ISSN (Online) 2747-0563
Volume 6 Tahun 2026

Evaluasi *Hybrid Retrieval* Berbasis Reciprocal Rank Fusion untuk Pencarian Informasi pada Dokumen Peraturan Daerah

Achmad Yusuf Yulestiono[1], Anggraini Puspita Sari[2*], Ardhon Rakhmadi[3]
Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
[1]22081010180@student.upnjatim.ac.id
[2] anggraini.puspita.if@upnjatim.ac.id
[3]ardhon.rakhmadi.fasilkom@upnjatim.ac.id
*Corresponding author email: anggraini.puspita.if@upnjatim.ac.id

Abstrak*— Dokumen Peraturan Daerah memiliki struktur panjang, bahasa normatif, dan banyak istilah administratif yang dapat menyulitkan proses pencarian informasi secara tepat. Penelitian ini bertujuan mengevaluasi kinerja dense retrieval, BM25, dan *hybrid retrieval* berbasis *Reciprocal Rank Fusion* (RRF) pada korpus dokumen Peraturan Daerah Kota Surabaya. Korpus penelitian terdiri atas 106 berkas PDF, dengan 104 dokumen berhasil diekstraksi menjadi teks non-kosong. Dokumen kemudian diproses menggunakan *fixed-size chunking* berukuran 200 kata dengan *overlap* 50 kata sehingga menghasilkan 9.763 *chunk*. *Dense retrieval* dibangun menggunakan model intfloat/multilingual-e5-small, BM25 digunakan sebagai pendekatan *sparse retrieval*, sedangkan *hybrid* RRF menggabungkan peringkat kedua metode dengan konstanta 60. Evaluasi dilakukan menggunakan 100 pertanyaan faktual single-hop dengan Top-K sebesar 5 pada level dokumen dan level *chunk*. Hasil *doc-level* menunjukkan BM25 dan hybrid RRF memperoleh Precision@5 sebesar 0,172, Recall@5 sebesar 0,86, dan F1@5 sebesar 0,2867, lebih tinggi dibanding dense retrieval dengan F1@5 sebesar 0,2667. Pada *chunk-level*, BM25 memperoleh F1@5 tertinggi sebesar 0,2133, diikuti *hybrid* RRF sebesar 0,2000 dan dense retrieval sebesar 0,1433. *Hasil* ini menunjukkan bahwa sinyal leksikal masih dominan pada kueri faktual, sedangkan *hybrid* RRF tetap kompetitif meskipun belum melampaui BM25.

Kata Kunci— ***Hybrid Retrieval, Reciprocal Rank Fusion, BM25, Dense Retrieval, Dokumen Regulasi Daerah.***

I. PENDAHULUAN

Dokumen regulasi daerah umumnya disusun dalam bahasa normatif, panjang, dan sarat rujukan. Dalam kondisi seperti ini, pencarian informasi tidak cukup hanya mengandalkan kecocokan kata secara dangkal, karena satu istilah dapat muncul pada banyak pasal dengan fungsi yang berbeda. Perkembangan transformasi digital mendorong kebutuhan akan sistem yang mampu mengelola dan menyajikan informasi secara lebih cepat, terstruktur, dan mudah diakses. Kebutuhan tersebut juga terlihat pada penelitian-penelitian sebelumnya yang menekankan pentingnya digitalisasi dan pengelolaan informasi terpusat dalam pengembangan sistem informasi [1]. Pada sistem tanya jawab berbasis *Retrieval-Augmented generation* (RAG), kualitas jawaban sangat ditentukan oleh kualitas dokumen atau potongan teks yang berhasil diambil pada tahap *retrieval* [2][3]. Oleh sebab itu, evaluasi terhadap strategi *retrieval* menjadi penting sebelum sistem dikembangkan lebih jauh ke tahap generasi jawaban. Dua pendekatan yang paling umum pada tahap *retrieval* adalah *sparse retrieval* dan *dense retrieval*. BM25 mewakili pendekatan *sparse* yang kuat untuk menangkap kecocokan leksikal dan istilah eksplisit, sedangkan *dense retrieval* memanfaatkan embedding untuk menangkap kemiripan semantik di luar kesamaan kata [4]. Pada dokumen regulasi, kedua pendekatan tersebut memiliki kelebihan dan keterbatasan masing-masing. BM25 cenderung unggul ketika kueri menyebut istilah hukum atau frasa regulatif secara eksplisit, sedangkan *dense retrieval* berpotensi membantu ketika kueri dinyatakan dalam bentuk parafrasa [3], [5].

Untuk memanfaatkan kekuatan kedua pendekatan tersebut secara bersamaan, sejumlah penelitian menggunakan *hybrid retrieval*. Salah satu strategi yang banyak dipakai adalah *Reciprocal Rank Fusion* (RRF), yaitu penggabungan peringkat dari beberapa retriever tanpa perlu menyamakan skala skor aslinya [6]. RRF telah dilaporkan efektif pada berbagai konteks *retrieval*, termasuk query variants, *hybrid dense-sparse retrieval*, dan sistem berbasis RAGeval [7], [8]. Namun, efektivitasnya tetap perlu diuji pada korpus dan karakteristik kueri yang spesifik.

Penelitian ini berfokus pada evaluasi *hybrid retrieval* berbasis RRF untuk pencarian informasi pada dokumen Peraturan Daerah Kota Surabaya. Evaluasi dilakukan dengan membandingkan *dense retrieval*, BM25, dan *hybrid* RRF pada korpus regulasi daerah yang diekstraksi dari PDF. Kontribusi utama penelitian ini adalah: (1) menyajikan implementasi evaluasi *retrieval* pada korpus regulasi daerah yang nyata, (2) membandingkan kinerja ketiga metode secara doc-level dan *chunk*-level, serta (3) memberikan pembacaan yang netral terhadap hasil eksperimen berdasarkan karakteristik kueri faktual single-hop yang digunakan.

II. TINJAUAN PUSTAKA

A. *Retrieval pada Dokumen Regulasi*

Retrieval pada dokumen regulasi memiliki karakteristik yang berbeda dibandingkan korpus umum. Isi dokumen cenderung panjang, menggunakan istilah normatif yang relatif baku, serta memuat banyak rujukan antarperaturan. Dalam situasi seperti ini, sistem pencarian tidak cukup hanya mengandalkan

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

kedekatan semantik, tetapi juga perlu mampu mengenali bentuk leksikal yang eksplisit, seperti nama jenis peraturan, nomor, tahun, dan istilah legal tertentu. Kajian pada domain legal juga menunjukkan bahwa kualitas *retrieval* berpengaruh langsung terhadap ketepatan konteks yang diteruskan ke tahap generasi atau analisis lanjutan [3].

B. Sparse Retrieval dengan BM25

BM25 merupakan salah satu pendekatan *sparse retrieval* yang paling banyak digunakan karena efektif dalam menangkap kecocokan istilah secara eksplisit. Pada kueri yang menyebut nama peraturan, frasa legal, atau entitas administratif tertentu, BM25 umumnya memberikan hasil yang stabil karena bobot relevansi ditentukan oleh frekuensi istilah, *inverse document frequency*, dan panjang dokumen [4].

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

Meskipun demikian, pendekatan ini memiliki keterbatasan ketika kueri dan dokumen menggunakan redaksi yang berbeda untuk konsep yang sama [9].

C. Dense Retrieval Berbasis Embedding

Dense retrieval memetakan kueri dan dokumen ke dalam representasi embedding sehingga hubungan semantik dapat ditangkap meskipun tidak terjadi kecocokan kata secara langsung. Pendekatan ini relevan untuk mengatasi variasi perumusan pertanyaan dan perbedaan redaksi antardokumen. Dalam praktiknya, kedekatan antara kueri dan dokumen umumnya dihitung menggunakan *cosine similarity*.

$$sim(q,d) = \frac{q \cdot d}{|q||d|} \quad (2)$$

Namun, pada domain regulasi, *dense retrieval* dapat kehilangan sinyal penting yang justru bersifat leksikal, misalnya nomor peraturan, istilah hukum, atau nama objek pajak yang sangat spesifik [9], [10]. Oleh karena itu, performanya tidak selalu lebih baik daripada pendekatan *sparse* ketika kueri bersifat faktual dan eksplisit.

D. Hybrid Retrieval dan Reciprocal Rank Fusion

Untuk memanfaatkan kekuatan *sparse retrieval* dan *dense retrieval* secara bersamaan, sejumlah penelitian menggunakan *hybrid retrieval*. Salah satu teknik penggabungan yang banyak dipakai adalah Reciprocal Rank Fusion (RRF), yaitu metode yang menggabungkan beberapa daftar peringkat tanpa menuntut normalisasi skor awal [11]. Metode ini tidak langsung menjumlahkan skor asli dari masing-masing retriever, melainkan menggabungkan posisi peringkat dokumen dari beberapa daftar hasil pencarian. Dengan cara tersebut, RRF relatif stabil karena tidak memerlukan normalisasi skor yang rumit antarmetode.

$$RRF(d) = \sum_{r \in R} \frac{1}{k+rank_r(d)} \quad (3)$$

Pendekatan ini dinilai praktis karena sederhana, stabil, dan tetap efektif ketika sumber ranking memiliki karakteristik yang berbeda. Pada penelitian lain, RRF juga digunakan untuk menggabungkan varian kueri maupun kombinasi *dense-sparse retrieval* dan menunjukkan peningkatan efektivitas pada beberapa skenario [6], [7], [8]. Berdasarkan pertimbangan tersebut, penelitian ini menggunakan RRF untuk mengevaluasi apakah penggabungan ranking *dense* dan BM25 dapat meningkatkan pencarian informasi pada dokumen Peraturan Daerah.

III. METODOLOGI

A. *Korpus dan Praproses*

Eksperimen pada penelitian ini menggunakan korpus dokumen regulasi daerah yang dihimpun dalam format PDF. Berdasarkan proses identifikasi awal, sistem menemukan 106 berkas PDF pada direktori dataset. Setelah dilakukan ekstraksi teks, sebanyak 104 dokumen berhasil diproses dan 2 dokumen teridentifikasi sebagai dokumen dengan teks kosong. Dari keseluruhan dokumen yang berhasil diproses tersebut, total halaman yang terekstrak mencapai 4.973 halaman. Hasil ini menunjukkan bahwa korpus yang digunakan memiliki cakupan yang cukup besar dan heterogen, baik dari sisi ukuran dokumen maupun jumlah halaman, sehingga relevan untuk menguji performa retrieval pada lingkungan dokumen regulasi yang nyata. Dengan detail terlihat pada tabel 1.

TABEL 1
KORPUS DAN KONFIGURASI EKSPERIMEN

Komponen	Nilai
Jumlah PDF	106
Dokumen terekstrak non-kosong	104
Dokumen teks kosong	2
Total halaman terekstrak	4.973
Metode *chunk*ing	Fixed-size, 200 kata
Overlap *chunk*	50 kata
Total *chunk*	9.763
Model *dense*	intfloat/multilingual-e5-small
Fusion	RRF, k = 60
Top-k evaluasi	5
Pertanyaan evaluasi	100 faktual, single-hop

Tahap praproses dilakukan dengan ekstraksi teks menggunakan PyMuPDF, dilanjutkan dengan normalisasi teks untuk mengurangi karakter anomali, spasi berlebih, dan artefak hasil

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

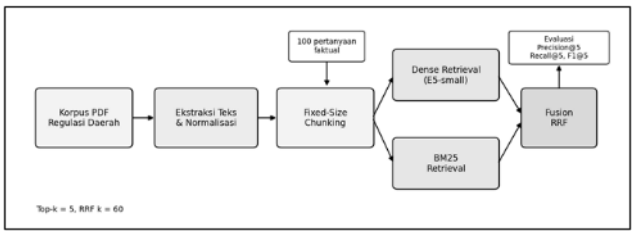
pembacaan PDF. Setelah tahap ini, dokumen dipecah menggunakan pendekatan fixed-size chunking dengan ukuran 200 kata dan overlap 50 kata. Konfigurasi tersebut menghasilkan 9.763 chunk dari 104 dokumen valid, dengan rata-rata 199,20 kata per chunk. Secara praktis, konfigurasi ini dipilih untuk menjaga keseimbangan antara kelengkapan konteks dan efisiensi retrieval, sehingga setiap chunk tetap cukup ringkas untuk diindeks namun masih memuat informasi yang memadai bagi proses pencarian.

Setelah proses ekstraksi, setiap dokumen dipecah menggunakan fixed-size *chunk*ing dengan ukuran 200 kata dan overlap 50 kata. Konfigurasi ini menghasilkan 9.763 *chunk* dari 104 dokumen. Pendekatan fixed-size dipilih untuk menjaga keseragaman unit indeks selama evaluasi *retrieval*, sehingga perbandingan antara *dense retrieval*, BM25, dan *hybrid* RRF tidak dipengaruhi oleh variasi strategi *chunk*ing.

Dari sisi distribusi ukuran, korpus menunjukkan variasi yang cukup tinggi. Rata-rata panjang dokumen berada pada 14.055,68 kata, dengan nilai maksimum mencapai 862.778 kata. Variasi ini mengindikasikan bahwa korpus tidak hanya memuat dokumen berukuran pendek, tetapi juga mencakup regulasi yang sangat panjang dan berpotensi menimbulkan tantangan retrieval tersendiri. Dalam konteks tersebut, proses chunking menjadi komponen penting karena berfungsi mengubah dokumen yang sangat panjang menjadi unit pencarian yang lebih terkontrol.

Dataset evaluasi yang digunakan terdiri atas 100 pertanyaan. Seluruh pertanyaan pada data uji ini berjenis faktual dan dikategorikan sebagai single-hop. Setiap pertanyaan telah dilengkapi dengan gold document dan gold chunk untuk mendukung evaluasi pada dua level, yaitu doc-level dan chunk-level. Komposisi tersebut membuat eksperimen terfokus pada kemampuan sistem dalam menemukan dokumen relevan dan chunk relevan secara langsung, tanpa menambahkan kompleksitas penalaran multi-hop. Dengan demikian, hasil evaluasi lebih mudah diinterpretasikan sebagai dampak dari perbedaan metode retrieval yang diuji.

Eksperimen membandingkan tiga pendekatan retrieval, yaitu dense retrieval, BM25, dan hybrid retrieval berbasis Reciprocal Rank Fusion (RRF). Untuk dense retrieval, penelitian ini menggunakan model intfloat/multilingual-e5-small sebagai pembentuk embedding. Sementara itu, BM25 digunakan sebagai representasi pendekatan sparse retrieval berbasis kecocokan leksikal. Pendekatan ketiga menggabungkan hasil peringkat dari dense retrieval dan BM25 menggunakan RRF. Pada seluruh skenario, jumlah dokumen atau chunk yang diambil ditetapkan sebesar top-k = 5, sedangkan konstanta pada RRF ditetapkan sebesar 60. Konfigurasi ini digunakan secara konsisten untuk seluruh eksperimen agar perbandingan antar metode dapat dilakukan secara adil yang dapat dilihat pada gambar 1.



GBR 1. ARSITEKTUR EVALUASI HYBRID RRF

Evaluasi dilakukan pada dua level. Evaluasi doc-level digunakan untuk menilai apakah sistem berhasil menemukan dokumen sumber yang benar, sedangkan evaluasi chunk-level digunakan untuk memeriksa apakah sistem mampu mengambil unit konteks yang tepat sesuai anotasi gold chunk.

B. *Dense Retrieval*

Komponen *dense retrieval* digunakan untuk menangkap kedekatan makna antara kueri dan dokumen, sehingga sistem tidak hanya bergantung pada kecocokan kata yang sama persis. Pada penelitian ini, dense *retrieval* dibangun menggunakan model *intfloat/multilingual-e5-small*. Model tersebut dipilih karena mendukung teks multibahasa, berukuran relatif ringan, dan cukup sesuai untuk eksperimen retrieval semantik pada dokumen regulasi.

Proses dense retrieval diawali dengan menyiapkan seluruh *chunk* hasil pemotongan dokumen yang memiliki teks valid. Setiap chunk kemudian diubah menjadi representasi vektor (*embedding*) menggunakan model E5. Sebelum proses encoding dilakukan, teks chunk diberi awalan passage: agar format masukan sesuai dengan skema pelatihan model E5. Seluruh *embedding* yang dihasilkan kemudian dinormalisasi sehingga setiap vektor berada pada skala yang seragam. Berdasarkan hasil eksperimen, embedding yang terbentuk memiliki dimensi 384.

Pada saat kueri diberikan oleh pengguna, sistem melakukan proses yang sama terhadap kueri tersebut. Kueri terlebih dahulu diberi awalan *query:*, kemudian diubah menjadi embedding dalam ruang vektor yang sama dengan embedding chunk. Setelah itu, sistem menghitung tingkat kemiripan antara embedding kueri dan seluruh *embedding* chunk menggunakan *cosine similarity* seperti persamaan 2. Nilai kemiripan yang lebih tinggi menunjukkan bahwa chunk tersebut secara semantik lebih dekat dengan isi kueri.

C. *BM25 Retrieval*

Komponen BM25 digunakan sebagai representasi sparse retrieval yang menekankan kecocokan leksikal antara kueri dan dokumen. Pendekatan ini penting pada dokumen regulasi karena banyak kueri mengandung istilah yang sangat spesifik, seperti nama peraturan, nomor dokumen, jenis pajak, atau frasa administratif yang perlu dicocokkan secara langsung.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

Proses BM25 dimulai dengan menggunakan kumpulan chunk yang sama seperti pada dense retrieval. Sebelum diindeks, setiap chunk menjalani tahap praproses sederhana, yaitu mengubah seluruh huruf menjadi huruf kecil, menghapus tanda baca yang tidak diperlukan, merapikan spasi, dan memecah teks menjadi token-token kata. Hasil tokenisasi inilah yang kemudian digunakan sebagai masukan untuk membangun indeks BM25.

Ketika kueri diberikan, kueri juga diproses dengan langkah yang sama, yaitu diubah menjadi huruf kecil, dibersihkan, lalu ditokenisasi. Setelah itu, BM25 menghitung skor relevansi setiap chunk terhadap kueri berdasarkan kemunculan istilah kueri dalam chunk, frekuensi istilah, serta panjang chunk. Dengan mekanisme ini, chunk yang memuat istilah penting dari kueri secara eksplisit akan memperoleh skor yang lebih tinggi.

D. *Hybrid Retrieval Berbasis RRF*

Hybrid retrieval pada penelitian ini dibangun dengan memadukan dua sumber sinyal yang berbeda, yaitu dense retrieval dan BM25. Dense retrieval digunakan untuk menangkap kedekatan makna antara kueri dan chunk, sedangkan BM25 berperan dalam menangkap kecocokan istilah secara eksplisit. Dengan menggabungkan keduanya, sistem diharapkan tidak hanya mampu menemukan chunk yang secara semantik relevan, tetapi juga tetap peka terhadap istilah hukum, nama regulasi, atau frasa administratif yang muncul secara langsung di dalam dokumen.

Pada setiap kueri, sistem terlebih dahulu menjalankan kedua retriever secara terpisah dan menghasilkan daftar peringkat dari masing-masing metode. Hasil peringkat tersebut kemudian digabungkan menggunakan Reciprocal Rank Fusion (RRF). Dalam pendekatan ini, sistem tidak langsung menjumlahkan skor asli dari dense retrieval dan BM25, melainkan menghitung skor gabungan berdasarkan posisi peringkat suatu chunk pada kedua daftar hasil. Dengan cara ini, chunk yang muncul relatif tinggi pada lebih dari satu retriever akan memperoleh skor fusion yang lebih besar, sehingga berpeluang naik ke urutan teratas pada hasil akhir.

Secara matematis, skor RRF dihitung dengan menjumlahkan nilai kebalikan peringkat dari masing-masing retriever. Pada penelitian ini, konstanta RRF ditetapkan sebesar 60 untuk mengurangi dominasi peringkat teratas dan menjaga kontribusi kedua metode tetap stabil. Nilai tersebut juga sesuai dengan praktik yang umum digunakan dalam penelitian terkait rank fusion. Setelah skor fusion seluruh chunk diperoleh, sistem mengurutkan hasil dari skor tertinggi ke skor terendah, lalu mengambil sejumlah chunk teratas sesuai nilai top-k = 5 sebagai keluaran akhir hybrid retrieval.

Pada penelitian ini, top-k hasil *retrieval* ditetapkan sebesar 5. Pemilihan top-k yang kecil dimaksudkan untuk menjaga konteks tetap ringkas dan memudahkan evaluasi Precision@5, Recall@5, dan F1@5. Dengan pengaturan ini, *hybrid* RRF dievaluasi bukan sebagai sistem generatif, melainkan sebagai mekanisme penggabungan peringkat untuk tugas pencarian informasi pada korpus regulasi daerah.

E. *Data Evaluasi dan Metrik*

Data evaluasi terdiri atas 100 pertanyaan yang seluruhnya berlabel faktual dan single-hop. Distribusi dokumen acuan menunjukkan terdapat 20 dokumen gold, masing-masing direpresentasikan oleh 5 pertanyaan. Dengan karakteristik tersebut, evaluasi dalam paper ini lebih menekankan kemampuan sistem menemukan dokumen atau *chunk* yang benar untuk pertanyaan yang bersifat eksplisit, bukan kemampuan menalar lintas dokumen atau multi-hop.

Komposisi seperti ini dipilih karena sesuai dengan tujuan penelitian yang berfokus pada evaluasi kinerja retrieval, bukan pada kemampuan generasi atau penalaran bertingkat. Dengan pertanyaan yang relatif langsung dan terikat pada satu sumber acuan, perbedaan performa antar metode dapat diamati dengan lebih jelas pada tahap pencarian. Selain itu, susunan dataset yang seimbang, yaitu lima pertanyaan untuk setiap dokumen acuan, membantu menjaga agar hasil evaluasi tidak terlalu dipengaruhi oleh dominasi dokumen tertentu. Dengan demikian, analisis yang dihasilkan lebih mencerminkan perilaku umum metode retrieval yang diuji terhadap korpus regulasi yang digunakan.

Evaluasi dilakukan pada dua level. Pertama, doc-level, yaitu menilai apakah dokumen acuan muncul pada lima hasil teratas. Kedua, *chunk*-level, yaitu menilai apakah *chunk* acuan muncul pada lima hasil teratas. Metrik yang digunakan adalah Precision@5, Recall@5, dan F1@5, lalu nilai akhir dilaporkan sebagai rata-rata makro seluruh kueri. Penggunaan dua level evaluasi ini penting karena keberhasilan menemukan dokumen yang benar belum tentu berarti sistem berhasil mengembalikan *chunk* yang tepat.

IV. HASIL DAN PEMBAHASAN

A. *Karakteristik Korpus dan Konfigurasi Eksperimen*

Panjang dokumen pada korpus cukup timpang. Rata-rata panjang dokumen mencapai 14.055,68 kata, tetapi terdapat dokumen yang sangat panjang hingga 862.778 kata. Kondisi ini menunjukkan bahwa pencarian informasi pada regulasi daerah tidak hanya menghadapi keragaman isi, tetapi juga variasi ukuran dokumen yang cukup ekstrem. Dalam situasi seperti ini, pemecahan dokumen ke dalam *chunk* menjadi langkah yang penting agar *retrieval* tidak langsung bekerja pada dokumen utuh yang terlalu panjang.

B. *Hasil Evaluasi Doc-Level*

Ringkasan hasil doc-level menunjukkan bahwa BM25 dan *hybrid* RRF memperoleh nilai yang sama, yaitu Precision@5 sebesar 0,172, Recall@5 sebesar 0,86, dan F1@5 sebesar

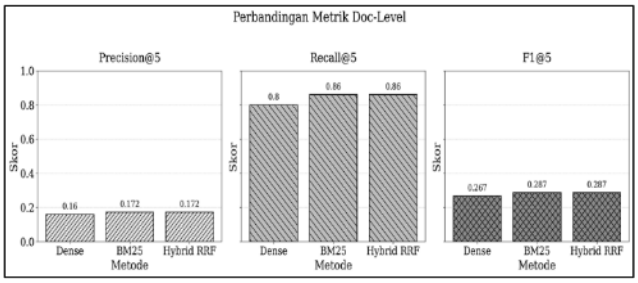
ISSN (Online) 2747-0563
Volume 6 Tahun 2026

0,2867. *Dense retrieval* berada sedikit di bawah keduanya dengan Precision@5 sebesar 0,160, Recall@5 sebesar 0,80, dan F1@5 sebesar 0,2667. Hasil ini menunjukkan bahwa pada level dokumen, penggabungan peringkat dengan RRF belum memberikan peningkatan di atas BM25, tetapi tetap mampu mempertahankan hasil terbaik yang sudah dicapai metode *sparse* yang diperlihatkan pada tabel 2.

TABEL 2

PERBANDINGAN SKOR DOC_LEVEL

Metode	Metrik
Dense	Precision@5 = 0,160; Recall@5 = 0,800; F1@5 = 0,2667
BM25	Precision@5 = 0,172; Recall@5 = 0,860; F1@5 = 0,2867
Hybrid RRF	Precision@5 = 0,172; Recall@5 = 0,860; F1@5 = 0,2867



GBR. 2. PERBANDINGAN METRIK DOC_LEVEL

Secara substantif, hasil pada gambar 2 mengindikasikan bahwa sinyal leksikal masih sangat dominan untuk kumpulan kueri yang digunakan. Seluruh pertanyaan evaluasi bersifat faktual dan single-hop, sehingga banyak di antaranya menggunakan frasa yang dekat dengan redaksi dokumen. Dalam kondisi seperti itu, BM25 sudah cukup kuat untuk menempatkan dokumen acuan pada daftar teratas. RRF tetap bermanfaat karena tidak menurunkan kinerja doc-level, namun kontribusi *dense retrieval* belum cukup kuat untuk mendorong kenaikan metrik di atas BM25.

C. *Hasil Evaluasi Chunk-Level*

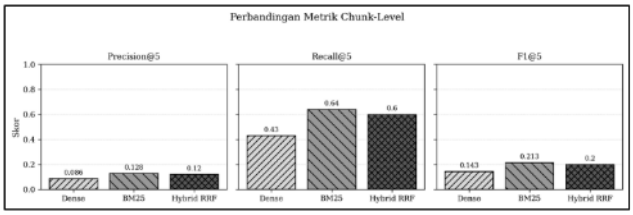
Pada *chunk*-level, perbedaan antarmetode terlihat lebih jelas. BM25 memperoleh Precision@5 sebesar 0,128, Recall@5 sebesar 0,64, dan F1@5 sebesar 0,2133. *Hybrid* RRF berada pada posisi kedua dengan Precision@5 sebesar 0,120, Recall@5 sebesar 0,60, dan F1@5 sebesar 0,2000. *Dense retrieval* menghasilkan nilai terendah, yaitu Precision@5 sebesar 0,086, Recall@5 sebesar 0,43, dan F1@5 sebesar 0,1433.

TABEL 3

PERBANDINGAN SKOR CHUNK_LEVEL

Metode	Metrik
Dense	Precision@5 = 0,086; Recall@5 = 0,430; F1@5 = 0,1433
BM25	Precision@5 = 0,128; Recall@5 = 0,640; F1@5 = 0,2133
Hybrid RRF	Precision@5 = 0,120; Recall@5 = 0,600; F1@5 = 0,2000

Hasil *chunk*-level pada tabel 3 memperlihatkan bahwa menemukan *chunk* yang tepat lebih sulit dibanding menemukan dokumen yang tepat. Pada level dokumen, *hybrid* masih mampu menyamai BM25, tetapi pada level *chunk* justru sedikit berada di bawah BM25. Hal ini dapat dibaca sebagai indikasi bahwa sinyal semantik dari *dense retrieval* memang membantu menemukan dokumen yang topiknya dekat, tetapi tidak selalu cukup presisi untuk mengangkat *chunk* acuan yang benar ke peringkat teratas.



GBR. 3. PERBANDINGAN METRIK CHUNK_LEVEL

Pada gambar 3 korpus Regulasi, *chunk* yang semantik mirip sering kali berasal dari pasal, konsideran, atau lampiran yang berbeda, sehingga fusion dapat membawa kandidat yang relevan secara umum namun tidak identik dengan *chunk* gold. Ringkasan kuantitatif hasil *chunk*-level disajikan pada Tabel III dan divisualisasikan pada Gbr. 3.

D. *Analisis Kualitatif*

Evaluasi Analisis per kueri juga menunjukkan pola yang konsisten dengan hasil agregat. Pada beberapa contoh kueri, seperti "Dokumen mana yang memuat ketentuan umum Pajak Daerah dan Retribusi Daerah Kota Surabaya?", "Regulasi apa yang menjadi pedoman teknis penilaian pajak bumi dan bangunan perdesaan dan perkotaan?", dan "Dokumen apa yang mengatur penghapusan sanksi administratif pajak daerah dalam rangka Hari Jadi Kota Surabaya ke-731?", *dense retrieval* tidak menempatkan dokumen acuan pada top-5. Sebaliknya, BM25 dan *hybrid* RRF berhasil menemukan dokumen acuan pada ketiga contoh tersebut.

Temuan tersebut menguatkan pembacaan bahwa pada kueri yang mengandung istilah regulatif spesifik, penanda administratif, atau frasa baku dari judul dokumen, BM25 masih lebih stabil dibanding *dense retrieval*. *Hybrid* RRF dalam hal

ini bekerja sebagai mekanisme penyangga: ia mampu mempertahankan keberhasilan BM25 pada doc-level, tetapi belum konsisten meningkatkan ketepatan *chunk*. Dengan demikian, RRF pada konfigurasi ini lebih tepat dibaca sebagai strategi penggabungan yang kompetitif, bukan otomatis lebih unggul dari BM25.

Penelitian ini juga memiliki keterbatasan. Evaluasi hanya dilakukan pada 100 pertanyaan faktual single-hop dan menggunakan fixed-size *chunk*ing. Karena itu, hasil yang diperoleh belum dapat digeneralisasi untuk kueri parafrastik yang lebih bebas, pertanyaan multi-hop, atau korpus yang di-*chunk* dengan strategi sadar struktur. Pengujian lanjutan perlu dilakukan pada variasi kueri yang lebih beragam, strategi *chunk*ing lain, serta skenario end-to-end yang melibatkan tahap generasi jawaban.

V. KESIMPULAN

Penelitian ini mengevaluasi *dense retrieval*, BM25, dan *hybrid retrieval* berbasis Reciprocal Rank Fusion pada korpus dokumen Peraturan Daerah Kota Surabaya. Pada konfigurasi eksperimen yang digunakan, BM25 dan *hybrid* RRF menghasilkan kinerja doc-level yang sama, dengan F1@5 sebesar 0,2867, sedangkan *dense retrieval* berada sedikit di bawahnya. Pada *chunk*-level, BM25 memperoleh hasil terbaik dengan F1@5 sebesar 0,2133, diikuti *hybrid* RRF sebesar 0,2000 dan *dense retrieval* sebesar 0,1433.

Temuan ini menunjukkan bahwa pada kueri faktual single-hop, sinyal leksikal masih memegang peranan utama. *Hybrid* RRF tetap relevan karena mampu mempertahankan performa doc-level yang kompetitif dan melampaui *dense retrieval* murni, tetapi belum menunjukkan peningkatan di atas BM25 pada korpus dan skenario evaluasi ini. Ke depan, evaluasi dapat diperluas ke kueri multi-hop, kueri parafrastik, dan strategi *chunk*ing yang lebih selaras dengan struktur regulasi agar potensi *hybrid retrieval* dapat diuji secara lebih komprehensif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing atas bimbingan, arahan, dan masukan yang diberikan selama proses penelitian dan penulisan artikel ini. Terima kasih juga disampaikan kepada Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur yang telah memberikan dukungan akademik selama penelitian berlangsung. Penghargaan turut disampaikan kepada semua pihak yang telah membantu dalam penyediaan data, pelaksanaan eksperimen, dan penyusunan artikel ini.

REFERENSI

[1] Anggraini Puspita Sari, M. M. Al Haromainy, and Ryan Purnomo, "Implementasi Metode Rapid Application Development Pada Aplikasi Sistem Informasi Monitoring Santri Berbasis Website," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 1, pp. 316–325, Mar. 2024, doi: 10.51454/decode.v4i1.348.

[2] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2005.11401

[3] M. Buffa, A. Ferrara, S. Picascia, D. Riva, and S. Castano, "Enhancing legal document building with Retrieval-Augmented Generation," *Computer Law & Security Review*, vol. 59, p. 106229, Nov. 2025, doi: 10.1016/j.clsr.2025.106229.

[4] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[5] Suharyadi and I. Saputra, "Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting," *Journal of Novel Engineering Science and Technology*, vol. 4, no. 02, pp. 71–85, Jul. 2025, doi: 10.56741/jnest.v4i02.1042.

[6] Z. Rackauckas, "RAG-Fusion: a New Take on Retrieval-Augmented Generation," Feb. 2024, doi: 10.5121/ijnlc.2024.13103.

[7] M. Hendriksen, G. Vries, and A. P. De; Potthast, "Open Web Search at LongEval 2023: Reciprocal Rank Fusion on Automatically Generated Query Variants Notebook for the LongEval Lab at CLEF 2023," 2023.

[8] B. Merchant, A. Khazi, and S. S. Sonawane, "Reciprocal Rank Fusion Based Hybrid Dense-Sparse Information Retrieval on Code-Mixed Banglish Social Media Text," 2025.

[9] L. Gao, Z. Dai, T. Chen, Z. Fan, B. Van Durme, and J. Callan, "Complementing Lexical Retrieval with Semantic Residual Embedding," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2004.13969

[10] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, "Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey", doi: 10.1109/ACCESS.2024.0429000.

[11] G. Cormack, C. Clarke, and S. Büttcher, *Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods*. 2009. doi: 10.1145/1571941.1572114.