JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer
ISSN: 2615-577X (Online)

Information Retrieval System untuk Pencarian Putusan Perdata Mahkamah Agung Menggunakan TF-IDF dan Cosine Similarity

Rafli Sandiar[1], Budi Endarto[2], Surya Priyambudi[3]

[1]*Program Studi Teknik Informatika, Fakultas Teknik, Universitas Wijaya Putra,*
Jalan Raya Benowo 1-3 Surabaya, Indonesia
[2]*Program Studi Ilmu Hukum, Fakultas Hukum, Universitas Wijaya Putra,*
Jalan Raya Benowo 1-3 Surabaya, Indonesia
[3]*Program Studi Sastra Inggris, Fakultas Bahasa & Sastra, Universitas Wijaya Putra,*
Jalan Raya Benowo 1-3 Surabaya, Indonesia
[1]22053028@student.uwp.ac.id
[2]budiendarto@uwp.ac.id
[3]surya@uwp.ac.id

Abstract— **The large number of civil case decisions published by the Supreme Court of Indonesia makes it difficult to find relevant information, and ordinary keyword search often returns inaccurate results. This study develops a web-based information retrieval system for searching Supreme Court civil decisions using TF-IDF and Cosine Similarity. The novelty of this research lies in the integration of query expansion based on a legal term dictionary to bridge the vocabulary gap between everyday language and legal terminology in decision documents, which has not been widely applied in the domain of Indonesian civil decisions. A collection of 2,454 decisions was gathered from the official Supreme Court decision directory by web scraping and processed through cleaning, case folding, tokenization, stopword removal, and stemming. TF-IDF is used to weight terms, Cosine Similarity is used to rank documents against the query. Evaluation was conducted using precision, recall, and F1-score with ground truth determined by matching legal dictionary terms against document preprocessing results; documents containing query-related terms were considered relevant. In a test with the query "perceraian" (divorce), the system achieved a precision of 66.67%, a recall of 100%, and an F1-score of 80%. It should be noted that this test was limited to a single query and five sample documents, so the results cannot be widely generalized. The system, implemented with Laravel and Python, helps users find civil decisions with search results sorted by relevance.**

Keywords—Information Retrieval, Civil Case Decisions, TF-IDF, Cosine Similarity, Query Expansion

Abstrak— **Banyaknya dokumen putusan perkara perdata di Mahkamah Agung membuat pencarian informasi yang tepat menjadi sulit, dan pencarian berbasis kata kunci biasa sering menghasilkan informasi yang tidak relevan. Penelitian ini bertujuan membangun sistem temu kembali informasi berbasis web untuk menemukan putusan perkara perdata Mahkamah Agung dengan metode TF-IDF dan Cosine Similarity. Kebaruan penelitian ini terletak pada integrasi query expansion berbasis kamus istilah hukum untuk menjembatani kesenjangan kosakata antara bahasa sehari-hari masyarakat dan terminologi hukum dalam dokumen putusan, yang belum banyak diterapkan pada domain putusan perdata Indonesia. Sebanyak 2.454 putusan diambil dari situs resmi Mahkamah Agung melalui web scraping, lalu diproses dengan pembersihan noise, case folding, tokenizing, penghapusan stopword, dan stemming. TF-IDF digunakan untuk memberi bobot pada kata, Cosine Similarity digunakan untuk mengukur kemiripan antara kueri dan dokumen. Evaluasi dilakukan dengan precision, recall, dan F1-score menggunakan ground truth yang ditetapkan melalui pencocokan istilah hukum pada kamus terhadap hasil preprocessing dokumen, dokumen yang memuat istilah terkait kueri dianggap relevan. Pada pengujian dengan kueri "perceraian", sistem memperoleh precision 66,67%, recall 100%, dan F1-score 80%. Perlu dicatat bahwa pengujian ini masih terbatas pada satu kueri dan lima dokumen contoh, sehingga hasil belum dapat digeneralisasi secara luas. Hasil penelitian berupa sistem berbasis web yang membantu pengguna mencari putusan hukum dengan tampilan hasil yang lebih terurut berdasarkan relevansi.**

***Kata kunci*— Information Retrieval, Putusan Perdata, TF-IDF, Cosine Similarity, Query Expansion**

I. PENDAHULUAN

Portal Direktori Putusan Mahkamah Agung (putusan.mahkamahagung.go.id) membuka akses publik terhadap ratusan ribu putusan dari berbagai tingkat dan lingkungan peradilan, termasuk perkara perdata. Hingga tahun 2024, lebih dari 500.000 putusan telah dipublikasikan secara daring. Pada volume sebesar ini, pencarian manual dengan kata kunci sederhana sering membawa pengguna ke hasil yang tidak relevan atau melewatkan putusan penting, karena mekanisme pencarian yang tersedia belum sepenuhnya mempertimbangkan relevansi isi dokumen [1]. Kesulitan

ini bertambah karena putusan perdata berstruktur panjang, berbahasa formal, dan sarat istilah hukum yang spesifik, suatu tantangan yang juga ditemukan pada pengolahan dokumen putusan pengadilan di berbagai negara [1].

Era digital menandai transisi dari industri tradisional menuju era ekonomi yang berbasis pada informasi terkomputerisasi. Revolusi teknologi, yang dipicu oleh perkembangan internet dan evolusi komputer, telah mendorong transformasi seluruh aspek kehidupan ke dalam era digital. Hal ini turut berdampak pada profesi hukum, yang kini harus selaras dengan arus perkembangan di era digital tersebut [2].

Salah satu lembaga yang memegang peran strategis dalam sistem hukum Indonesia adalah Mahkamah Agung. Sebagai puncak kekuasaan kehakiman, Mahkamah Agung berwenang mengadili perkara pada tingkat kasasi, menguji peraturan perundang-undangan di bawah undang-undang terhadap undang-undang, dan melakukan pengawasan tertinggi atas penyelenggaraan peradilan, sebagaimana diatur dalam Undang-Undang Nomor 3 Tahun 2009 tentang Perubahan Kedua atas Undang-Undang Nomor 14 Tahun 1985 tentang Mahkamah Agung. Salah satu bidang yang paling banyak ditangani adalah perkara perdata, yaitu sengketa keperdataan antarpihak seperti sengketa hak milik, waris, perjanjian, dan perbuatan melawan hukum. Putusan atas perkara-perkara tersebut menjadi bahan rujukan bagi praktisi hukum, akademisi, dan masyarakat yang ingin memahami hak-hak keperdataannya, sehingga kemudahan menemukan putusan yang tepat menjadi kebutuhan yang nyata.

Tantangan tersebut menjadi fokus bidang *Information Retrieval* (IR) atau temu kembali informasi, yaitu bidang yang mempelajari metode untuk menemukan kembali informasi yang paling relevan dari sekumpulan dokumen berdasarkan kueri pengguna[3] Seiring bertambahnya dokumen digital, kebutuhan untuk menemukan informasi yang benar-benar relevan secara cepat dan tepat semakin besar [4] Salah satu metode yang banyak digunakan adalah TF-IDF (*Term Frequency–Inverse Document Frequency*), yang memberi bobot lebih tinggi pada kata yang sering muncul dalam sebuah dokumen tetapi jarang ditemukan pada keseluruhan koleksi [5] Sebelum pembobotan, teks lazimnya melalui *preprocessing* yang terbukti berpengaruh pada akurasi temu kembali, khususnya pada teks berbahasa Indonesia yang memiliki morfologi kompleks [6], dan konsistensi bentuk kata dasar hasil *stemming* turut meningkatkan akurasi pengukuran kemiripan [7]. Kemiripan antara kueri dan dokumen kemudian diukur dengan *Cosine Similarity*, yaitu sudut antara dua vektor yang tidak dipengaruhi panjang teks [8].

Kombinasi TF-IDF dan *Cosine Similarity* telah terbukti efektif pada sistem *retrieval* arsip digital [5], sistem rekomendasi artikel berita [9], dan penyaringan aspirasi publik [10]. Pada domain hukum, pendekatan ini mencapai akurasi 93,6% dalam mengukur kemiripan dokumen, meskipun masih terbatas dalam memahami makna semantik istilah hukum [4]. Pendekatan lain seperti BM25 [11] dan *semantic search* [12] memerlukan kalibrasi parameter atau sumber daya komputasi yang jauh lebih besar. Penelitian ini tetap memilih TF-IDF dan *Cosine Similarity* karena ringan, mudah diimplementasikan, dan bobotnya mudah diinterpretasikan sehingga mendukung transparansi sistem kepada pengguna. Era saat ini secara umum dikenal sebagai perubahan metode kerja yang menekankan pada pengelolaan data dan sistem kerja industri melalui kemajuan teknologi serta komunikasi, sekaligus peningkatan efisiensi kerja yang berkaitan dengan interaksi manusia [13].

Persoalan lain muncul dari perbedaan bahasa antara pengguna dan dokumen. Masyarakat awam cenderung memakai istilah sehari-hari, misalnya “cerai”, sedangkan putusan menggunakan istilah hukum baku seperti “gugat cerai” atau “talak”. Pencarian yang hanya mencocokkan kata secara persis akan melewatkan dokumen yang sebenarnya relevan ketika kata yang diketik pengguna tidak sama dengan kata yang tertulis dalam putusan. Karena itu, sistem pencarian putusan membutuhkan tahap yang menjembatani kesenjangan kosakata tersebut, selain metode pembobotan dan pengukuran kemiripan yang mampu mengurutkan dokumen berdasarkan tingkat relevansinya.

Meskipun demikian, penerapan IR pada pencarian putusan perdata Mahkamah Agung masih sangat terbatas [14]. Penelitian ini merancang dan membangun sistem IR berbasis web untuk pencarian putusan perdata Mahkamah Agung menggunakan TF-IDF, *Cosine Similarity*, dan *query expansion* berbasis kamus istilah hukum, serta mengukur kinerjanya dengan *precision*, *recall*, dan *F1-score*. Sistem ini ditujukan bagi advokat, hakim, jaksa, akademisi, mahasiswa, dan masyarakat umum yang tidak berlatar belakang hukum, agar dapat menemukan putusan yang relevan tanpa bergantung pada istilah hukum yang teknis.

Beberapa penelitian terdahulu telah menerapkan TF-IDF dan Cosine Similarity pada sistem temu kembali informasi, seperti pada arsip digital [5], rekomendasi artikel berita [9], dan sistem aspirasi publik [10]. Namun, penerapan Information Retrieval pada putusan perdata Mahkamah Agung Indonesia masih sangat terbatas [1]. Penelitian yang ada umumnya berfokus pada peringkasan teks putusan [1] atau summarization otomatis [2], [11], bukan pada sistem pencarian yang mempertimbangkan kesenjangan kosakata antara bahasa masyarakat umum dan terminologi hukum. Padahal, masyarakat awam cenderung menggunakan istilah sehari-hari seperti "cerai", sedangkan dokumen putusan menggunakan istilah hukum baku seperti "gugat cerai" atau "talak". Oleh karena itu, penelitian ini menawarkan kebaruan berupa integrasi query expansion berbasis kamus istilah hukum ke dalam sistem Information Retrieval untuk putusan perdata Mahkamah Agung, yang bertujuan menjembatani kesenjangan kosakata tersebut sekaligus meningkatkan relevansi hasil pencarian.[15]

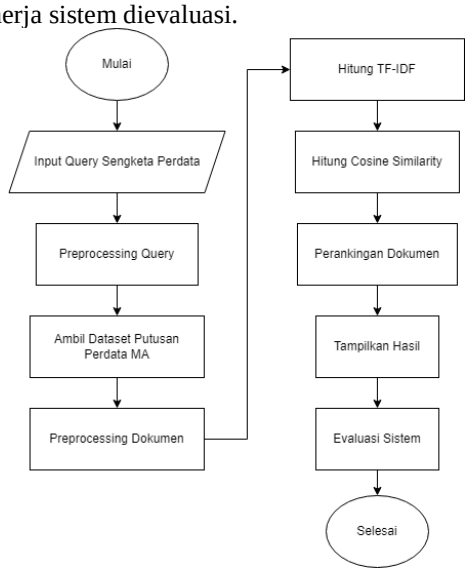
Berdasarkan uraian tersebut, permasalahan yang dikaji adalah bagaimana merancang sistem Information

Retrieval untuk pencarian putusan perdata Mahkamah Agung menggunakan TF-IDF dan Cosine Similarity, serta seberapa akurat sistem tersebut diukur dengan precision, recall, dan F1-score. Penelitian dibatasi pada dokumen putusan perkara perdata yang tersedia untuk umum pada portal resmi Mahkamah Agung sebanyak 2.454 dokumen, pada platform web yang dibangun dengan Laravel 11 dan Python, serta pada pencarian terhadap teks putusan. Secara teoretis, penelitian ini menambah referensi penerapan Information Retrieval pada dokumen hukum berbahasa Indonesia; secara praktis, sistem yang dihasilkan membantu praktisi hukum, akademisi, mahasiswa, dan masyarakat umum menemukan putusan yang relevan dengan lebih cepat. Artikel ini disusun sebagai berikut. Bagian II menguraikan metode penelitian, Bagian III memaparkan hasil dan pembahasan, dan Bagian IV memuat kesimpulan serta saran.

II. METODE PENELITIAN

A. Rancangan Penelitian

Penelitian ini menggunakan metode eksperimen dengan pendekatan kuantitatif yang bersifat aplikatif, karena menghasilkan sistem pencarian berbasis web yang diuji langsung terhadap koleksi putusan perdata. Seluruh proses bertumpu pada perhitungan numerik, yaitu pembobotan kata dengan TF-IDF dan pengukuran kemiripan dengan *Cosine Similarity*. Alur sistem diperlihatkan pada Gambar 1. Kueri pengguna di-*preprocessing*, dataset putusan diambil dan diberi *preprocessing* yang sama, dokumen dibobot dengan TF-IDF, kemiripan dihitung dengan *Cosine Similarity*, dokumen diurutkan, hasil ditampilkan, dan kinerja sistem dievaluasi.



Gambar. 1 Alur penelitian dan proses pencarian putusan perdata

B. Pengumpulan Data

Data berupa dokumen putusan perkara perdata yang diambil dari Direktori Putusan Mahkamah Agung Republik Indonesia [16] menggunakan teknik web scraping dengan bahasa pemrograman Python. Kriteria pemilihan dokumen meliputi: (1) jenis perkara perdata (Pdt.P, Pdt.G, dan Pdt.Sus); (2) dipublikasikan dalam rentang tahun 2020–2024; (3) tersedia dalam format PDF yang dapat diekstraksi; dan (4) memiliki metadata lengkap (nomor putusan, jenis perkara, asal pengadilan, dan tahun). Proses scraping menghasilkan 2.454 dokumen. Validasi hasil ekstraksi dilakukan secara otomatis dengan memeriksa keberadaan field wajib (nomor putusan, jenis perkara, dan isi teks) pada setiap dokumen; dokumen yang tidak memenuhi kriteria dikeluarkan dari koleksi. Dokumen yang tersedia dalam format PDF diekstraksi menjadi teks, kemudian disimpan dalam basis data MySQL beserta nama berkas, nomor putusan, jenis perkara, asal pengadilan, tahun putusan, isi putusan, dan hasil preprocessing.

C. Pembobotan TF-IDF dan Cosine Similarity

Bobot setiap term pada dokumen dihitung dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada (1)–(3), dengan f(t,d) adalah frekuensi term t pada dokumen d, N jumlah dokumen, dan DF(t) jumlah dokumen yang memuat term t.

$$\mathrm{TF}(t,d) = \frac{f(t,d)}{\sum_{t'} f(t',d)} \tag{1}$$

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{DF}(t)}\right) \tag{2}$$

$$w_{t,d} = \mathrm{TF}(t,d) \times \mathrm{IDF}(t) \tag{3}$$

Kueri dan dokumen direpresentasikan sebagai vektor bobot TF-IDF, kemudian kemiripannya dihitung dengan *Cosine Similarity* pada (4), dengan Q adalah vektor kueri dan D vektor dokumen. Nilainya berada pada rentang 0 sampai 1; semakin mendekati 1 semakin relevan dokumen tersebut. Dokumen yang bernilai lebih dari nol diurutkan dari kemiripan tertinggi ke terendah dan ditampilkan sebagai hasil pencarian.

$$\mathrm{sim}(Q,D) = \frac{Q \cdot D}{\|Q\|\|D\|} = \frac{\sum_i q_i d_i}{\sqrt{\sum_i {q_i}^2}\sqrt{\sum_i {d_i}^2}} \tag{4}$$

D. Perancangan Sistem

Antarmuka pengguna dibangun dengan *framework* Laravel 11, sedangkan pengolahan data, perhitungan TF-IDF, dan *Cosine Similarity* dijalankan secara terpisah dengan Python (Jupyter Notebook). Basis data MySQL terdiri atas 14 tabel; enam di antaranya merupakan tabel aplikasi, yaitu putusans (data putusan), users, saved_putusans (putusan tersimpan), search_histories (riwayat pencarian), viewed_documents (dokumen yang dibuka), dan feedbacks (masukan pengguna), sedangkan sisanya adalah tabel bawaan Laravel. Pengembangan dilakukan pada laptop berprosesor Intel Core i7 dengan RAM 32 GB.

E. Evaluasi Sistem

Kinerja diukur dengan precision, recall, dan F1-score [13] pada persamaan (5)–(7). Ground truth ditetapkan melalui dua tahap: (1) pencocokan otomatis istilah pada kamus istilah hukum terhadap hasil preprocessing

dokumen; dan (2) verifikasi manual oleh peneliti untuk memastikan dokumen yang dianggap relevan benar-benar memuat konteks yang sesuai dengan kueri, bukan sekadar kemunculan kata kunci. Dokumen yang memuat istilah terkait kueri dan lolos verifikasi manual dianggap relevan. TP adalah dokumen relevan yang ditemukan, FP dokumen tidak relevan yang ikut ditampilkan, dan FN dokumen relevan yang tidak ditemukan.

Rancangan evaluasi diperluas dengan menggunakan beberapa kueri yang mewakili beragam jenis perkara perdata, yaitu: "perceraian" (perdata permohonan), "waris" (perdata gugatan), "sengketa tanah" (perdata gugatan), "hak asuh anak" (perdata permohonan), dan "wanprestasi" (perdata gugatan). Namun, pada penelitian ini pelaporan hasil difokuskan pada kueri "perceraian" sebagai studi kasus, sedangkan kueri lainnya digunakan untuk menguji konsistensi sistem.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1} = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (7)$$

III. HASIL DAN PEMBAHASAN

A. Hasil Pengumpulan Data

Proses *scraping* menghasilkan 2.454 dokumen putusan perkara perdata yang seluruhnya tersimpan di basis data dan menjadi koleksi sistem. Komposisinya ditunjukkan pada Tabel I; perkara permohonan (Pdt.P) dan gugatan (Pdt.G) mendominasi koleksi. Untuk menjelaskan perhitungan secara rinci, digunakan lima dokumen contoh (D1–D5) yang berasal dari PN Surabaya, PN Simalungun, PN Pekalongan, PN Palembang, dan PN Jakarta Selatan.

TABEL I
KOMPOSISI DATASET PUTUSAN PERDATA

Jenis Perkara	Jumlah	Persentase
Perdata Permohonan (Pdt.P)	1.429	58,23%
Perdata Gugatan (Pdt.G)	1.022	41,65%
Perdata Khusus (Pdt.Sus)	3	0,12%
Total	**2.454**	**100%**

B. Hasil Preprocessing

Tabel II memperlihatkan contoh perubahan teks sebelum dan sesudah *preprocessing*. Hasilnya jauh lebih ringkas: kata tanpa nilai informasi dibuang, kata berimbuhan diubah menjadi kata dasar (misalnya "bercerai" dan "perceraian" menjadi "cerai"), sehingga hanya term yang membedakan isi dokumen yang diproses pada tahap pembobotan.

TABEL II
CONTOH HASIL PREPROCESSING

Sebelum	Sesudah
Bahwa sebelumnya Pemohon pernah berumah tangga dengan seorang perempuan bernama [nama] namun sudah bercerai sebagaimana Kutipan Akta Perceraian yang diterbitkan oleh Dinas Kependudukan dan Pencatatan Sipil Pemerintah Kota Surabaya	mohon rumah tangga perempuan cerai kutip akta cerai terbit dinas duduk catat sipil perintah kota surabaya

C. Query Expansion dan Perankingan Dokumen

Sebelum diproses, kueri diperluas dengan istilah yang bermakna serupa berdasarkan kamus istilah hukum yang disusun pada sistem. Kamus ini dibangun secara manual oleh peneliti dengan merujuk pada Kamus Besar Bahasa Indonesia (KBBI) dan Peraturan Mahkamah Agung terkait, kemudian divalidasi. Aturan perluasan adalah sebagai berikut: setiap istilah umum dalam kueri dipetakan ke satu atau lebih istilah hukum yang memiliki makna setara. Contoh: "perceraian" dipetakan menjadi "perceraian, cerai, gugat cerai, talak". Perluasan ini mengurangi kesenjangan kosakata antara istilah yang digunakan masyarakat umum dan istilah yang dipakai dalam putusan.

Pengujian menggunakan kueri "perceraian". Berdasarkan kamus istilah hukum, kueri ini diperluas menjadi empat istilah seperti pada Tabel III. Dengan perluasan ini sistem tidak hanya mencari kata "perceraian", tetapi juga istilah lain yang berhubungan makna dalam putusan.

TABEL III
HASIL QUERY EXPANSION

Kueri Awal	Kueri Setelah Expansion
perceraian	perceraian, cerai, gugat cerai, talak

Hasil perangkingan pada lima dokumen contoh disajikan pada Tabel IV. Dokumen D2 menempati peringkat pertama dengan nilai *Cosine Similarity* 0,453744, diikuti D5 (0,297950) dan D1 (0,035235). Dokumen D3 dan D4 bernilai 0,0000 sehingga tidak ditampilkan, karena sistem hanya menampilkan dokumen dengan kemiripan lebih dari nol. Hasil ini menunjukkan bahwa TF-IDF dan *Cosine Similarity* mampu mengurutkan dokumen berdasarkan tingkat relevansinya terhadap kueri.

TABEL IV
HASIL RANKING DOKUMEN

Rank	Kode	Nomor Putusan	Jenis Perkara	Nilai Similarity
1	D2	100	Pdt.G	0,453744
2	D5	1012	Pdt.G	0,297950
3	D1	1008	Pdt.P	0,035235

D. Hasil Pengujian

Kueri hasil perluasan, yaitu "perceraian perceraian cerai gugat cerai talak", digunakan sebagai kueri pengujian. Berdasarkan *ground truth*, dokumen yang relevan adalah D1 dan D2. Sistem menampilkan tiga dokumen (D2, D5, D1), sehingga diperoleh TP = 2, FP = 1, dan FN = 0 seperti pada Tabel V.

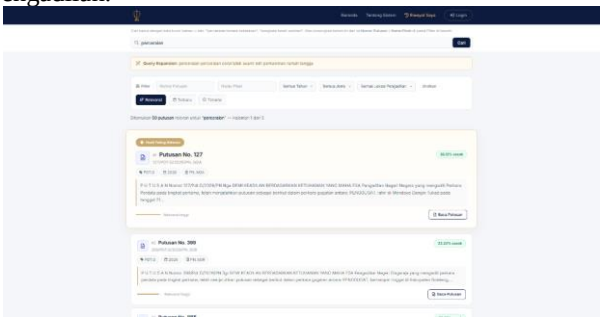
TABEL V
HASIL EVALUASI SISTEM

Parameter	Nilai
Jumlah dokumen relevan (*ground truth*)	2 (D1, D2)
Jumlah dokumen ditemukan	3
True Positive (TP)	2
False Positive (FP)	1
False Negative (FN)	0
Precision	0,6667 (66,67%)
Recall	1 (100%)
F1-Score	0,8 (80%)

Recall sebesar 100% menunjukkan bahwa seluruh dokumen relevan berhasil ditemukan. *Precision* sebesar 66,67% menunjukkan bahwa satu dari tiga dokumen yang ditampilkan (D5) tidak termasuk *ground truth*; hal ini wajar karena TF-IDF dan *Cosine Similarity* menilai kemiripan berdasarkan kata, sehingga dokumen yang memuat kata serupa tetapi berkonteks berbeda dapat ikut muncul. *F1-score* sebesar 80% menunjukkan keseimbangan yang baik antara ketepatan dan kelengkapan hasil. Perlu dicatat bahwa pengujian ini menggunakan satu kueri dan lima dokumen contoh, sehingga generalisasi hasil memerlukan pengujian dengan lebih banyak kueri.

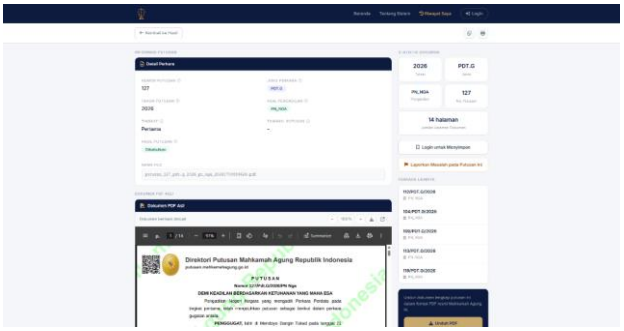
E. Implementasi Sistem

Sistem yang dibangun diberi nama Virlo.id. Pada halaman hasil pencarian (Gambar 2), sistem menampilkan kotak notifikasi *query expansion* yang secara transparan memperlihatkan kata kunci hasil perluasan, lencana persentase kecocokan yang merepresentasikan nilai *Cosine Similarity*, label "Hasil Paling Relevan" pada peringkat pertama, serta panel filter berdasarkan nomor putusan, nama pihak, tahun, jenis perkara, dan lokasi pengadilan.



Gambar. 2 Halaman hasil pencarian sistem Virlo.id

Halaman detail putusan (Gambar 3) menampilkan informasi hasil ekstraksi otomatis, yaitu nomor, jenis perkara, tahun, asal pengadilan, tingkat proses, dan hasil putusan, disertai dokumen PDF asli dan daftar perkara lain dari pengadilan yang sama. Selain itu, sistem menyediakan halaman "Tentang Sistem" yang menjelaskan metode kepada pengguna awam, fitur riwayat, penyimpanan putusan, dan pengiriman masukan, serta panel administrasi dengan fitur *create*, *read*, *update*, *delete* (CRUD) untuk mengelola data hasil *scraping*.



Gambar. 3 Halaman detail putusan sistem Virlo.id

IV. KESIMPULAN

Penelitian ini telah membangun sistem *Information Retrieval* berbasis web untuk pencarian putusan perdata Mahkamah Agung menggunakan TF-IDF, *Cosine Similarity*, dan *query expansion*. Sebanyak 2.454 dokumen putusan berhasil dikumpulkan melalui *web scraping* dan diolah dengan tahapan *preprocessing* yang meliputi pembersihan *noise*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Pada pengujian dengan kueri "perceraian", sistem memperoleh *precision* 66,67%, *recall* 100%, dan *F1-score* 80%, yang menunjukkan bahwa seluruh dokumen relevan dapat ditemukan dengan keseimbangan ketepatan yang baik. Sistem juga menampilkan metode secara transparan kepada pengguna melalui notifikasi *query expansion* dan nilai kecocokan pada hasil pencarian.

Untuk penelitian selanjutnya, disarankan: (1) pengujian dengan jumlah kueri yang lebih banyak dan penelusuran manual dokumen relevan pada seluruh basis data untuk menetapkan ground truth yang lebih komprehensif; (2) penyempurnaan kamus istilah hukum dengan melibatkan lebih banyak ahli hukum; (3) penerapan pendekatan yang lebih mampu menangkap makna semantik seperti word embedding atau model berbasis transformer (misalnya IndoBERT) untuk meningkatkan precision; (4) pengukuran waktu respons sistem secara empiris untuk menguji klaim kecepatan; dan (5) pemanfaatan fitur pelaporan data tidak akurat oleh pengguna untuk memperbaiki kualitas data hasil scraping.

REFERENSI

[1] M. Koniaris, D. Galanis, E. Giannini, and P. Tsanakas, "Evaluation of automatic legal text summarization techniques for Greek case law," *Information*, vol. 14, no. 4, p. 250, Apr. 2023, doi: 10.3390/info14040250.

[2] B. Endarto, A. Syahrul Alam, and S. Atmojo, "Expert system development for law firm using information and communication technology in Indonesia," *KnE Social Sciences*, Mar. 2023, doi: 10.18502/kss.v8i4.12937.

[3] A. Widianto and E. Pebriyanto, "Document similarity using term frequency-inverse document frequency representation and cosine similarity," *Journal of Dinda Data Science, Information Technology, and Data Analytics*, vol. 4, no. 2, pp. 149–153, 2024.

[4] I. H. Junaidi, Sopingi, and S. Sumarlinda, "Sistem retrieval e-arsip Tirta Asata menggunakan algoritma vector space model," *JURNAL FASILKOM*, vol. 14, no. 2, pp. 361–369, Aug. 2024, doi: 10.37859/jf.v14i2.7358.

[5] K. A. Musodo, "Sistem pencarian dokumen berbasis algoritma Nazief Adriani, vector space model, tf-idf serta cosine sebagai

kontribusi kegiatan pengabdian masyarakat untuk perpustakaan umum," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 4, no. 1, pp. 102–110, May 2024, doi: 10.36080/kresna.v4i1.118.

[6] J. Jumadi, D. S. Maylawati, L. D. Pratiwi, and M. A. Ramdhani, "Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process," *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 3, p. 032044, Mar. 2021, doi: 10.1088/1757-899X/1098/3/032044.

[7] A. D. Hartanto, Y. Pristyanto, A. N. Rohman, E. Pujastuti, A. Nurmasani, and I. A. Astuti, "Measuring of scientific document abstraction similarity using Rabin-Karp and Poter stemmer," in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, IEEE, Nov. 2023, pp. 49–54. doi: 10.1109/ICIMCIS60089.2023.10348988.

[8] A. R. Lahitani, "Automated essay scoring menggunakan cosine similarity pada penilaian esai multi soal," *Jurnal Kajian Ilmiah*, vol. 22, no. 2, pp. 107–118, May 2022, doi: 10.31599/jki.v22i2.1121.

[9] A. A. Huda, R. Fajarudin, and A. Hadinegoro, "Sistem rekomendasi content-based filtering menggunakan tf-idf vector similarity untuk rekomendasi artikel berita," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2511.

[10] R. Rismayani, H. SY, T. Darwansyah, and I. Mansyur, "Implementasi algoritma text mining dan cosine similarity untuk desain sistem aspirasi publik berbasis mobile," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 2, pp. 169–176, Aug. 2022, doi: 10.34010/komputika.v11i2.6501.

[11] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[12] R. C. Belwal, S. Rai, and A. Gupta, "Text summarization using topic-based vector space model and semantic measure," *Information Processing & Management*, vol. 58, no. 3, p. 102536, May 2021, doi: 10.1016/j.ipm.2021.102536.

[13] Surya Priyambudi and M. Harist Murdani, "The Development of E-Learning Model for College Students in the Industrial Era 4.0," *Journal of Education and Practice*, no. 11, 2020, doi: 10.7176/jep/11-31-09.

[14] M. Adha and M. Nasucha, "Peringkasan teks putusan pengadilan berbahasa Indonesia menggunakan algoritma textrank," *SINTECH JOURNAL*, vol. 8, no. 1, pp. 12–27, Apr. 2025, doi: 10.31598/sintechjournal.v8i1.1760.

[15] A. Rachmawati, M. Hafid, and D. Irmayanti, "Analisis clustering penyebaran corona virus disease 2019 (covid-19) di Indonesia tahun 2021 menggunakan algoritma k-means," *Jurnal Aplikasi dan Teori Ilmu Komputer*, vol. 7, no. 1, pp. 42–50, Mar. 2024, doi: 10.17509/jatikom.v7i1.49050.

[16] Direktori Putusan MA, "Putusan nomor 452 K/Pdt/2026," 2026.