

PERBANDINGAN KINERJA METODE TF-IDF DAN BM25 PADA PENCARIAN ARTIKEL BERITA BAHASA INDONESIA

Muhammad Al Adib¹, Yiska Dayanti Zagoto², Bualazatulo Laia³, Roslina⁴

^{1,2,3}Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Potensi Utama, Medan,
Indonesia

⁴Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Politeknik Negeri Medan, Medan,
Indonesia

e-mail: ¹dibsowden@gmail.com, ²zagotoyiska@gmail.com, ³buallalaia57@gmail.com,
⁴roslina@polmed.ac.id

ABSTRAK

Perkembangan media daring meningkatkan jumlah artikel berita berbahasa Indonesia dan menimbulkan kebutuhan akan sistem pencarian yang efektif serta efisien. Penelitian ini membandingkan kinerja *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Best Matching 25* (BM25) pada *Indonesia News Dataset* (2025). Dari 80.472 artikel, sebanyak 80.463 dokumen digunakan setelah penghapusan sembilan artikel dengan isi kosong. Evaluasi dilakukan menggunakan 10 kueri multikategori. *Ground truth* dibentuk melalui *pooling* 20 hasil teratas dari masing-masing metode, menghasilkan 304 pasangan kueri-dokumen yang dinilai secara manual oleh satu penilai dengan skala relevansi 0–2. Efektivitas diukur menggunakan *Precision@k*, *Recall@k*, *Mean Average Precision* (MAP), *Mean Reciprocal Rank* (MRR), dan *normalized Discounted Cumulative Gain* (nDCG), sedangkan waktu pencarian diuji sebanyak 30 kali setelah tiga kali *warm-up*. Pada Top-10, BM25 memperoleh *Precision@10* sebesar 0,79, *Recall@10* sebesar 0,351630, *MAP@10* sebesar 0,728635, dan *nDCG@10* sebesar 0,630976, lebih tinggi daripada nilai TF-IDF pada metrik yang sama. TF-IDF memperoleh MRR sebesar 0,90, sedikit lebih tinggi daripada BM25 sebesar 0,85. Pada implementasi yang digunakan, rata-rata waktu pencarian BM25 adalah 0,010191 detik, sedangkan TF-IDF 0,066014 detik. Secara deskriptif, BM25 cenderung lebih baik pada sebagian besar metrik Top-10 dan waktu pencarian, tetapi hasil dipengaruhi karakteristik kueri dan konfigurasi eksperimen.

Kata kunci: BM25; pencarian berita berbahasa Indonesia; pemeringkatan dokumen; sistem temu kembali informasi; TF-IDF.

ABSTRACT

The growth of online media has increased the number of Indonesian-language news articles and created a need for effective and efficient retrieval systems. This study compares Term Frequency-Inverse Document Frequency (TF-IDF) and Best Matching 25 (BM25) using the Indonesia News Dataset (2025). Of 80,472 articles, 80,463 documents were retained after removing nine articles with empty content. The evaluation used 10 multi-category queries. Ground truth was constructed by pooling the top 20 results from each method, producing 304 query-document pairs manually assessed by one evaluator using relevance labels from 0 to 2. Retrieval effectiveness was measured using Precision@k, Recall@k, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG). Retrieval time was measured over 30 repetitions after three warm-up runs. At Top-10, BM25 achieved a Precision@10 of 0.79, Recall@10 of 0.351630, MAP@10 of 0.728635, and nDCG@10 of 0.630976, which were higher than the corresponding TF-IDF scores. TF-IDF obtained a slightly higher MRR of 0.90 than

BM25 at 0.85. In the evaluated implementation, BM25 required 0.010191 seconds on average, whereas TF-IDF required 0.066014 seconds. Descriptively, BM25 performed better on most Top-10 metrics and retrieval time, but the results depended on query characteristics and experimental configuration.

Keywords: *BM25; document ranking; Indonesian news retrieval; information retrieval; TF-IDF.*

1. PENDAHULUAN

Perkembangan media daring telah meningkatkan jumlah artikel berita berbahasa Indonesia yang tersedia dalam bentuk digital. Kondisi tersebut mempermudah masyarakat memperoleh informasi, tetapi sekaligus menimbulkan kesulitan ketika pengguna harus menemukan artikel yang sesuai dari koleksi dokumen dalam jumlah besar. Penelitian mengenai pengolahan berita Indonesia menunjukkan bahwa ribuan artikel dari berbagai portal berita dapat dikumpulkan dan diproses untuk memperoleh informasi tertentu [1]. Pada bidang lain, sistem pencarian juga telah diterapkan pada dokumen peraturan pemerintah berbahasa Indonesia [2] dan dokumen rekam medis [3]. Hal tersebut menunjukkan bahwa sistem *information retrieval* diperlukan untuk mencocokkan kueri pengguna dengan isi dokumen, menghitung tingkat relevansi, dan menempatkan dokumen yang paling sesuai pada urutan teratas. Artikel berita mempunyai panjang dokumen, topik, gaya penulisan, dan frekuensi penggunaan istilah yang berbeda-beda. Di dalam artikel berita terdapat istilah penting, seperti nama tokoh, lokasi, lembaga, peristiwa, komoditas, dan kebijakan. Perbedaan karakteristik tersebut menyebabkan setiap istilah tidak dapat dianggap mempunyai tingkat kepentingan yang sama. Salah satu metode pembobotan istilah yang banyak digunakan adalah *Term Frequency-Inverse Document Frequency* atau TF-IDF. Metode ini memberikan bobot berdasarkan frekuensi kemunculan istilah dalam suatu dokumen dan tingkat kelangkaan istilah tersebut pada keseluruhan koleksi dokumen [4], [5]. Dengan TF-IDF, dokumen dan kueri direpresentasikan dalam bentuk vektor, kemudian tingkat kemiripannya dapat dihitung menggunakan *cosine similarity*.

Penerapan TF-IDF dan *cosine similarity* telah dilakukan pada berbagai jenis dokumen. Heryawan et al. menggunakan TF-IDF dan *Vector Space Model* untuk membangun sistem pencarian dokumen rekam medis [3]. Al Rasyid dan Ningsih menerapkan TF-IDF dan *cosine similarity* untuk pencarian informasi pada dataset destinasi wisata [6]. Sementara itu, Tanuwijaya et al. menggunakan kedua metode tersebut untuk melakukan penyaringan dokumen berita berbahasa Indonesia [7]. Beberapa penelitian tersebut menunjukkan bahwa TF-IDF relatif sederhana, mudah diterapkan, dan dapat digunakan untuk mengukur kesamaan leksikal antara kueri dan dokumen. Namun, hasil pemeringkatan TF-IDF masih dipengaruhi oleh frekuensi istilah dan kemiripan vektor tanpa mekanisme khusus untuk mengendalikan kejenuhan frekuensi istilah dan variasi panjang dokumen.

Metode lain yang banyak digunakan dalam pemeringkatan dokumen adalah *Best Matching 25* atau BM25. BM25 merupakan metode pemeringkatan probabilistik yang memanfaatkan frekuensi istilah dan *inverse document frequency*, sekaligus mempertimbangkan kejenuhan frekuensi istilah serta normalisasi panjang dokumen [8]. Karakteristik tersebut membuat BM25 relevan diterapkan pada artikel berita karena setiap artikel dapat mempunyai panjang dan distribusi istilah yang berbeda. Faisal et al. menggunakan BM25 sebagai salah satu model *retriever* dalam pencarian komponen

peraturan pemerintah Indonesia [2]. Satria et al. menerapkan pembobotan BM25 untuk mengelompokkan hasil pencarian dokumen skripsi berbahasa Indonesia [9], sedangkan Kheng et al. menerapkan BM25 untuk menghasilkan pemeringkatan informasi lowongan pekerjaan berdasarkan kueri pengguna [10].

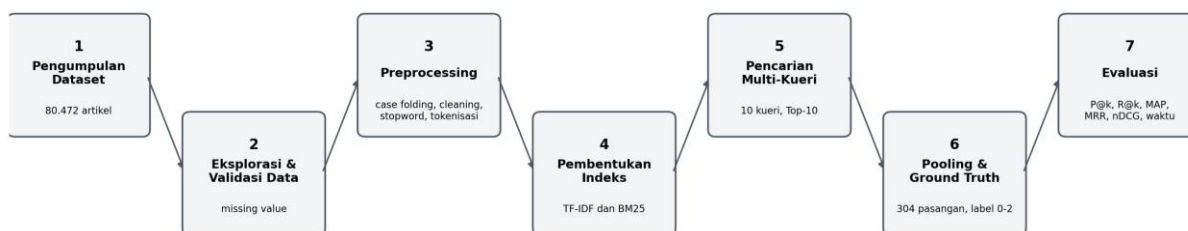
Penelitian terdahulu telah menerapkan TF-IDF dan BM25 pada berbagai jenis dokumen. Namun, pada penelitian yang ditinjau, TF-IDF lebih banyak diterapkan pada pencarian menggunakan *cosine similarity*, sedangkan BM25 diterapkan pada pencarian atau pengelompokan dokumen secara terpisah [2], [3], [6]–[9]. Sebagian penelitian juga membandingkan pendekatan leksikal dengan metode pembelajaran mesin atau model berbasis BERT [1], [2], bukan membandingkan TF-IDF dan BM25 secara langsung pada koleksi artikel berita berbahasa Indonesia. Oleh karena itu, perbandingan kedua metode masih perlu dilakukan untuk mengetahui metode yang memberikan keseimbangan lebih baik antara efektivitas hasil pencarian dan efisiensi waktu pemrosesan. Evaluasi yang hanya menggunakan satu kueri, sejumlah kecil hasil pencarian teratas, dan satu kali pengukuran waktu belum cukup untuk menggambarkan kemampuan metode pada variasi kebutuhan informasi pengguna. Oleh karena itu, evaluasi sistem pencarian perlu melibatkan sejumlah kueri dan lebih dari satu ukuran evaluasi karena setiap metrik menggambarkan aspek kualitas pemeringkatan yang berbeda [11], [12]. Selain jumlah dokumen relevan yang berhasil ditemukan, posisi dokumen relevan pada daftar hasil pencarian juga perlu diperhatikan.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan membandingkan efektivitas dan efisiensi metode TF-IDF dan BM25 dalam pencarian artikel berita berbahasa Indonesia. Pengujian dilakukan menggunakan 10 kueri yang mewakili beberapa kategori berita. Efektivitas kedua metode dievaluasi menggunakan Precision@5, Precision@10, Recall@5, Recall@10, MAP@5, MAP@10, MRR, nDCG@5, nDCG@10, dan tingkat kesamaan hasil Top-10. Sementara itu, aspek efisiensi dievaluasi melalui pengukuran waktu pencarian yang dilakukan secara berulang.

Kontribusi penelitian ini adalah menyediakan perbandingan langsung antara TF-IDF dan BM25 pada dataset berita berbahasa Indonesia melalui evaluasi multi-kueri, penyusunan *ground truth* berbasis *pooling*, penggunaan beberapa metrik evaluasi pemeringkatan, serta pengukuran waktu pencarian secara berulang. Hasil penelitian diharapkan dapat memberikan dasar pemilihan metode pencarian dokumen yang efektif, efisien, mudah direplikasi, dan sesuai untuk lingkungan komputasi dengan sumber daya relatif terbatas.

2. METODE PENELITIAN

Penelitian menggunakan pendekatan eksperimen komparatif. Kedua metode menerima dokumen, preprocessing, vocabulary, dan kumpulan kueri yang sama. Perbedaan hanya terletak pada fungsi pembobotan dan pemeringkatan. Tahapan penelitian terdiri atas pengumpulan data, eksplorasi, preprocessing, pembentukan indeks, pencarian multi-kueri, pooling dan penilaian relevansi, evaluasi efektivitas, pengujian waktu, serta analisis hasil.



Gambar 1. Alur penelitian

2.1 Dataset dan Persiapan Data

Dataset diperoleh dari platform Kaggle dengan nama Indonesia News Dataset (2025). Dataset awal terdiri atas 80.472 artikel dan 10 atribut. Atribut Content digunakan sebagai dokumen pencarian, sedangkan Judul digunakan untuk menampilkan identitas artikel. Atribut source dan Waktu digunakan sebagai informasi tambahan pada candidate pool. Pemeriksaan missing value menemukan sembilan data kosong pada Content. Setelah data kosong dan teks tanpa isi dihapus, korpus akhir terdiri atas 80.463 dokumen.

Tabel 1. Atribut utama dataset

Atribut	Deskripsi	Penggunaan
Judul	Judul artikel	Informasi hasil pencarian
Waktu	Waktu publikasi	Metadata tambahan
Link	Tautan artikel	Identifikasi sumber
Content	Isi artikel	Dokumen utama pencarian
tag1-tag5	Kategori/tag	Tidak digunakan dalam model
source	Media sumber	Metadata tambahan

2.2 Lingkungan Pengujian

Seluruh eksperimen dijalankan pada Google Colab menggunakan CPU. Konfigurasi lingkungan dan library dicatat agar eksperimen dapat direplikasi. Nilai random seed ditetapkan sebesar 42. Waktu preprocessing dan pembentukan indeks dilaporkan terpisah dari waktu pencarian.

Tabel 2. Lingkungan komputasi

Komponen	Spesifikasi
Lingkungan komputasi	Google Colab
Python	3.12.13
Platform	Linux x86_64
CPU	1 physical core; 2 logical CPU

Komponen	Spesifikasi
Lingkungan komputasi	Google Colab
RAM	12,67 GB
NumPy	2.0.2
pandas	2.2.2
SciPy	1.16.3
scikit-learn	1.6.1
Random seed	42

2.3 Preprocessing Teks

Preprocessing diterapkan secara identik pada dokumen dan kueri agar bentuk token konsisten. Tahapannya meliputi case folding, penghapusan URL, alamat surel, tag HTML, tanda baca, angka, dan spasi berlebih; penghapusan stopwords Bahasa Indonesia; serta tokenisasi berdasarkan spasi. Stopword diperoleh dari Sastrawi. Kata negasi seperti tidak, bukan, belum, jangan, dan tanpa dipertahankan karena dapat mengubah makna. Token dengan panjang kurang dari dua karakter dibuang. Stemming tidak digunakan agar proses lebih ringan dan bentuk istilah berita tetap dipertahankan. Pemilihan tahapan preprocessing perlu disesuaikan dengan karakteristik data karena stopwords removal dan stemming dapat memengaruhi hasil pengukuran kemiripan dokumen [13].

Rata-rata panjang dokumen berkurang dari 330,59 kata menjadi 247,09 token setelah preprocessing. Proses preprocessing seluruh korpus membutuhkan 58,215 detik. Setelah penerapan $\min_df=2$ dan $\max_df=0,95$, vocabulary yang terbentuk terdiri atas 90.281 fitur unigram. Pembentukan matriks frekuensi dan indeks kedua metode membutuhkan 14,993 detik.

2.4 Kumpulan Kueri Pengujian

Pengujian menggunakan 10 kueri yang mewakili beberapa kategori berita. Kueri dimasukkan sekaligus melalui satu blok input dan diproses menggunakan pipeline yang sama dengan dokumen. Pemilihan kueri multikategori bertujuan mengurangi ketergantungan hasil pada satu topik tertentu.

Tabel 3. Daftar kueri pengujian

ID	Kategori	Kueri
Q01	Ekonomi	harga cabai
Q02	Ekonomi	kenaikan harga beras
Q03	Politik	pemilihan kepala daerah
Q04	Olahraga	hasil pertandingan sepak bola
Q05	Teknologi	perkembangan kecerdasan buatan

ID	Kategori	Kueri
Q06	Bencana	gempa bumi Indonesia
Q07	Kesehatan	kasus demam berdarah
Q08	Pendidikan	kurikulum pendidikan nasional
Q09	Hukum	kasus korupsi pejabat
Q10	Lingkungan	banjir dan perubahan iklim

2.5 Pembobotan dan Pemeringkatan TF-IDF

TF-IDF merepresentasikan setiap dokumen sebagai vektor bobot istilah. Term Frequency menyatakan frekuensi istilah t pada dokumen d , sedangkan Inverse Document Frequency menurunkan kontribusi istilah yang muncul pada banyak dokumen.

$$TF(t, d) = f(t, d) \quad (1)$$

$$IDF(t) = \log \left(\frac{N + 1}{df(t) + 1} \right) + 1 \quad (2)$$

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (3)$$

Dokumen dan kueri dinormalisasi dengan norma L2. Skor kemiripan dihitung menggunakan cosine similarity. Konfigurasi yang digunakan adalah unigram, min_df=2, max_df=0,95, use_idf=True, smooth_idf=True, sublinear_tf=False, dan norm=l2. Penerapan VSM dan TF-IDF pada sistem pencarian dokumen berbahasa Indonesia juga telah digunakan untuk menentukan dokumen yang paling relevan terhadap kueri [14].

$$\cos(q, d) = \frac{q \cdot d}{\|q\| \|d\|} \quad (4)$$

2.6 Pembobotan dan Pemeringkatan BM25

BM25 menghitung relevansi dokumen terhadap kueri menggunakan frekuensi istilah, IDF, kejenuhan frekuensi, dan normalisasi panjang dokumen. Parameter k_1 mengendalikan kejenuhan term frequency, sedangkan b mengatur pengaruh panjang dokumen. Penelitian ini menggunakan $k_1=1,5$ dan $b=0,75$.

$$IDF(t) = \log \left(1 + \frac{N - df(t) + 0,5}{df(t) + 0,5} \right) \quad (5)$$

$$BM25(d, q) = \sum_{t \in q} IDF(t) \times \frac{f(t, d)(k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \frac{|d|}{avgdl} \right)} \quad (6)$$

Vocabulary dan matriks frekuensi dasar dibuat menggunakan konfigurasi yang sama dengan TF-IDF. Hal ini menjaga keadilan perbandingan karena kedua metode bekerja pada token dan ruang fitur yang sama.

2.7 Pembentukan Candidate Pool dan Ground Truth

Ground truth dibentuk menggunakan metode pooling. Untuk setiap kueri, 20 dokumen teratas dari TF-IDF dan 20 dokumen teratas dari BM25 digabungkan, sedangkan dokumen duplikat hanya dinilai satu kali. Proses tersebut menghasilkan 304 pasangan kueri–dokumen. Penilaian relevansi dilakukan secara manual oleh peneliti sebagai satu penilai berdasarkan kesesuaian antara kueri dan informasi pada judul, snippet, serta isi artikel. Label 0 diberikan apabila artikel tidak membahas kebutuhan informasi pada kueri, label 1 apabila artikel membahas topik yang berkaitan tetapi hanya memenuhi sebagian maksud kueri, dan label 2 apabila artikel secara langsung dan jelas memenuhi kebutuhan informasi pada kueri. Penilaian awal dilakukan menggunakan judul dan snippet. Apabila tingkat relevansi belum dapat ditentukan berdasarkan kedua informasi tersebut, isi artikel lengkap diperiksa sebelum label akhir ditetapkan. Dokumen dengan label 1 atau 2 dianggap relevan pada metrik biner, sedangkan nilai 0–2 digunakan sebagai graded relevance pada nDCG. Penilaian ini merupakan ground truth berbasis candidate pool. Oleh karena itu, Recall@k dihitung terhadap seluruh dokumen relevan yang telah dinilai di dalam pool, bukan terhadap seluruh dokumen relevan yang mungkin terdapat dalam korpus. Karena penilaian dilakukan oleh satu penilai, potensi subjektivitas menjadi salah satu keterbatasan penelitian.

2.8 Metrik Evaluasi Efektivitas

Sistem mengambil 20 dokumen teratas dari masing-masing metode untuk keperluan pembentukan *candidate pool*. Sementara itu, evaluasi efektivitas dilakukan pada kedalaman $k=5$ dan $k=10$ karena penelitian berfokus pada kualitas dokumen yang muncul pada peringkat awal hasil pencarian. Dengan demikian, Top-20 digunakan untuk memperluas cakupan penilaian relevansi, sedangkan Top-5 dan Top-10 digunakan sebagai batas evaluasi performa pemeringkatan. Metrik yang digunakan meliputi Precision@k, Recall@k, Average Precision@k, Mean Average Precision (MAP), Reciprocal Rank (RR), Mean Reciprocal Rank (MRR), Discounted Cumulative Gain (DCG), dan normalized Discounted Cumulative Gain (nDCG). Precision dan recall merupakan ukuran yang umum digunakan untuk menilai efektivitas sistem temu kembali informasi [15].

$$\text{Precision@}k = \frac{\text{Jumlah dokumen relevan pada Top} - k}{k} \quad (7)$$

$$\text{Recall@}k = \frac{\text{Jumlah dokumen relevan pada Top} - k}{\text{Jumlah dokumen relevan dalam pool}} \quad (8)$$

$$\text{AP@}k = \frac{\sum_{i=1}^k P @ i \times \text{rel}(i)}{\min(R, k)} \quad (9)$$

$$\text{MAP@}k = \frac{1}{|Q|} \sum_{q \in Q} \text{AP@}k(q) \quad (10)$$

$$RR(q) = \frac{1}{\text{peringkat dokumen relevan pertama}}, MRR = \frac{1}{|Q|} \sum_{q \in Q} RR(q) \quad (11)$$

$$DCG@k = \sum_{i=1}^k \frac{2^{rel(i)} - 1}{\log_2(i + 1)}, \quad nDCG@k = \frac{DCG@k}{IDCG@k} \quad (12)$$

2.9 Prosedur Pengujian Waktu

Waktu pencarian diukur menggunakan `time.perf_counter()`. Setiap metode dan kueri dijalankan tiga kali sebagai warm-up, kemudian diulang sebanyak 30 kali. Pengukuran dimulai ketika sistem menghitung skor seluruh dokumen dan berakhir setelah dokumen teratas diperoleh. Waktu preprocessing serta pembangunan indeks tidak dimasukkan ke dalam waktu pencarian. Seluruh metode dijalankan pada dataset dan lingkungan komputasi yang sama.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Persiapan Data dan Pembentukan Indeks

Setelah menghapus sembilan artikel dengan Content kosong, jumlah dokumen yang digunakan adalah 80.463. Rata-rata jumlah kata turun sebesar 25,26%, dari 330,59 menjadi 247,09. Pengurangan tersebut terutama berasal dari penghapusan stopword, URL, angka, tanda baca, dan token pendek. Vocabulary akhir terdiri atas 90.281 fitur dengan 12.093.988 nilai non-zero pada matriks TF-IDF. Hasil ini menunjukkan bahwa preprocessing mengurangi ukuran teks yang diproses tanpa mengubah jumlah dokumen secara berarti.

Tabel 4. Ringkasan preprocessing dan indeks

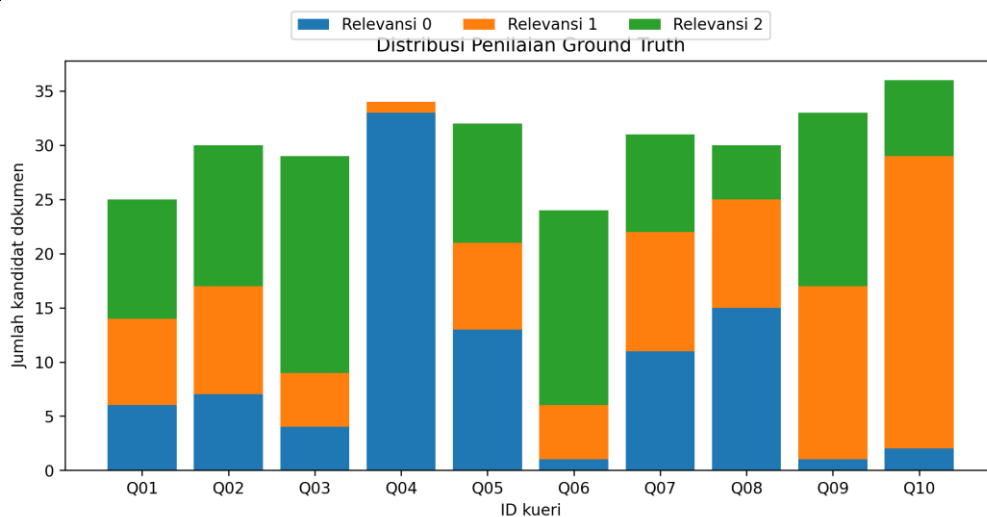
Indikator	Nilai
Dokumen awal	80.472
Content kosong	9
Dokumen akhir	80.463
Rata-rata kata sebelum	330,59
Rata-rata token sesudah	247,09
Jumlah fitur	90.281
Non-zero TF-IDF	12.093.988
Waktu preprocessing	58,215 detik
Waktu pembentukan indeks	14,993 detik

3.2 Distribusi Ground Truth

Dari 304 pasangan kueri-dokumen, sebanyak 93 dokumen diberi label 0, 101 dokumen diberi label 1, dan 110 dokumen diberi label 2. Secara biner, 211 pasangan dianggap relevan. Distribusi antar-kueri tidak seimbang karena setiap topik memiliki jumlah kandidat relevan yang berbeda.

Tabel 5. Distribusi label relevansi per kueri

ID	Kueri	Rel. 0	Rel. 1	Rel. 2	Total
Q01	harga cabai	6	8	11	25
Q02	kenaikan harga beras	7	10	13	30
Q03	pemilihan kepala daerah	4	5	20	29
Q04	hasil pertandingan sepak bola	33	1	0	34
Q05	perkembangan kecerdasan buatan	13	8	11	32
Q06	gempa bumi Indonesia	1	5	18	24
Q07	kasus demam berdarah	11	11	9	31
Q08	kurikulum pendidikan nasional	15	10	5	30
Q09	kasus korupsi pejabat	1	16	16	33
Q10	banjir dan perubahan iklim	2	27	7	36



Gambar 2. Distribusi penilaian ground truth

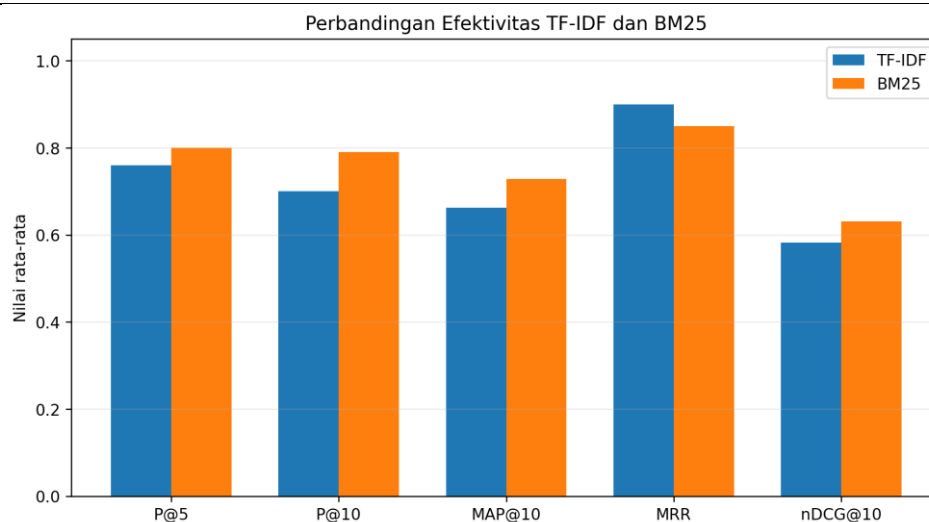
Kueri Q04, yaitu “hasil pertandingan sepak bola”, hanya memiliki satu dokumen relevan dalam candidate pool dan dokumen tersebut tidak muncul pada Top-10 kedua metode. Kondisi ini menunjukkan bahwa kueri tersebut lebih sulit dibandingkan kueri lain. Penyebab yang mungkin adalah kelangkaan artikel yang benar-benar menyebut hasil atau skor pertandingan, atau penggunaan istilah yang terlalu umum. Kueri sulit tetap dipertahankan karena memberikan informasi mengenai keterbatasan metode.

3.3 Evaluasi Efektivitas Keseluruhan

Tabel 6 menampilkan rata-rata metrik pada 10 kueri. Pada Top-5, BM25 menghasilkan Precision@5, Recall@5, dan nDCG@5 yang lebih tinggi. TF-IDF hanya sedikit lebih tinggi pada MAP@5. Pada Top-10, BM25 memperoleh nilai rata-rata lebih tinggi pada Precision@10, Recall@10, MAP@10, dan nDCG@10, sedangkan TF-IDF memperoleh nilai MRR yang sedikit lebih tinggi. Nilai MRR TF-IDF sebesar 0,90, sedangkan BM25 sebesar 0,85. Perbedaan MRR tersebut terutama dipengaruhi oleh kueri Q08, ketika dokumen relevan pertama BM25 muncul pada peringkat kedua. Hasil ini merupakan perbandingan deskriptif berdasarkan 10 kueri pengujian.

Tabel 6. Rata-rata efektivitas pada 10 kueri

Metode	P@5	R@5	MAP@5	nDCG@5	P@10	R@10	MAP@10	MRR	nDCG@10
TF-IDF	0.760	0.165	0.760	0.582	0.700	0.302	0.663	0.900	0.582
BM25	0.800	0.177	0.756	0.638	0.790	0.352	0.729	0.850	0.631



Gambar 3. Perbandingan metrik efektivitas utama

BM25 memperoleh Precision@10 sebesar 0,79, lebih tinggi dibandingkan TF-IDF sebesar 0,70. BM25 juga memperoleh Recall@10 sebesar 0,351630, MAP@10 sebesar 0,728635, dan nDCG@10 sebesar 0,630976, sedangkan TF-IDF memperoleh nilai berturut-turut sebesar 0,302329, 0,662766, dan 0,582197. Secara deskriptif, hasil tersebut menunjukkan bahwa BM25 menemukan lebih banyak dokumen relevan dan cenderung menempatkan dokumen dengan tingkat relevansi tinggi pada posisi yang lebih

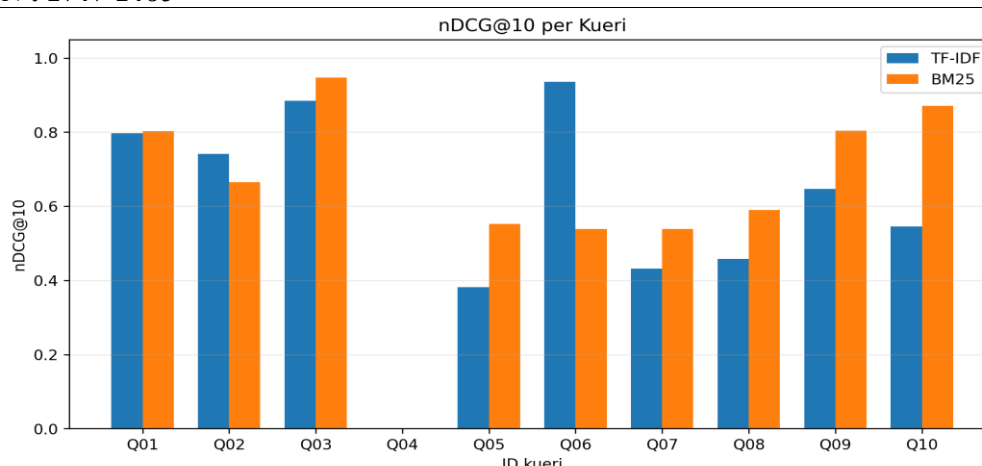
baik dalam Top-10. Namun, TF-IDF memperoleh nilai MRR yang lebih tinggi, sehingga kecenderungan kinerja BM25 tidak berlaku pada seluruh aspek evaluasi.

3.4 Evaluasi Per Kueri

Analisis per kueri diperlukan karena nilai rata-rata dapat menyembunyikan perbedaan karakteristik topik. Tabel 7 menampilkan metrik utama pada Top-10 untuk masing-masing kueri dan metode.

Tabel 7. Metrik Top-10 per kueri

ID	Metode	P@10	R@10	AP@10	RR	nDCG@10
Q01	TF-IDF	0.90	0.474	0.890	1.00	0.797
Q01	BM25	1.00	0.526	1.000	1.00	0.802
Q02	TF-IDF	0.80	0.348	0.758	1.00	0.742
Q02	BM25	0.70	0.304	0.670	1.00	0.665
Q03	TF-IDF	0.90	0.360	0.866	1.00	0.884
Q03	BM25	1.00	0.400	1.000	1.00	0.948
Q04	TF-IDF	0.00	0.000	0.000	0.00	0.000
Q04	BM25	0.00	0.000	0.000	0.00	0.000
Q05	TF-IDF	0.40	0.211	0.215	1.00	0.382
Q05	BM25	0.80	0.421	0.648	1.00	0.552
Q06	TF-IDF	0.90	0.391	0.900	1.00	0.936
Q06	BM25	0.90	0.391	0.835	1.00	0.539
Q07	TF-IDF	0.60	0.300	0.569	1.00	0.431
Q07	BM25	0.80	0.400	0.716	1.00	0.539
Q08	TF-IDF	0.50	0.333	0.429	1.00	0.458
Q08	BM25	0.70	0.467	0.416	0.50	0.590
Q09	TF-IDF	1.00	0.312	1.000	1.00	0.647
Q09	BM25	1.00	0.312	1.000	1.00	0.804
Q10	TF-IDF	1.00	0.294	1.000	1.00	0.545
Q10	BM25	1.00	0.294	1.000	1.00	0.871



Gambar 4. Perbandingan nDCG@10 per kueri

BM25 menunjukkan peningkatan yang cukup besar pada Q05 “perkembangan kecerdasan buatan”, dengan Precision@10 sebesar 0,80 dibandingkan TF-IDF sebesar 0,40, sedangkan AP@10 meningkat dari 0,215 menjadi 0,648. Hasil ini mengindikasikan bahwa mekanisme saturasi frekuensi dan normalisasi panjang dokumen pada BM25 membantu menangani kueri teknologi yang istilahnya dapat muncul dalam berbagai konteks. BM25 juga memperoleh nilai nDCG@10 yang lebih tinggi pada Q03, Q07, Q08, Q09, dan Q10. Pada Q10, kedua metode memiliki nilai Precision@10 dan Recall@10 yang sama, tetapi BM25 memperoleh nDCG@10 sebesar 0,871, lebih tinggi dibandingkan TF-IDF sebesar 0,545, sehingga perbedaannya terletak pada susunan dokumen sangat relevan, bukan pada jumlah dokumen relevan yang ditemukan. Sebaliknya, TF-IDF memperoleh hasil yang lebih tinggi pada Q02 dan Q06. Pada Q06 “gempa bumi Indonesia”, TF-IDF memperoleh nDCG@10 sebesar 0,936, sedangkan BM25 sebesar 0,539, yang menunjukkan bahwa normalisasi panjang dokumen pada BM25 tidak selalu menghasilkan urutan yang lebih sesuai. Performa tersebut dapat dipengaruhi oleh panjang artikel, frekuensi kata “gempa”, variasi lokasi, dan distribusi istilah “Indonesia” dalam korpus. Sementara itu, pada Q04 kedua metode memperoleh nilai 0 pada seluruh metrik Top-10. Hasil ini tidak menunjukkan bahwa artikel sepak bola tidak tersedia dalam dataset, tetapi mengindikasikan bahwa dokumen yang ditemukan belum sesuai dengan maksud kueri yang menekankan hasil pertandingan. Oleh karena itu, kueri yang lebih spesifik, seperti “skor pertandingan timnas Indonesia” atau “hasil pertandingan Persib”, dapat digunakan pada penelitian selanjutnya untuk menguji pengaruh tingkat spesifikasi kueri.

3.5 Kesamaan Hasil Top-10

Kesamaan hasil dianalisis menggunakan jumlah dokumen yang sama, Jaccard similarity, dan kesamaan peringkat pertama. Metrik ini bukan pengganti penilaian relevansi, tetapi membantu menjelaskan seberapa jauh kedua metode menghasilkan daftar yang berbeda.

Tabel 8. Kesamaan hasil TF-IDF dan BM25

ID	Dokumen sama pada Top-10	Jaccard	Top-1 sama
Q01	8	0.667	Ya
Q02	4	0.250	Ya
Q03	6	0.429	Ya
Q04	3	0.176	Tidak
Q05	4	0.250	Tidak
Q06	6	0.429	Tidak
Q07	7	0.538	Ya
Q08	5	0.333	Tidak
Q09	2	0.111	Tidak
Q10	3	0.176	Tidak

Dokumen peringkat pertama kedua metode sama pada Q01, Q02, Q03, dan Q07. Kesamaan terbesar terdapat pada Q01 dengan delapan dokumen yang sama dan Jaccard similarity 0,667. Kesamaan terendah terdapat pada Q09 dengan dua dokumen yang sama dan Jaccard 0,111. Variasi tersebut memperlihatkan bahwa walaupun kedua metode sama-sama bersifat leksikal, normalisasi panjang dan saturasi frekuensi pada BM25 dapat menghasilkan susunan dokumen yang berbeda dari TF-IDF.

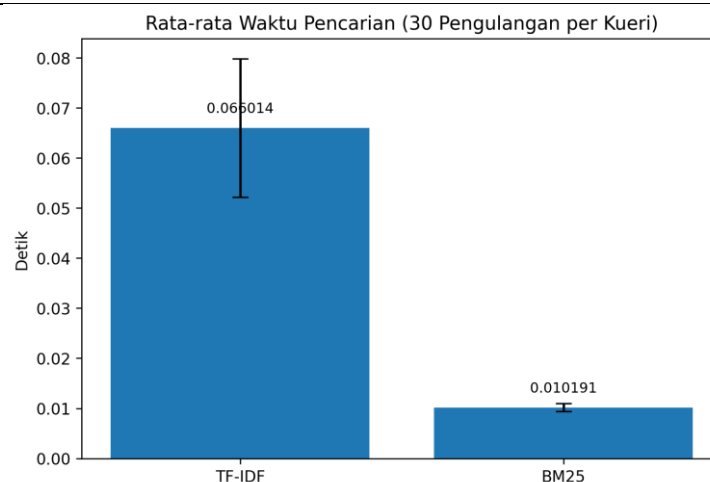
3.6 Perbandingan Waktu Pencarian

Setiap kueri diuji sebanyak 30 kali setelah tiga kali warm-up. Rata-rata waktu TF-IDF adalah 0,066014 detik, sedangkan BM25 adalah 0,010191 detik. Dengan demikian, BM25 sekitar 6,48 kali lebih cepat atau memerlukan waktu sekitar 84,56% lebih rendah pada implementasi dan lingkungan pengujian ini.

Tabel 9. Ringkasan waktu pencarian

Metode	Rata-rata waktu pencarian (detik)	Rata-rata simpangan baku per kueri (detik)	Median waktu rata-rata per kueri (detik)
TF-IDF	0,066014	0,013805	0,062198
BM25	0,010191	0,000836	0,009924

Rata-rata simpangan baku per kueri diperoleh dengan menghitung simpangan baku dari 30 pengulangan pada setiap kueri, kemudian merata-ratakannya untuk seluruh kueri. Median waktu rata-rata per kueri merupakan median dari rata-rata waktu pencarian masing-masing kueri.



Gambar 5. Perbandingan rata-rata waktu pencarian

Perbedaan waktu dipengaruhi oleh implementasi yang digunakan. TF-IDF menghitung perkalian matriks sparse antara seluruh dokumen dan vektor kueri, sedangkan BM25 pada program ini hanya memperbarui skor dokumen yang mengandung istilah kueri melalui struktur matriks CSC. Oleh karena itu, kesimpulan efisiensi perlu dibatasi pada implementasi, perangkat, dan konfigurasi penelitian ini, bukan dianggap sebagai keunggulan universal semua implementasi BM25.

3.7 Pembahasan Umum

Secara deskriptif, BM25 memperoleh nilai rata-rata lebih tinggi pada Precision@10, Recall@10, MAP@10, dan nDCG@10 serta membutuhkan waktu pencarian yang lebih rendah pada implementasi yang digunakan. Hasil tersebut menunjukkan kecenderungan keseimbangan efektivitas dan efisiensi yang lebih baik bagi BM25 pada skenario pengujian ini. Hasil tersebut dapat dikaitkan dengan karakteristik BM25 yang menerapkan saturasi frekuensi istilah dan normalisasi panjang dokumen sehingga pengulangan istilah tidak meningkatkan skor secara berlebihan serta artikel yang lebih panjang tidak secara otomatis memperoleh keuntungan. Meskipun demikian, TF-IDF tetap memberikan hasil yang kompetitif karena memperoleh MAP@5 yang sedikit lebih tinggi, MRR sebesar 0,90, serta menghasilkan urutan dokumen yang lebih baik pada kueri Q02 dan Q06. Temuan ini menunjukkan bahwa kinerja setiap metode dipengaruhi oleh karakteristik kueri dan tidak sebaiknya dinilai hanya berdasarkan satu metrik. Sistem yang mengutamakan kemunculan dokumen relevan pertama dapat mempertimbangkan MRR, sedangkan sistem yang menampilkan beberapa hasil teratas dapat lebih memperhatikan Precision, Recall, MAP, dan nDCG. Hasil evaluasi juga dipengaruhi oleh ground truth berbasis *candidate pool* yang hanya mencakup 20 hasil teratas dari masing-masing metode, sehingga dokumen relevan di luar pool belum dinilai. Selain itu, penilaian berdasarkan judul dan *snippet* masih berpotensi menimbulkan ambiguitas. Oleh karena itu, penelitian selanjutnya perlu melibatkan lebih dari satu penilai, menghitung tingkat kesepakatan antarpemilai, serta menyediakan dataset, kode sumber, daftar kueri, dan ground truth sebagai material pendukung untuk meningkatkan reliabilitas dan replikasi penelitian.

4. KESIMPULAN

Penelitian ini membandingkan efektivitas dan efisiensi metode TF-IDF dan BM25 dalam pencarian 80.463 artikel berita berbahasa Indonesia menggunakan 10 kueri multikategori dan ground truth berbasis candidate pool yang terdiri atas 304 pasangan kueri–dokumen dengan skala relevansi 0–2. Berdasarkan nilai rata-rata evaluasi Top-10, BM25 memperoleh Precision@10 sebesar 0,79, Recall@10 sebesar 0,351630, MAP@10 sebesar 0,728635, dan nDCG@10 sebesar 0,630976, sedangkan TF-IDF memperoleh nilai berturut-turut sebesar 0,70, 0,302329, 0,662766, dan 0,582197. Hasil tersebut menunjukkan kecenderungan kualitas pemeringkatan Top-10 yang lebih baik pada BM25 dalam skenario pengujian ini. Namun, TF-IDF memperoleh MRR sebesar 0,90, sedikit lebih tinggi dibandingkan BM25 sebesar 0,85, yang menunjukkan bahwa TF-IDF lebih konsisten dalam menempatkan dokumen relevan pertama pada posisi awal. Dari sisi efisiensi, rata-rata waktu pencarian BM25 sebesar 0,010191 detik, sedangkan TF-IDF sebesar 0,066014 detik. Dengan demikian, implementasi BM25 pada penelitian ini sekitar 6,48 kali lebih cepat. Perbedaan waktu tersebut terbatas pada struktur data, optimasi program, perangkat keras, dan konfigurasi eksperimen yang digunakan sehingga tidak dapat digeneralisasikan pada seluruh implementasi TF-IDF dan BM25. Penelitian ini masih terbatas pada 10 kueri, satu konfigurasi parameter, satu lingkungan komputasi, satu penilai relevansi, ground truth berbasis candidate pool, serta belum disertai pengujian signifikansi statistik. Penelitian selanjutnya dapat menambah jumlah dan variasi kueri, melibatkan lebih dari satu penilai, menghitung kesepakatan antarpemilai, melakukan pengujian signifikansi, mengoptimalkan parameter BM25, serta membandingkan metode leksikal dengan pendekatan dense atau hybrid retrieval.

DAFTAR PUSTAKA

- [1] N. P. R. Adiati *et al.*, “Comparative study of predictive models for hoax and disinformation detection in Indonesian news,” *International Journal of Advances in Intelligent Informatics*, vol. 10, no. 3, p. 504, Aug. 2024, doi: 10.26555/ijain.v10i3.878.
- [2] D. R. Faisal, F. Darari, and R. A. Ryanda, “Granularity-aware legal question answering: a case study of Indonesian government regulations,” *International Journal of Advances in Intelligent Informatics*, vol. 10, no. 3, p. 359, Aug. 2024, doi: 10.26555/ijain.v10i3.1105.
- [3] L. Heryawan, D. Novitaningrum, K. R. Nastiti, and S. N. Mahmudah, “Medical Record Document Search with TF-IDF and Vector Space Model (VSM),” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 3, pp. 847–852, Jun. 2024, doi: 10.18517/ijaseit.14.3.19606.
- [4] V. Vichianchai and S. Kasemvilas, “A New Term Frequency with Gaussian Technique for Text Classification and Sentiment Analysis,” *Journal of ICT Research and Applications*, vol. 15, no. 2, pp. 152–168, Oct. 2021, doi: 10.5614/itbj.ict.res.appl.2021.15.2.4.
- [5] Pande Made Risky Cahya Dinatha and Nur Aini Rakhmawati, “Komparasi Term Weighting dan Word Embedding pada Klasifikasi Tweet Pemerintah Daerah,”

-
- Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 9, no. 2, pp. 155–161, May 2020, doi: 10.22146/jnteti.v9i2.90.
- [6] R. Al Rasyid and D. H. U. Ningsih, “Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata,” *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, pp. 170–178, Jan. 2024, doi: 10.35870/jtik.v8i1.1416.
- [7] W. Tanuwijaya, C. E. Setiawan, H. Irsyad, and A. Rahman, “Implementasi TF-IDF dan Cosine Similarity untuk Penyaringan Dokumen Berita Program Makan Siang Gratis Pemerintah Indonesia,” *DEVICE: JOURNAL OF INFORMATION SYSTEM, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY*, vol. 6, no. 2, pp. 322–334, Dec. 2025, doi: 10.46576/device.v6i2.6724.
- [8] K. A. Hambarde and H. Proença, “Information Retrieval: Recent Advances and Beyond,” *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.
- [9] R. A. Satria, I. Indriati, and S. Sutrisno, “Pengelompokan Hasil Pencarian Skripsi Berbahasa Indonesia Menggunakan Metode DBSCAN dengan Pembobotan BM25,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 4, pp. 781–790, Aug. 2023, doi: 10.25126/jtiik.2024106899.
- [10] T. Kheng, J. S. Asri, S. Wahyu, and Y. Yulhendri, “Penerapan Algoritma BM25 dalam Pencarian Lowongan Pekerjaan pada Website Job Portal,” *Bulletin of Computer Science Research*, vol. 5, no. 5, pp. 1029–1038, Aug. 2025, doi: 10.47065/bulletincsr.v5i5.760.
- [11] X. Zhang *et al.*, “MIRACL : A Multilingual Retrieval Dataset Covering 18 Diverse Languages,” *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1114–1131, Sep. 2023, doi: 10.1162/tacl_a_00595.
- [12] M. Markwald, J. Liu, and R. Yu, “Constructing and meta-evaluating state-aware evaluation metrics for interactive search systems,” *Information Retrieval Journal*, vol. 26, no. 1–2, p. 10, Dec. 2023, doi: 10.1007/s10791-023-09426-1.
- [13] A. E. Budiman and A. Widjaja, “Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 6, no. 3, Dec. 2020, doi: 10.28932/jutisi.v6i3.2892.
- [14] N. P. Yunita, “Aplikasi Pencarian Hadis Menggunakan Vector Space Model Dengan Pembobotan TF-IDF Dan Confix-Stripping Stemmer,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 3, pp. 665–676, Jul. 2023, doi: 10.25126/jtiik.2023106736.
- [15] T. T. Prabowo, “Efektivitas Sistem Temu Kembali Informasi Perpustakaan Digital Institut Seni Indonesia (ISI) Yogyakarta dalam Tinjauan Recall dan Precision,” *Media Pustakawan*, vol. 28, no. 1, pp. 37–48, Apr. 2021, doi: 10.37014/medpus.v28i1.1087.