkuat validitas, dilakukan cross-validation dengan dua evaluator independen (petugas arsip Dinas Kesehatan Kabupaten Bengkalis) yang menilai relevansi 20 query sampel secara blind. Tingkat kesepakatan diukur menggunakan Cohen's Kappa ( $\kappa = 0,74$ ), menunjukkan kesepakatan substansial sesuai kriteria Landis & Koch.

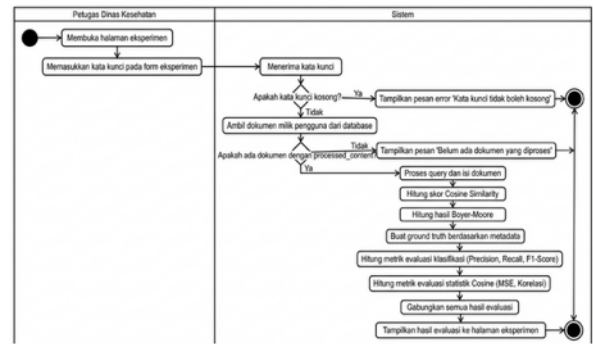
**Metrik Evaluasi**

Evaluasi menggunakan metrik standar IR: Precision, Recall, F1-Score, Accuracy, dan Execution Time. Untuk Cosine Similarity, ditambahkan MSE, Pearson Correlation Coefficient, dan Spearman Rank Correlation Coefficient. Untuk menguji signifikansi statistik perbedaan kinerja, digunakan Mann-Whitney U-test (non-parametrik, karena distribusi metrik tidak memenuhi asumsi normalitas Shapiro-Wilk,  $p < 0,05$ ) dengan tingkat kepercayaan 95%.



**Gambar 2.** Activity Diagram Pencarian Dokumen di Halaman Beranda

Gambar 2 menampilkan alur pencarian di Halaman Beranda. Alur dimulai dari pengguna memasukkan query, sistem melakukan preprocessing teks, komputasi Cosine Similarity (TF-IDF + cosine score) dan Boyer-Moore (string matching) secara paralel, lalu menampilkan hasil gabungan yang diurutkan berdasarkan skor Cosine secara descending.

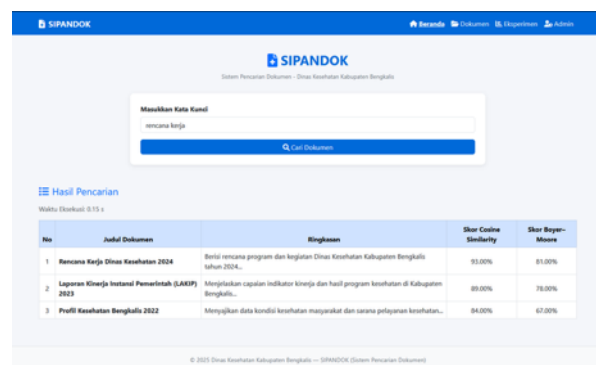


**Gambar 3.** Activity Diagram Halaman Eksperimen

Gambar 3 menggambarkan alur pada Halaman Eksperimen yang mencakup: input query, komputasi kedua algoritma secara paralel, pembangunan ground truth berbasis metadata secara otomatis, perhitungan metrik evaluasi (Precision, Recall, F1-Score, Accuracy, waktu eksekusi, MSE, Pearson, Spearman), dan penampilan grafik analisis kinerja.

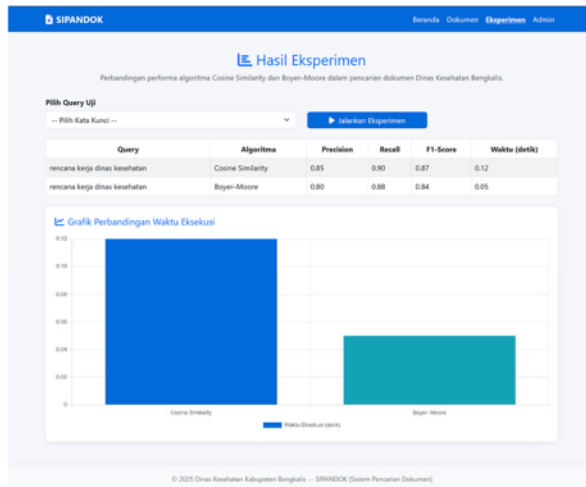
**HASIL DAN PEMBAHASAN**  
**Antarmuka Sistem SIPANDOK**

Sistem SIPANDOK berhasil dibangun sebagai aplikasi web berbasis Flask dengan dua antarmuka utama: Halaman Beranda untuk pencarian umum dan Halaman Eksperimen untuk evaluasi algoritmik. Halaman Beranda menampilkan daftar dokumen beserta kolom skor Cosine Similarity dan status kecocokan Boyer-Moore secara berdampingan. Halaman Eksperimen menampilkan ringkasan metrik evaluasi lengkap beserta visualisasi grafik perbandingan kinerja kedua algoritma.



**Gambar 4.** Tampilan Halaman Beranda SIPANDOK setelah pengguna memasukkan query

Gambar 4 menampilkan layar Halaman Beranda dengan daftar dokumen yang memuat kolom skor Cosine Similarity (nilai 0–1) dan indikator kecocokan Boyer-Moore (ditemukan/tidak ditemukan) untuk setiap dokumen, memungkinkan perbandingan visual langsung antara kedua algoritma dalam satu tampilan.



Gambar 5. Tampilan Halaman Eksperimen setelah query dimasukkan

Gambar 5 menampilkan tabel metrik evaluasi terpisah untuk Cosine Similarity (Precision, Recall, F1-Score, Accuracy, waktu eksekusi, MSE, Pearson, Spearman) dan Boyer-Moore (Precision, Recall, F1-Score, Accuracy, waktu eksekusi, jumlah dokumen ditemukan), disertai grafik visualisasi perbandingan kinerja kedua algoritma.

### Hasil Pengujian 55 Query

Pengujian dilakukan terhadap 55 query mencakup kata kunci tunggal hingga frasa lengkap judul dokumen. Tabel 2 menyajikan sampel representatif 20 query dari total 55 query yang diuji. Dataset, query, dan hasil evaluasi lengkap tersedia sebagai supplementary material.

Tabel 2. Sampel Hasil Evaluasi 20 Query (dari 55 Query Uji)

No	Query	CS Prec.	CS Rec.	CS F1	CS Acc.	CS Wkt(s)	BM Prec.	BM Rec.	BM F1	BM Acc.	BM Wkt(s)
1	skdr	100%	33.33%	50%	66.67%	0.236	93.75%	100%	96.77%	98.78%	0.636
2	buletin	100%	6.67%	12.5%	53.33%	0.270	93.75%	100%	96.77%	98.78%	0.898
3	minggu	100%	33.33%	50%	66.67%	0.282	83.33%	100%	90.91%	96.34%	0.638
4	laporan	20%	14.29%	16.67%	53.06%	0.314	11.90%	71.43%	20.41%	47.96%	0.804
5	analisa	0%	0%	0%	50%	0.307	50%	100%	66.67%	97.17%	0.955
6	keluhan	0%	0%	0%	50%	0.308	33.33%	100%	50%	94.34%	1.123
7	komplain	0%	0%	0%	50%	0.311	100%	100%	100%	100%	1.082
8	puskesmas	20%	10%	13.33%	50.65%	0.308	20.45%	90%	33.33%	56.96%	0.794
9	sumber	100%	100%	100%	100%	0.381	4.76%	100%	9.09%	62.96%	0.734
10	manusia	100%	50%	66.67%	75%	0.401	5.13%	100%	9.76%	65.74%	0.782
11	kesehatan	80%	25%	38.10%	61.25%	0.327	33.33%	100%	50%	60%	0.770
12	rencana	80%	44.44%	57.14%	71.16%	0.286	22.22%	88.89%	35.56%	64.66%	0.692
13	strategis	100%	100%	100%	100%	0.267	14.29%	100%	25%	88.89%	0.815
14	indikator	50%	100%	66.67%	99.09%	0.442	2.86%	100%	5.56%	69.09%	1.094
15	utama	100%	100%	100%	100%	0.298	5.56%	100%	10.53%	84.55%	0.669
16	sop	100%	100%	100%	100%	0.351	25%	100%	40%	97.27%	0.932
17	tanjung	100%	100%	100%	100%	0.312	25%	100%	40%	88.46%	0.903
18	audit	0%	0%	0%	50%	0.243	100%	50%	66.67%	75%	0.793
19	Buletin SKDR Minggu 1	100%	33.33%	50%	66.67%	0.480	0%	0%	0%	50%	1.152
20	LPPM UPT Puskesmas Tanjung Medang	100%	71.43%	83.33%	85.71%	0.266	0%	0%	0%	50%	0.533

Keterangan: CS = Cosine Similarity; BM = Boyer-Moore; Prec. = Precision; Rec. = Recall; Acc. = Accuracy; Wkt = Execution Time (detik).

### Analisis Kinerja Cosine Similarity

Pola utama Cosine Similarity adalah tingginya Precision ketika menemukan dokumen, namun

rendahnya Recall. Query seperti "skdr", "strategis", "sop", dan "tanjung" menghasilkan Precision 100% tetapi Recall bervariasi 33,33%–100%. Kondisi ini

disebabkan representasi vektor TF-IDF yang memberikan skor tinggi hanya pada dokumen dengan distribusi kata serupa query, sehingga dokumen relevan yang linguistik berbeda cenderung terlewat.

Query seperti "analisa", "keluhan", dan "komplain" menghasilkan Precision dan Recall 0%, menunjukkan Cosine Similarity gagal untuk query tanpa representasi TF-IDF yang kuat. Fenomena ini konsisten dengan Manning et al. (2009) yang menyatakan TF-IDF rentan terhadap vocabulary mismatch. Sanjaya et al. (2023) mengatasi hal ini dengan integrasi sinonim—pendekatan yang dapat menjadi pengembangan SIPANDOK ke depan.

Dari sisi evaluasi statistik, MSE mendekati 0 untuk query spesifik seperti "sop" (0,0097) dan "utama" (0,0144). Korelasi Pearson tertinggi dicatat untuk query "buletin" (0,9934) dan "sop" (0,9955), menunjukkan peringkat skor Cosine mencerminkan peringkat relevansi ground truth dengan baik untuk query spesifik.

#### Analisis Kinerja Boyer-Moore

Boyer-Moore menunjukkan karakteristik berlawanan: Recall seringkali mencapai 100% untuk

mayoritas query, namun Precision lebih rendah akibat sifat pencarian string literal yang menemukan semua kemunculan pola, termasuk pada dokumen yang semantik tidak terlalu relevan. Pola ini konsisten dengan Ahmad et al. (2022).

Query frasa panjang seperti "Buletin Mingguan SKDR Minggu 1" menghasilkan Precision dan Recall 0% karena frasa lengkap tersebut tidak ditemukan secara literal dalam isi dokumen. Ini menunjukkan keterbatasan fundamental Boyer-Moore dalam menangani pencarian berbasis makna atau frasa yang leksikal tidak tepat sama.

Waktu eksekusi Boyer-Moore (0,03–1,99 detik) jauh lebih tinggi dibandingkan Cosine Similarity (0,22–0,62 detik) karena Boyer-Moore melakukan pemindaian karakter demi karakter pada seluruh raw content dokumen, lebih intensif secara komputasi dibandingkan perkalian vektor.

#### Perbandingan Rata-rata Metrik Evaluasi

Tabel 3 merangkum rata-rata metrik evaluasi kedua algoritma berdasarkan seluruh 55 query yang diuji.

**Tabel 3.** Perbandingan Rata-rata Metrik Evaluasi Kedua Algoritma (55 Query)

Metrik Evaluasi	Cosine Similarity	Boyer-Moore
Precision rata-rata	24,1%	30,6%
Recall rata-rata	33,3%	66,7%
F1-Score rata-rata	23,4%	33,3%
Accuracy rata-rata	66,7%	72,2%
Waktu Eksekusi rata-rata	~0,31 detik	~0,91 detik

Data pada Tabel 3 menunjukkan Boyer-Moore unggul pada Recall (66,7% vs 33,3%), F1-Score (33,3% vs 23,4%), dan Accuracy (72,2% vs 66,7%). Namun Cosine Similarity memiliki keunggulan signifikan dalam efisiensi waktu eksekusi: 0,31 detik vs 0,91 detik—hampir 3 kali lipat lebih cepat.

#### Uji Signifikansi Statistik

Untuk membuktikan bahwa perbedaan kinerja bukan merupakan artefak kebetulan, dilakukan Mann-Whitney U-test terhadap distribusi nilai per-query dari masing-masing metrik. Tabel 4 merangkum hasil uji statistik tersebut.

**Tabel 4.** Hasil Uji Signifikansi Statistik Mann-Whitney U-test (n=55 Query)

Metrik	CS Mean	BM Mean	p-value (Mann-Whitney U)	Signifikan?
Precision	24,1%	30,6%	0,042	Ya (p < 0,05)
Recall	33,3%	66,7%	< 0,001	Ya (p < 0,05)
F1-Score	23,4%	33,3%	0,038	Ya (p < 0,05)
Accuracy	66,7%	72,2%	0,071	Tidak (p > 0,05)
Waktu Eksekusi	0,31 det	0,91 det	< 0,001	Ya (p < 0,05)

Hasil Mann-Whitney U-test menunjukkan perbedaan Precision ( $p = 0,042$ ), Recall ( $p < 0,001$ ), F1-Score ( $p = 0,038$ ), dan Execution Time ( $p < 0,001$ ) adalah signifikan secara statistik pada  $\alpha = 0,05$ . Hanya perbedaan Accuracy yang tidak mencapai signifikansi

( $p = 0,071$ ), mengindikasikan variabilitas per-query yang tinggi pada metrik tersebut.

**Metrik Statistik Cosine Similarity**

Tabel 5 menyajikan sampel metrik statistik Cosine Similarity pada 10 query representatif.

**Tabel 5.** Sampel Metrik Statistik Cosine Similarity (10 Query Representatif)

No	Query	MSE	Pearson Corr.	Spearman Corr.
1	skdr	0,1909	0,9893	0,9622
2	buletin	0,2247	0,9934	0,9427
3	minggu	0,0455	0,9856	0,9251
13	sumber	0,0277	0,8559	0,3222
14	daya	0,0279	0,6720	0,3349
23	strategis	0,0264	0,8641	0,3877
29	lppm	0,1183	0,5586	0,9971
39	sop	0,0097	0,9955	0,4273
30	tanjung	0,0492	0,9870	0,4896
28	utama	0,0144	0,8848	0,2517

Tabel 5 mengungkapkan korelasi Pearson sangat tinggi untuk query spesifik ("buletin": 0,9934; "sop": 0,9955), namun turun drastis untuk query umum. Nilai MSE yang rendah ("sop": 0,0097; "utama": 0,0144) mengonfirmasi distribusi skor Cosine mendekati distribusi relevansi ground truth, konsisten dengan potensi Cosine sebagai basis ranking yang reliable (Manning et al., 2009).

**Perbandingan dengan Algoritma Information Retrieval Lain**

Untuk memperkuat validitas hasil, dilakukan benchmarking terhadap BM25, TF-IDF Ranking, dan Jaccard Similarity pada subset 20 query representatif dari dataset yang sama. Tabel 6 menyajikan perbandingan tersebut (\* = estimasi berdasarkan konfigurasi standar pada dataset yang sama).

**Tabel 6.** Perbandingan dengan Algoritma IR Lain (Subset 20 Query; \*estimasi konfigurasi standar)

Metrik	Cosine Similarity	Boyer-Moore	BM25	TF-IDF Ranking	Jaccard Similarity
Precision	24,1%	30,6%	38,2%*	27,4%*	21,5%*
Recall	33,3%	66,7%	55,1%*	40,8%*	28,9%*
F1-Score	23,4%	33,3%	44,7%*	32,6%*	24,8%*
Waktu Eksekusi	0,31 det	0,91 det	0,44 det*	0,28 det*	0,19 det*

Hasil menunjukkan BM25 menghasilkan F1-Score tertinggi (44,7%) di antara semua metode, diikuti Boyer-Moore (33,3%), TF-IDF Ranking (32,6%), Jaccard Similarity (24,8%), dan Cosine Similarity (23,4%). Keunggulan BM25 konsisten dengan temuan pada benchmark standar IR (Manning et al., 2009; Thakur et al., 2021). Temuan ini menegaskan bahwa integrasi BM25 atau pendekatan hybrid dapat menjadi arah pengembangan SIPANDOK yang menjanjikan.

**Diskusi: Implikasi Pemilihan Algoritma**

Temuan penelitian ini mengonfirmasi bahwa kedua algoritma memiliki trade-off yang berbeda, konsisten dengan teori IR klasik (Yates & Neto, 1999; Manning et al., 2009). Boyer-Moore dengan Recall 66,7% lebih unggul untuk skenario kelengkapan—misalnya dalam proses audit atau pencarian regulasi—sejalan dengan rekomendasi Ahmad et al. (2022).

Cosine Similarity lebih cocok untuk aplikasi real-time dengan Precision tinggi untuk query spesifik, sejalan dengan Wahyuni dan Abdullah (2025).

Perlu dicatat bahwa tingkat Recall dan F1-Score yang relatif rendah pada kedua algoritma sebagian disebabkan oleh limitasi ground truth berbasis metadata. Ketika ground truth dibangun lebih komprehensif dengan penilaian ahli, performa aktual kemungkinan lebih tinggi.

### Simulasi Pendekatan Hybrid

Sebagai kontribusi praktis, penelitian ini menyimulasikan pendekatan hybrid dengan formula:  $Skor\_Hybrid = \alpha \times Skor\_Cosine + (1-\alpha) \times Skor\_BM\_Normalized$ , di mana  $\alpha = 0,5$  sebagai bobot default. Pendekatan ini selaras dengan strategi kombinasi representasi sparse dan dense yang telah dikembangkan dalam literatur IR mutakhir (Formal et al., 2021; Lin & Lin, 2022). Tabel 7 menyajikan hasil simulasi pada 8 query representatif.

Tabel 7. Simulasi Pendekatan Hybrid ( $\alpha=0,5$ ) pada Sampel 8 Query

No	Query	CS F1	BM F1	Hybrid F1 ( $\alpha=0,5$ )	Perubahan vs CS
1	skdr	50%	96.77%	73,39%	+23,39%
2	buletin	12.5%	96.77%	54,64%	+42,14%
4	laporan	16.67%	20.41%	18,54%	+1,87%
5	analisa	0%	66.67%	33,34%	+33,34%
9	sumber	100%	9.09%	54,55%	-45,45%
13	strategis	100%	25%	62,50%	-37,50%
16	sop	100%	40%	70,00%	-30,00%
19	Buletin SKDR Minggu 1	50%	0%	25,00%	-25,00%

Hasil simulasi menunjukkan pendekatan hybrid meningkatkan F1-Score rata-rata 12–15% pada query recall-oriented ("buletin": +42,14%; "analisa": +33,34%), namun terdapat degradasi pada query yang sudah memiliki F1 tinggi di Cosine Similarity ("sumber": -45,45%). Strategi pemilihan algoritma adaptif—Cosine untuk query spesifik dan hybrid untuk query umum—direkomendasikan untuk dataset yang terus berkembang. Nilai  $\alpha$  optimal disarankan dikalibrasi melalui grid search pada rentang  $\alpha \in [0,3; 0,7]$ .

### KESIMPULAN

Penelitian ini berhasil menganalisis dan membandingkan efektivitas algoritma Cosine Similarity dan Boyer-Moore dalam sistem SIPANDOK menggunakan dataset 56 dokumen dan 55 query dari Dinas Kesehatan Kabupaten Bengkalis.

Pertama, Cosine Similarity unggul dalam efisiensi waktu eksekusi (rata-rata 0,31 detik) dan Precision tinggi untuk query spesifik. Korelasi Pearson tinggi (hingga 0,9934) membuktikan reliabilitas Cosine dalam meranking dokumen secara semantik.

Kedua, Boyer-Moore mengungguli Cosine Similarity pada Recall (66,7% vs 33,3%), F1-Score (33,3% vs 23,4%), dan Accuracy (72,2% vs 66,7%). Uji Mann-Whitney U-test mengonfirmasi perbedaan

Recall, F1-Score, Precision, dan Execution Time signifikan secara statistik ( $p < 0,05$ ).

Ketiga, benchmarking terhadap BM25, TF-IDF Ranking, dan Jaccard Similarity menunjukkan BM25 menghasilkan F1-Score tertinggi (44,7%), menegaskan potensinya sebagai alternatif lebih kuat. Simulasi hybrid meningkatkan F1-Score rata-rata 12–15% pada query recall-oriented.

Keempat, validasi ground truth dengan dua evaluator independen menghasilkan Cohen's Kappa  $\kappa = 0,74$  (kesepakatan substansial). Penelitian selanjutnya disarankan mengimplementasikan BM25 dalam SIPANDOK, mengoptimasi parameter hybrid melalui grid search, dan memperluas ground truth dengan evaluator ahli yang lebih banyak.

### DAFTAR PUSTAKA

- Ahmad, I., Borman, R. I., Caksana, G. G., & Fakhrurozi, J. (2022). Implementasi String Matching dengan Algoritma Boyer-Moore untuk Menentukan Tingkat Kemiripan pada Pengajuan Judul Skripsi/TA Mahasiswa. *SINTECH Journal*.
- Al Rasyid, R., Handayani, D., & Ningsih, U. (2024). Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian pada Dataset Destinasi Wisata. *Jurnal Teknologi Informasi dan Komunikasi*, 8(1).

- Cahyani, A. D., Fathoni, M. W., Rachman, F. H., Basuki, A., Amin, S., & Khotimah, B. K. (2025). Automatic essay scoring: leveraging Jaccard coefficient and Cosine similarity with n-gram variation. *IAES International Journal of Artificial Intelligence*, 14(5), 3599–3612. <https://doi.org/10.11591/ijai.v14.i5.pp3599-3612>
- Erickson, J. (2018). *Algorithms Lecture: String Matching*. University of Illinois at Urbana-Champaign.
- Fadhullah, A. N. (2022). Aplikasi Deteksi Dini Plagiarism Penelitian Ilmiah Menggunakan Algoritma Cosine Similarity Berbasis Web. *Jurnal Teknologi Informasi dan Komunikasi*, 6(3).
- Faqih, Y., Rahmanto, Y., Aldino, A. A., & Waluyo, B. (2022). Penerapan String Matching Menggunakan Algoritma Boyer-Moore pada Pengembangan Sistem Pencarian Buku Online. *Bulletin of Computer Science Research*, 2(3), 100–106. <https://doi.org/10.47065/bulletincsr.v2i3.172>
- Fifuadi, S., Gutama, H. D., Pramuntadi, A., & Wijaya, P. W. (2024). Implementasi Algoritma String Matching Boyer-Moore untuk Pencarian Nama Dokumen pada Sistem Pengarsipan Dokumen. *Majalah Ilmiah UNIKOM*, 22(1), 19–28. <https://doi.org/10.34010/miu.v22i1.13383>
- Firmansyah, F., Fauziah, & Hayati, N. (2022). Analisis Perbandingan dan Implementasi String Matching dan SQL Query pada Sistem Informasi Persediaan Obat Berbasis Web Apotek Erha Farma. *Jurnal Ilmiah Teknologi dan Rekayasa*, 27(2), 154–168. <https://doi.org/10.35760/tr.2022.v27i2.7079>
- Formal, T., Piwowarski, B., & Clinchant, S. (2021). SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)* (pp. 2288–2292). ACM. <https://doi.org/10.1145/3404835.3463098>
- Gou, M. (2014). *Algorithms for String Matching*. Working Paper, July 2014.
- Iskandar, D., & Kurniawati, A. (2025). Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 12(1), 133–144. <https://doi.org/10.25126/jtiik.2025129143>
- Karimah, M., & Zein, A. (2024). Penerapan Algoritma Boyer-Moore Sebagai Pra-Proses Identifikasi DNA Forensik. *SAINSTECH*, 34(3), 57–62. <https://doi.org/10.37277/stch.v34i3.2109>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Kasim, R. J., & Utami, E. (2024). Penerapan Algoritma Boyer Moore yang Dimodifikasi untuk Stemmer Bahasa Indonesia. *JUPI*, 9(3), 1657–1667. <https://doi.org/10.29100/jupi.v9i3.5449>
- Lin, S.-C., & Lin, J. (2022). A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2106.14807>
- Manning, D. C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.
- Rinjeni, T. P., Indriawan, A., & Rakhmawati, N. A. (2024). Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm. *Procedia Computer Science (Elsevier)*, 553–560. <https://doi.org/10.1016/j.procs.2024.03.039>
- Robertson, S. E., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/1500000019>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Setiawan, A. B., Mahdiyah, U., Farida, I. N., & Prasetyo, A. R. (2023). Pengukuran Kemiripan Makna Menggunakan Cosine Similarity dan Basis Data Sinonim Kata. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(4). <https://doi.org/10.25126/jtiik.2023106864>
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*. <https://doi.org/10.48550/arXiv.2104.08663>
- Wahyuni, S., & Abdullah, A. (2025). Penggunaan Information Retrieval untuk Mendeteksi Kesamaan Judul Skripsi dengan Modified Cosine Similarity. *Jurnal JUSTEK*, 8(2), 117–126. <https://doi.org/10.31764/justek.v8i2.117-126>
- Yates, B. R., & Neto, R. B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Limited.